1-1-2012

A Meta-Analysis of Testing Accommodations for Students with Disabilities: Implications for High-Stakes Testing

Michelle Vanchu-Orosco
University of Denver

A META-ANALYSIS OF TESTING ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES: IMPLICATIONS FOR HIGH-STAKES TESTING

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Michelle Vanchu-Orosco

November 2012

Advisor: Dr. Kathy Green

Author: Michelle Vanchu-Orosco
Title: A META-ANALYSIS OF TESTING ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES: IMPLICATIONS FOR HIGH-STAKES TESTING
Advisor: Dr. Kathy Green
Degree Date: November 2012

Abstract

Test accommodations are designed to ensure the comparability of test scores between students and their typically developing counterparts by eliminating as much construct-irrelevant variance and construct-irrelevant difficulty as possible. Although those involved in test creation endeavor to create tests with suitable accommodations for students with disabilities, there is lack of consensus regarding accommodation efficacy. Using meta-analysis and meta-regression to summarize previous research, this study examined whether test accommodations differentially boost test scores of students with disabilities, and whether accommodated conditions provided a more effective and valid assessment of students with disabilities. Results from the meta-analysis of 34 studies (119 effect sizes) lend support to the differential boost hypotheses, whereby students with disabilities ($\overline{ES} = 0.30$, k = 62, $p < 0.001$) are positively impacted by test accommodations while their typically developing peers ($\overline{ES} = 0.17$, k = 57, $p < 0.001$) gain little from test accommodations.

Presentation assessment accommodations ($\overline{ES} = 0.22$, k = 41, $p < 0.001$) had a small statistically significant impact on the performance of students with disabilities, while use of timing/scheduling accommodations ($\overline{ES} = 0.47$, k = 17, $p < 0.001$) had a small, bordering on medium, statistically significant impact on these students. The effect for presentation accommodations intensified when narrowing the focus to students with learning disabilities ($\overline{ES} = 0.36$, k = 23, $p < 0.001$) but not for timing/scheduling accommodations ($\overline{ES}$ = 0.48, k = 13, $p < 0.001$). Overall results for setting (k = 1) and response (k = 3) accommodations were not available as there were too few studies for an overall comparison.

The results of meta-regression analyses examining the effects of assessment accommodations on test scores for students with disabilities showed that 42% of the heterogeneity in test score could be explained by an overall model examining population description, test characteristic, results dissemination, and researcher-manipulated (test accommodation effect size for students with disabilities) variables. Population description and test characteristic variable sets explained the greatest amounts of variability for mean increase in test score, R^2=0.22 and R^2 =0.35 respectively; researcher-manipulated variable (test accommodation) and research dissemination explained little variance, R^2 =0.07 and R^2 =0.01, respectively.

Acknowledgments

I wish to thank my dissertation committee for their support and guidance through the process of conducting this research. I appreciate the kind assistance of Dr. Kathy Green, my dissertation advisor and mentor, Dr. Monica Martinussen, who was always available to provide guidance as I struggled through the many stages that make up a meta-analysis and meta-regression, and Dr. Karen Riley, who guided me through the ever-changing field of special education. I owe thanks to Dr. Nick Cutforth, Antonio Olmos-Gallo, and Dr. Don Bacon who provided their support for this endeavor. I am grateful to Amelia Dinache for her hard work on the inter-rater reliability portion of this dissertation. I am indebted to Bridget Julian for her eyes and editing skills, helping at very short-notice. Finally, I wish to thank my wonderful family, friends, and my husband, Ty, for their love and support during my years as a graduate student.

Table of Contents

List of Tables

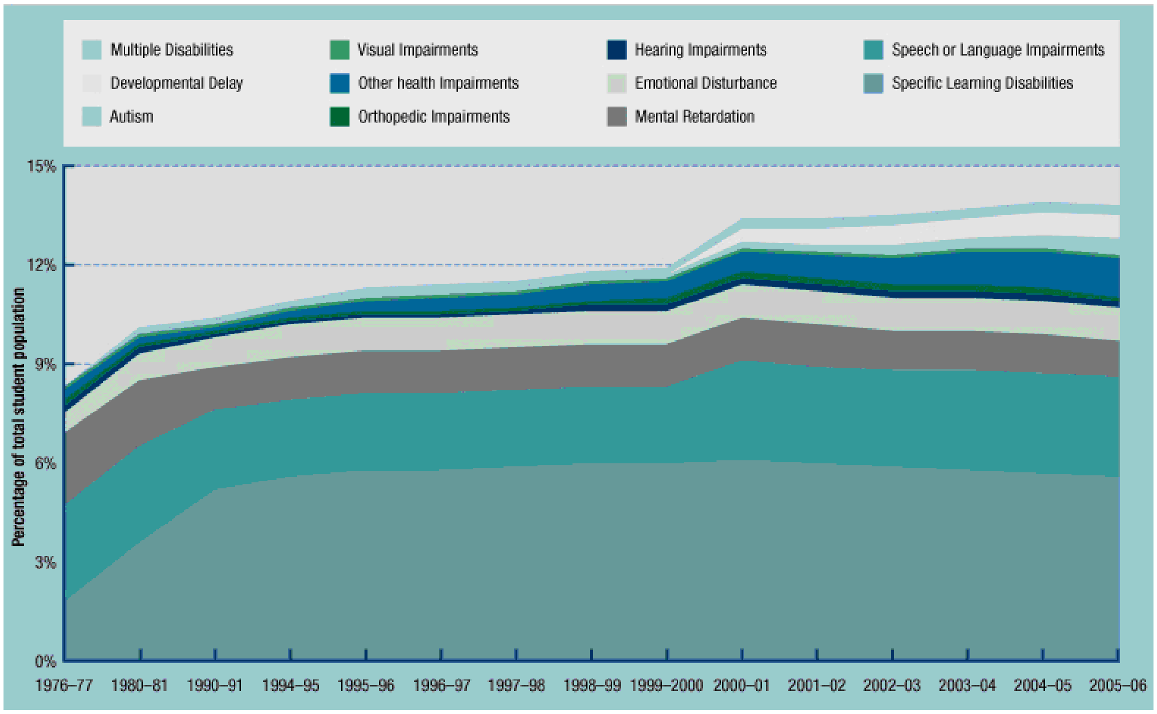
List of Figures

Chapter One

Rationale

The No Child Left Behind Act of 2001 (Public Law 107-110), generally referred to as NCLB, was enacted to ensure that all students learn. Consequently, in an effort to understand what students have learned, there has been an increase in the measurement of student achievement, coupled with an increased emphasis on the assessment of *all* students. States wishing to receive federal funding for their schools have been required to create assessments of basic skills and to test *all* of their students at certain, predetermined grades. The assessments provide one component for the Average Yearly Progress (AYP) reports necessary to ensure funding for schools. Thus, "…the goal [of high-stakes testing] has changed from differentiated standards for a small elite and the larger masses to one of high standards for *all* students" (Linn, 2001, p. 31, emphasis added). This change in direction has led to standardized, high-stakes testing of increasingly larger numbers of special education students.

Concurrently, with the increased emphasis on the assessment of all students, the number of students identified as requiring special education services has increased. In 1977, just over 8% of the total student population was receiving special education services. By 2006 this figure rose to nearly 14% (Dillon, 2007), with approximately 13.5% in K–12 schools receiving special education services (Figure 1: Dillion, 2007). Students with learning disabilities comprise the largest group of students with disabilities, at 6% of the total population of students with disabilities, and represent a diverse population with a wide range of skill strengths and deficits (Fuchs, Fuchs, & Capizzi, 2005). This trend appears to be continuing with recent increases in the identification of children with disabilities, such as autism, receiving national coverage in the popular news; e.g., The New York Times article on 'autism guru' Andrew Wakefield (Dominus, 2011).



Note: Data is for selected years: 1976-77, 1990-91, and 1995 through 2006 (Dillion, 2007)

Figure 1: Prevalence rates of students with disabilities, by disability type, 1977 – 2006.

Students requiring special education services are often referred to as students with special needs, students with disabilities, disabled students, or differently-abled students. Students with disabilities include students who are visually impaired (including blindness), hearing impaired (including deafness), cognitively impaired (including mental retardation), physically/orthopedically impaired (e.g., cerebral palsy, spina bifida,), speech or language impaired, seriously emotionally disturbed (e.g., attention deficit disorder (ADD)), autistic, traumatically brain injured, have other health impairments, or are specifically learning disabled. Such students, once found eligible for special education services under federal and state eligibility/disability standards, receive an Individualized Education Plan (IEP). Laws concerning the identification, funding, and provision of services of such students include the Individuals with Disabilities Education Act (IDEA 2004, Public Law 108-446 reauthorized in 2004), Section 504 of the Rehabilitation Act of 1973, and the Americans with Disabilities Act (ADA).

To provide a way to include students with disabilities in testing efforts, the development and use of *suitable* testing accommodations have been implemented. These accommodations provide a way to include these students in testing efforts, allowing them to perform at optimal levels, and be appropriately assessed. Test accommodations refer to a "… change to testing materials, setting, or procedures that *does not* alter what is being measured" (Thurlow, 2007, p. 2) and are used to promote fairness in testing (Sireci, Li, & Scarpati, 2003). Additionally, the use of accommodations for students with disabilities is thought to allow for the elimination of construct-irrelevant variance (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000a) which, in turn, "… level[s] the playing field so that the format of the test or the test administration conditions do not unduly prevent such students from demonstrating their 'true' knowledge, skills, and abilities" (Sireci et al., 2003, p. 3).

There is a "… great diversity in the way accommodations are created and implemented…" (Sireci et al., 2003, p. 62) with the most common types of testing accommodations for students with disabilities including, but not limited to:

- Presentation – oral test administration,
- Presentation – changes in test content (e.g., simplified language),

- Presentation – changes in test format (e.g., Braille, large print),
- Response – students write directly in test booklet,
- Response – students dictate response (e.g., use scribe),
- Setting – separate room for testing,
- Setting – individual administration,
- Timing/Scheduling – extended/unlimited administration time,

  and
- Timing/Scheduling – break up test administration into separate sessions.

As high-stakes decisions are made using assessment results, the effectiveness of accommodations designed to allow access to assessments and increase the accuracy of student results have been examined. In an effort to provide the most efficacious and appropriate testing accommodations for students requiring special education services, educational researchers have examined differences between these students and their typically developing counterparts for the various types of accommodations (see Bolt & Thurlow, 2006; Helwig & Tindal, 2003; Kosciolek & Ysseldyke, 2000). While these studies provide much-needed research in this area, they are limited to an examination of one or two accommodations for a relatively small sample of students requiring special education services and their typically developing peers. To address this and other shortcoming(s), several summaries of the research literature have been carried out. In particular, the National Center on Educational Outcomes (NCEO) produces a new technical report, summarizing the research literature, approximately every three years. For the most part, these reviews have not provided any firm conclusions regarding the effectiveness of the testing accommodations examined, with most reviews yielding mixed results. As Sireci et al. summarized, “[o]ne thing that is clear from our review is that there are no unequivocal conclusions that can be drawn regarding the effects, in general, of accommodations on students’ test performance” (2003, p. 48).

Prior to NCLB, in an effort to synthesize information on the effects of test accommodations, Chiu and Pearson (1999) conducted a meta-analysis of research looking into the effects of test accommodations for both students requiring special education services and students with limited English proficiency. Their findings did not support the use of testing accommodations for either population of students.

While original research, reviews of the research literature, and meta-analyses have added to our knowledge of testing accommodations for students requiring special education services, they have not provided a definitive understanding of the types of accommodations that are the most useful for these students.

Problem Statement

Students with disabilities are often excluded from the high-stakes tests needed to fulfill annual yearly progress (AYP) obligations for state and federal funding. High-stakes tests, taken without accommodations, generally do not represent these students’ *true* abilities. Such tests introduce construct-irrelevant variance as a type of systematic error (Messick, 1989, 1990, 1995) when students with disabilities are faced with modes of testing (e.g., paper and pencil) with which they are not facile. Construct-irrelevant variance is considered one of two primary threats to construct validity as a “contaminant with respect to score interpretation” (Messick, 1989, p. 34). In addition, construct-irrelevant difficulty, where “aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups” and “… [lead] to construct scores that are invalidly low for those individuals adversely affected" (p. 34) affects test scores for students with disabilities.

Test accommodations are designed to ensure the comparability of test scores between students with disabilities and their typically developing counterparts by eliminating as much construct-irrelevant variance and construct-irrelevant difficulty as possible. While researchers, measurement specialists, and test designers have endeavored to create tests with appropriate accommodations, there is no consensus as to whether or not test accommodations for students with disabilities are indeed effective.

The present study is important because it is an attempt to synthesize previous research in a manner; i.e., meta-analysis of the aggregate research on test accommodations for students with disabilities, that has only been attempted once in the past (see Chiu & Pearson, 1999), presenting what could be more objective results when compared to narrative syntheses of the research literature. As standardized test scores are used to assess AYP and provide school districts and schools with much needed educational funding as well as assessing individual growth and achievement, they must be both *accurate* and *adequate* measures of student knowledge for *all* students. When such tests are inadequate, inaccurate, or invalid measures of student knowledge, the inherent repercussions are manifold. Such repercussions include inadequate or inaccurate placement of students, loss of funding, teacher loss of jobs, and potential school closures.

With extant research limited by the number of accommodations that are addressed and the size of the samples drawn in a single study, it is difficult to draw generalized conclusions about the efficacy of test accommodations. With the introduction of NCLB, numerous studies have been completed. Some of this research points to an interaction between student characteristics and the type of accommodation.

> The interaction hypothesis states that (a) when test accommodations are given to the [students with disabilities] who need them, their test scores will improve, relative to the scores they would attain when taking the test under standard conditions; and (b) students without disabilities will not exhibit higher scores when taking the test with those accommodations (Sireci, Scarpati, & Li, 2005, p. 458).

Most research in this area is restricted by small sample sizes, as classification of students as “students with disabilities” occurs for less than 14% of the general student population. As well, most research and synthesis reports in this area generally aggregate students with disabilities with English language learners (ELL). Currently available research only allows for general accommodation decision-making and implementation guidelines, thus “more empirical study is warranted to further investigate the effects of testing accommodations for students with disabilities” (Bolt & Thurlow, 2004, p. 151).

Purpose of the Study

The purpose of the study was to: (a) determine whether there is empirical support to suggest provision of testing accommodations produces more effective assessment of students with disabilities (b) provide an estimate of the strength of this effect and (c) contribute to the understanding of the effects of test accommodations for this population of students.

Lack of consensus in the research literature regarding the efficacy of test accommodations for students with disabilities has prompted this researcher to investigate the issue of effective test accommodation for students with disabilities using meta-analysis. With the introduction of NCLB, numerous studies have been completed and serve as data points for the present research. Meta-analysis of research on testing accommodation practices allow us to understand which accommodations are being used, in which situations, and for what types of students. This technique also allows us to aggregate data across studies thus providing more power to detect effects that may not be apparent in an individual study, possibly because of the small sample sizes that plague studies focusing on students with disabilities.

In an effort to understand the ramifications of testing accommodations for students with disabilities, this research focused on studies, or portions of studies, examining students with disabilities or students with disabilities and their typically developing peers. Variables reflecting presentation, response, setting, and timing/scheduling test accommodations for paper and pencil tests were included. This study examined studies, or portions of studies, focusing on paper and pencil tests only. Computer-based testing (CBT) and other non-paper and pencil tests were considered inherently different from paper and pencil tests and were not included. Additionally, testing accommodations that are most effective for paper and pencil tests may not be effective for these other types of tests. Studies between 1999 and 2011 were selected for the meta-analysis to further, and not overlap, Chiu and Pearson's (1999) meta-analytic research. This research adds to the existing body of research and research syntheses and extends the original work of Chiu and Pearson (1999) by narrowing the focus from English Language Learners and students with disabilities populations on a variety of different assessments to students with disabilities on standardized, paper and pencil assessments only. Further, meta-regression analyses and graphic representations, not available to Chiu and Pearson in 1999, provide a unique contribution to research in this area.

Sireci et al.s (2005) notion of an interaction hypothesis has been incorporated within the framework of the present meta-analysis. As well, several summaries of the research have provided additional direction regarding research findings on types of accommodations being used, and information on studies in this area. To further our understanding of test accommodations for students with disabilities, salient variables were entered into a meta-regression analysis. Meta-regression was incorporated into this study in order to integrate the effects of multiple, potentially related predictors in an effort to yield a summary of overall prediction of the most effective testing accommodations, as well as examining residual variance and assessing the generalizability of the effects of these accommodations on students with disabilities and typically developing students.

Research Hypotheses

In the current study, the following hypotheses are addressed for the meta-analytic portion of the research:

- Research Hypothesis 1: Is there empirical support for effects of test accommodations for the target group, students with disabilities, as opposed to their typically developing peers?
- Research Hypothesis 2: As measured by effect size, does each of the following constitute an effective accommodation for students with disabilities?
    - Presentation test accommodations?
    - Response test accommodations?
    - Setting test accommodations?
    - Timing/Scheduling test accommodations?

The following research hypothesis is addressed through the meta-regression portion of the current research:

- Research Hypothesis 3: Which type of accommodation(s)–Presentation, Response, Setting, or Timing/Scheduling–more effectively remove construct-irrelevant variance from target students' test scores?

Null hypotheses.

The following null hypotheses are addressed in the meta-analytic portion of the research:

- Research Hypothesis 1: There is *no empirical support* for effects of test accommodations for the target group, students with disabilities, as opposed to their typically developing peers
- Research Hypothesis 2: Test accommodations are not effective.
    - Presentation test accommodations *do not increase* access to test items for target students
    - Response test accommodations *do not increase* access to test items for target students
    - Setting test accommodations *do not increase* access to test items for target students
    - Timing/Scheduling test accommodations *do not increase* access to test items for target students

The following null hypothesis was addressed in the meta-regression portion of the current research:

- Research Hypothesis 3: No test accommodations effectively remove construct-irrelevant variance from target students' test scores

Review of the Literature

Students with disabilities.

There are 13 special education categories listed in federal special education law (Individuals with Disabilities Act reauthorization of 2004, PUBLIC LAW 108–446, 2004). The disabilities cited in the legislation include

> mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as 'emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities (Part A (SEC. 602) (3) (A) (i), 118 STAT.2652, 2004, see Appendix A for the statute in its entirety).

While not in the same definitional area of this law, specific learning disabilities are further spelled out as

> … a disorder in 1 or more of the basic psychological processes involved in understanding or in using language, spoken or written, which disorder may manifest itself in the imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations

and "… includes such conditions as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia" but not "… learning problem[s] that [are] primarily the result of visual, hearing, or motor disabilities, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage" (IDEA, Part A – (30) (A), (B), and (C) (118 STAT.2657 – 118 STAT.2658)). The No Child Left Behind Act of 2001 relies on the definition "under section 602(3) of the Individuals with Disabilities Education Act" (TITLE I A: (111) (b)(2) (C) (v) (II) (cc), 115 STAT. 1451, 2001) when referring to children, or students, with disabilities. As well, the Council for

Exceptional Children (CEC), one of the major organizations worldwide for those involved in the field of Special Education, refers to the same legislation when discussing students with disabilities.

It should be noted that the identification of certain disabilities, such as specific learning disability and emotional disturbance, are often thought to be more subjective (National Association of Special Education Teachers) than disabilities with obvious associated medical or physical conditions such as deafness, blindness, and orthopedic impairments. As well, some of these designations; for example, specific learning disability and emotional disturbance, can be more dynamic and temporary. Students with specific learning disabilities or emotional disturbances may move out of or back into these conditions. Based on the preceding definition, it appears that students with disabilities are indeed a very diverse group.

While other definitions for students with disabilities exist; for example, in countries other than the United States, they were not applied within the scope of this research. Additionally, studies using definitions for students with disabilities found in the research under meta-analysis that could not be aligned with the definition previously cited were removed from the analysis.

Educational legislation and students with disabilities.

The Individuals with Disabilities Act reauthorization of 2004 (PUBLIC LAW 108–446, 2004), or IDEA, and No Child Left Behind (PUBLIC LAW 107-110, 2002), or NCLB, two relatively recent major laws affecting education in the United States have heavily impacted services for, and the assessment of, students with disabilities.

NCLB (2001) requires that educators be accountable for making sure all students, including students with disabilities, meet high expectations. Under TITLE I A (1111) (b)(2) (C) (v) (II) (cc), NCLB breaks out separate measurable annual objectives for students with disabilities as part of state, district, and school accountability for the adequate yearly progress of all students (see Appendix B for the statute in its entirety). Adequate yearly progress (AYP) includes the same high academic standards for all public school students with the expectation of continuous and substantial academic progress, and requires each student to become proficient in mathematics, reading/language skills, and science, with the exception of *low-achieving students*. According to the Council for Exception Children (CEC, 2002), *low-achievers* has not been defined in NCLB. Whether low-achieving students refer to all students with disabilities, a subset of students with disabilities, or some other groups of students is not made clear in the legislation. CEC (2002) believes the definitions in this section of the legislation

> …appear to have the same meaning as *child with a disability* under Sec. 602 of the IDEA …[b]ut judging by the nature of all further stipulations respecting students with disabilities, IDEA eligible and served children constitute the target population being cited (p. 8).

IDEA (2004) focuses on providing a free and appropriate public education (FAPE) to children with diagnosed disorders that impact their ability to learn in a regular classroom setting. As part of FAPE, IDEA Part D (2004) outlines activities to be used to improve the education of children with disabilities. A three-pronged approach for an effective educational system for students with disabilities should:

> (A) maintain *high academic achievement* standards and clear performance goals for children with disabilities, consistent with the standards and expectations for all students in the educational system, and provide for appropriate and effective strategies and methods to ensure that all children with disabilities have the opportunity to achieve those standards and goals;

> (B) clearly define, in objective, measurable terms, the school and post-school results that children with disabilities are expected to achieve; and
> (C) promote transition services and coordinate State and local education, social, health, mental health, and other services, in addressing the full range of student needs, particularly the needs of children with disabilities who need significant levels of support to participate and learn in school and the community ((SEC. 650) (4) (A), (B), and (C), 118 STAT. 2763, 2004), (see Appendix A for the statute in its entirety).

IDEA (2004) provides funding, at the state level, for assessment activities including appropriate accommodations or alternative assessments used to "assess[…] the performance of children with disabilities, in accordance with sections 1111(b) and 6111 of the Elementary and Secondary Education Act of 1965" (Part B (SEC. 611) (e) (2) (C) (x), 118 STAT.2667– 118 STAT.2668, 2004). This is also covered in NCLB (2001) as measurable objectives for all students in statewide assessment programs, including students with disabilities, with provisions for funding assessment accommodations for limited English proficiency (LEP) students and students with disabilities.

Both NCLB and IDEA provide information on assessment of students with disabilities, albeit each with a different focus. As part of AYP, NCLB proposes assessed, measurable objectives of academic standards for accountability include a

> single minimum percentage of students who are required to meet or exceed the proficient level on the academic assessments that applies separately to each group of students described in subparagraph (C) (v) (NCLB, TITLE I A (111) (b)(2)(G)(iii), 115 STAT. 1448),

of which students with disabilities constitute one group. This annual improvement cannot be less than 95% of each of the (C) (v) groups. While there is frequent mention of assessment as it pertains to statewide testing and the Elementary and Secondary Education Act of 1965, or its current reauthorization, NCLB (2001), much of the legislature is concerned with assessment information necessary to develop Individualized

Education Programs (IEPs) for students with disabilities; i.e., use of developmental and other assessments. While developmental and other assessments can be considered high-stakes tests for the student with disabilities, for purposes of the current study high-stakes tests refer to assessments of achievement used for decisions at the school, school district, state, or federal level.

At the federal level, NCLB (2001) and IDEA (2004) have pushed an agenda of assessing improved student achievement through a series of accountability structures. This generally plays out at the state level, as high-stakes tests comprise state assessment programs.

Notwithstanding a lack of definitional clarity of low-achieving students in NCLB, NCLB relying on clarification of this population in IDEA (1997), the full inclusion for students with disabilities is no longer the same type of *choice* it had been prior to the enactment of IDEA's predecessor, PL 94-142 of 1975 (Education of All Handicapped Children Act), with these two pieces of legislation (Crawford & Tindal, 2006; Thurlow, Lazarus, Thompson, & Blount Morse, 2005). Schools, districts, and states are no longer able to exclude students with disabilities, as a group, from assessment requirements; this, in turn, ensures equitable access to assessment and instruction (Baker, 2008). While school districts may decide to exclude some students with disabilities from state-mandated assessments, and states may decide to exclude some students with disabilities from federally mandated assessments, this is becoming more difficult to justify, especially when state, district, and school grant money is tied to AYP as defined in NCLB.

Assessment inclusion for students with disabilities.

Inclusion of students with disabilities in school, district, state, and federal assessment programs, discussed in the following sections, covers the calls for inclusion, the impact of exclusion, and a brief history of inclusion in high-stakes assessment programs for these students.

Calls for inclusion in assessments.

While recognition of the importance of providing services for students with disabilities in the general educational system had been a hotly debated topic for a number of years in the United States, steps toward including students with disabilities in that educational system reached fruition with passage of PL 94-142, the Education of All Handicapped Children Act of 1975. This legislation provided students with disabilities access to the regular educational system. Provisions within this act included a free and public education (FAPE) in the least restrictive environment (LRE) for students with disabilities, and introduced the individualized educational programs (IEP). Students with disabilities now had access to the educational system but were not included in the ongoing district, state, and federal assessment programs.

In the early 1990s, prior to President Bill Clinton's signing IDEA (1997) into law, opinions about including students with disabilities in district, state, and national level assessments differed; in some instances, radically. In 1992, Allington and McGill-Frazen were among the first to document issues with statewide assessment programs, citing lack of inclusion of students with disabilities as potential corruption of assessment results. Other early calls for students with disabilities' inclusion in assessment programs by researchers such as Algozzine (1993), McGrew, Thurlow, Shriner, and Spiegel (1992), Reschly (1993), and Reynolds (1993) were prefaced by the belief that no student, including students with disabilities, should be excluded from testing. Algozzine (1993) argued that excluding students "… violates the spirit and practice of full inclusion" (p. 8) and suggested accommodations or modifications offered to a student be offered to all students. Reynolds (1993) felt universal assessment practices, which allowed for full inclusion, should be used for imperative domains such as language, mathematics, social skills, and self-dependence. McGrew et al. (1992), in their examination of students with disabilities inclusion in federal and state assessment databases, held that it was imperative all students with disabilities able to participate in national and state assessments must participate, as "[t]here is … concern that we … only value who we can measure" (p. 3), emphasizing a need to value students with disabilities. Reschly (1993), in an exploration of advantages and disadvantages of full exclusion, full inclusion, and allowing two percent of students to be excluded, argued that "implementation of liberal accommodations policies would probably increase the perception of fairness and the assessment programs' credibility" (p. 9). As well, the National Center on Educational Outcomes (NCEO) proposed a complex model of six educational outcomes, the assessment of which was considered useful in guiding state and federal agencies educational resource and program policy decisions and reflected commitment to the inclusion of students with disabilities in the assessment of these outcomes to the maximum extent possible (Gilman, Thurlow, & Ysseldyke, 1993; Ysseldyke & Thurlow, 1993).

Perhaps one of the strongest advocates for inclusive models of assessment for students with disabilities, Algozzine (1993) stated "… difference[s] in performance across comparison groups [would be] due to naturally-occurring differences in characteristics of comparison groups" (p. 12) if all students were included in assessment programs. He noted that differences in inclusion practices for students with disabilities in assessment programs between states made state comparisons on standardized assessments virtually meaningless. As an advocate for the full inclusion perspective, Algozzine stressed that permitting IEP data to stand in for state and national assessments taken by general education students and establishing different performance standards for students with disabilities are "… discriminatory, selective practices that … violate the sentiments of full inclusion" (p. 13).

Reschly (1993) proposed a partial inclusion assessment model he felt might counter issues found with total exclusion, or barring students with disabilities' access to standardized state and national assessments. Within this model, students with severe disabilities, constituting approximately two percent of the student population, would be excluded. All other students with disabilities would be included, but would be given the lowest score possible if they did not participate. With such a model, students who would not benefit from participation in the assessment process would not be forced to complete the assessment. Reschly believed such a practice might be considered more equitable and be seen to foster more accurate comparisons of educational units, such as districts and states, when reporting standardized assessments results.

In opposition to full inclusion, based primarily on technological considerations, Merwin (1993) stated that excluding students with disabilities from testing could be justified as "… students in special education comprise such a small number of students that their exclusion [would] not affect state and national comparisons" (p. 8) and that excluding students with disabilities would "… affect group averages less than excluding other subgroups, such as children from low socioeconomic status groups" (p. 8).

In counterpoint, McGrew et al. (1992) declared that it was time to "…address the numerous political and technical hurdles that must be overcome in order for these students to participate more fully in our national and state data collection programs" (p. 8) given the enormity of state and federal support for educational programs for students with disabilities with "… over 4.5 million school-age youngsters receive[ing] some form of special education services, services that are provided at significant expense to our educational system" (p. 10). Thus, an examination of student performance was not only warranted, it was necessary. Algozzine (1993), echoing this sentiment, argued that while considering the inclusion of students with disabilities in federal and state assessments of educational outcomes may not be easy; full inclusion of these students should not be viewed simply as a technical question. Federal and state assessment programs should not dismiss the use of assessment accommodations as they present technical issues that cannot be addressed by psychometric practice. Rather, "… all tests and testing procedures lack perfect technical adequacy" (Algozzine, 1993, p. 13) so we should "simply take a step in some direction" (p. 14). The direction Algozzine (1993) pointed to was to "… avoid any practices that produce, encourage, foster, or facilitate separation among students" (p. 14). To that end, he suggested all students take all tests with any assessment accommodation allowed on one test being allowed on *all tests* for *all students*. In more recent research on design patterns for improving accessibility for test takers with disabilities Hansen and Mislevy state that "… there is a moral imperative to ensure that all students, including individuals with disabilities, have access to assessment products and services" (2008, p. 1).

When IDEA, the Individuals with Disabilities Education Act, was signed into law, the notion of "improving results" was added to the lexicon of access for students with disabilities. The amendments

> reflect[ed] a concern about the standards to which [students with disabilities] [were] held, and about the extent to which they participate[d] in state and district assessments, the primary means that education [uses] to demonstrate educational results (Ysseldyke, Thurlow, Kozleski, & Reschly, 1998, p. 14)

and required states to report on the performance of students with disabilities. Such participation and reporting not only allows for monitoring performance of students with disabilities through the demonstration of improving or declining results; it allows districts and states the ability to provide concrete evidence when justifying the costs of education for students with disabilities. With such legislature and the growing recognition of "… the value of large scale federally funded studies to assess student progress" (McGrew et al., 1992, p. 2) as part of the effort to measure the overall quality of its educational system in United States, students with disabilities' access to district, state, and federal assessment programs has been an issue for over a decade.

It should be noted that the extent to which students with disabilities are included in assessment programs continues to be complicated by domains being assessed, unresolved issues regarding the purpose(s) of assessment and inferences that will be made based on assessment, the type and severity of student's disability, and the measurement procedures used. All of these considerations need to be accounted for when assessing students, as it is the competency under consideration that should be assessed, not the student's disability.

While it was beyond the scope of this research to determine which content areas should be assessed in district, state, and federal assessment programs, research in the areas of language and mathematics was examined as these are considered to be necessary skills in the information and digital ages. As skills in these areas are considered basic to everyday life, understanding the progress of *all* students and program efficacy in teaching these skills cannot be overlooked.

Impact of exclusion from assessment programs.

Prior to the implementation of NCLB (2001), research consistently showed that students with disabilities were not included in district and state assessments; and if these students were included in the assessment process their test scores were not always reported (Elliott, Erickson, Thurlow, & Shriner, 2000). Educational researchers and policy analysts have forwarded several reasons for excluding students with disabilities from district, state, and national assessment programs, particularly *large-scale, high-stakes assessment programs*. Tindal and Fuchs (2000) stated that

> … for many [students with disabilities], the outcomes assessed within general education accountability systems have been viewed as irrelevant to setting and skills required for successful post-school adjustments (p. 9),

further arguing that this notion is reinforced by PL 94-142 (1975) in which student with disabilities' IEPs becomes an individually-referenced, separate apparatus for describing progress for the student with disabilities, with this system of assessment being removed from any existing general assessment systems. Additionally, many schools and school districts have excluded students with disabilities from their general assessment programs in an effort to ensure they do not report poor school progress (McGrew et al., 1992; Reschly, 1993; Tindal & Fuchs, 2000). Alternatively, schools which have included students with disabilities in their assessment program and have reported poor progress have been known to blame the victim, placing failure on the student with disabilities then isolating or removing the student from the school's educational mainstream (Reynolds, 1993).

Exclusion of students with disabilities from assessment programs has often been unwarranted (Reschly, 1993) with two related negative outcomes. One of the outcomes, placing emphasis on producing positive school-level/district-level assessment results in high-stakes decision-making processes, has been the possible discrimination against some students due to existing background characteristics, specifically disabilities, whereby "… conditions [are] ripe for … unwarranted exclusion of students with disabilities or low achievement" (Reschly, 1993, p. 45). Such unwarranted exclusion has been carried out in an attempt to raise average levels of performance on assessments as students with disabilities generally perform at much lower levels than same-grade/age peers have. Unwarranted exclusion is exemplified when students with disabilities with IEP reading goals are excluded from standardized literacy assessments. Methods to exclude students with disabilities from assessments may be a straightforward directive while other exclusion methods may be much more subtle. Anecdotal information provided to Reschly (1993) indicated methods to exclude students with disabilities from assessment efforts took the form of (i) encouraging the student to stay at home on "test day", (ii) marking the student absent on "test day" although they were present, or (iii) having test booklets for students with disabilities invalidated as their answer sheet was not appropriately completed. While the previous examples of exclusionary practices are discriminatory, some types of exclusionary practices are perfectly acceptable; e.g., deciding against assessing the literacy performance of middle school students with extremely low cognitive functioning who do not have literacy goals as their skill levels are below the average skill levels of kindergarten-aged students. Excluding such students from the literacy assessment, perhaps providing them with access to an alternative assessment, is generally considered a more appropriate course of action as including such students would not provide useful information about these students nor their program.

Consequences of exclusion run the gamut from issues with district, state, and national estimates of student performance to the myth of difference between students with disabilities and their typically developing counterparts. To start, many researchers question the accuracy of assessment when not all students participate in the assessment program (Crawford & Tindal, 2006; Elliott et al., 2000; McGrew et al., 1992). McGrew et al. (1992) pointed out that, treating students with disabilities as outliers in data, assessment programs “make it difficult to produce accurate national and state statistical estimates for this population [and] it also raises questions about bias being present in most national and state education statistical estimates that are reported” (p. 29). As Elliott et al. (2000) point out, “[w]ithout the inclusion of all students in accountability systems, incomplete data are reported” (p. 40). Inferences made from assessment results from programs that exclude students with disabilities are questionable. Additionally, exclusion practices are not uniform across districts or states, further complicating any comparisons or generalizations that could be made from the assessment data collected. Policy makers cannot make knowledgeable decisions about students with disabilities and programs for students with disabilities and curriculum based on incomplete information.

This issue is further complicated by the fact that students with disabilities are often excluded from norming samples for standardized tests. As well, most standardized tests are normed without including accommodations. Thus, when students with disabilities are measured using these assessments, they are generally outside the range assessed by the test. As this subgroup is generally not adequately represented, intervention information is suspect.

> Perhaps the primary reason for concern about the exclusion of students with disabilities from state and district assessments [has been] the lack of accountability for the results of education for these students. Intentional exclusion of students, either from testing or from reporting, [means] that there [is] no data available on the results of education for students with disabilities (Yssledyke et al., 1998, p. 15).

Without such data, judgments about student performance or the adequacy of programs for students with disabilities cannot be made. Students with disabilities must be allowed access to assessment programs if we are required, and desire, to see and interpret the results of these assessments to provide systematic information about individual performance for a student with disabilities, aggregate performance for students with disabilities, and the performance of educational programs and curriculum aimed at students with disabilities.

Other documented consequences of exclusion of students with disabilities from assessment programs include increases in retention at grade level, rates of referral to special education, and spurious comparisons among school districts (Thurlow, McGrew, Tindal, Thompson, Ysseldyke, & Elliot, 2000; Ysseldyke et al., 1998). Exclusion from the assessment process often results in exclusion from curriculum or reform initiatives designed to *improve* students' performance (Elliott et al., 2000; Ysseldyke et al., 1998). Further, McGrew et al. (1992), hold that it is imperative all students with disabilities, who are able, participate in national and state assessments as “[t]here is … concern that we … only value who we can measure” (p. 3) with those not being measured becoming non-students, and possibly non-people.

While there has been progress in the area of inclusion, and more states expressly prohibit exclusion of students, exclusionary practices still exist. Christensen, Lazarus, Crone, and Thurlow (2008) found that almost one-third of all states in 2007 provided some reasons students may be excluded from statewide assessment accountability programs. Further, they noted this was an increase from the previous examination of state policies on participation of students with disabilities in 2005.

A brief history of inclusion in high-stakes assessment programs.

With Section 504 of the Rehabilitation Act of 1974 and Title 1 of the Elementary and Secondary Education Act educational accountability came to the forefront. With the increased emphasis on educational accountability “… appropriate testing and reporting of assessment results … increased in importance to educators and policymakers across the nation” (Bolt & Thurlow, 2004, p. 141). With the significant expansion of assessment activities and increasing use of state-level assessments for accountability purposes in the 1990s (Elliott et al., 2000) calls for inclusion of students with disabilities in state accountability systems intensified, leading to inclusion of more students with disabilities in state assessment programs. However, there was little or no documentation on the actual participation rates of students with disabilities, or progress on goals and standards set for *all* learners, on these assessments (Elliott et al., 2000). Additionally, prior to 1996, of the total number of state-level assessments carried out, students with disabilities participation rates could be provided for less than 40% of these assessments (Elliott et al., 2000).

In an effort to better understand inclusion and participation rates of students with disabilities, in January of 1998, 44 people from various educational stakeholder groups met in Washington D.C. to, among other things, "… identify key issues and make recommendations related to assessment practices, research and development" (Ysseldyke et al., 1998, p. 9) and other areas impacted by IDEA 1997. The meeting was convened by the National Center on Educational Outcomes (NCEO) with the Council of Chief State School Officers (CCSSO) and the National Association of Directors of Special Education (NASDSE) also participating. The report generated by this meeting was in response to concerns "about the standards to which [students with disabilities] are held, … the extent to which they participate in state and district assessments, [and] the primary means that education has used to demonstrate educational results" (Ysseldyke et al., p. 14). New requirements generated in IDEA 1997 necessitated that students with disabilities be included in state and district-wide assessments with provision of appropriate accommodations where necessary (Thurlow et al., 2000; Ysseldyke et al., 1998). With the passage of this legislation, the general trend in state-wide assessment programs for those states with assessment programs, general and alternate, was toward "inclusiveness of [students with disabilities] in assessments, rather than toward delineating limitations on either who participates or the accommodations that they can use" (Thurlow et al., p. 162). IEPs began taking on a more pivotal role and were required to include statements about individual modifications to state or district-wide assessments for individual students with disabilities; or, if warranted, participation of a student with disabilities in an alternate assessment instead of the general state/district-wide assessments (Thurlow et al.; Ysseldyke et al.). Federal funding for states and districts now hinged on participation of students with disabilities in statewide assessment programs (IDEA, 1997 Part B funding; Thurlow et al., 2000). As some states reported on participation rates for students with disabilities and performance of student with disabilities for statewide assessments separately, concerns about the accuracy of results reported were raised. For example, a district could report there were 200 students with disabilities and then post the assessment results of students with disabilities based on a fraction (e.g., one-half) of those students taking the statewide assessment (Elliott et al., 2000). Thus, districts and states were called upon to report participation rates for students with disabilities as well as student performance using *standardized reporting procedures*.

Federal legislation changed in the late 1990s (IDEA, 1997) through early 2000 (NCLB, 2001 and IDEA, 2004) partially based on the premise that *all* students can learn, the notion of providing outcomes-based information for students with disabilities education in public accountability systems (Tindal & Fuchs, 2000), and calls for inclusion and participation of students with disabilities in district-wide, state-wide and federal assessment programs. Inclusion of students with disabilities in these mandates focused on accountability systems dealing with improvement of student achievement. The legislation clearly stated that students with disabilities had to be included in state/district-wide assessment programs, with states/districts having to report on (i) participation rates for state/district-wide assessments and (ii) student performance on state/district-wide assessments. Once IDEA (1997) was signed into law, educators had to find ways to include, or in some cases legally exclude, students with disabilities in assessment programs. It was no longer possible to exempt students with disabilities from participating in district and statewide assessments without appropriate documentation or some indication of how their learning would be assessed (Elliott et al., 2000). Now that total exemption was no longer an option, states began looking at how to make decisions about partial participation, out-of-level testing, and alternate assessments (Thurlow et al., 2000).

School accountability for improving education outcomes for *all students* has almost exclusively been addressed through state-wide assessment programs (Thurlow et al., 2005), with inclusion of students with disabilities in these assessment programs as a way for schools to monitor improvement of programs designed for this particular population of students. Inclusion of students with disabilities in statewide assessment programs was "considered essential to improving education opportunities for [students with disabilities] and to providing meaningful and valuable information about student performance to schools and communities" (Thurlow et al., p. 233). With the interplay of statewide assessment programs and school accountability, as well as federal legislation mandating assessment participation decisions for students with disabilities be made by local IEP teams, state policymakers were placed in charge of defining what participation for students with disabilities would look like. State guidelines for inclusion and participation of students with disabilities usually included rules about which assessment accommodations could and could not be used, as well as which students could be excluded from testing (Crawford & Tindal, 2006). Bolt and Thurlow (2004) "… anticipated that nearly all students with disabilities can participate in statewide assessments with appropriate accommodations, with only about 10% of these students requiring the use of an alternate assessment" (p. 142).

Beginning in 1993, NCEO began tracking and analyzing state policies encompassing assessment and accommodation policies for students with disabilities, providing information on the kind and amount of access students with disabilities had to statewide and federal general assessment programs. Each time the NCEO reported on state policies there were significant changes resulting from the report, as statewide accountability efforts began to include statewide assessments in efforts to improve educational programs for all students (Thurlow et al., 2005).

Between 1995 and 1997 there were 34 new or revised policies about participation of students with disabilities in statewide assessment programs (Thurlow et al., 2000). Early NCEO reports showed that 40 of 50 states had active policies on the participation of students with disabilities in state assessment programs. Of the ten states that did not have assessment programs, five were developing or had suspended assessment programs while three were revising participation policies. As well, 36 of 40 states relied on the IEP team's decision, looked at additional criteria (e.g., meaningfulness of testing for students with disabilities, certification of a medical condition, examination of the motivation for a student with disabilities to be like his/her peers, adverse effects of testing on students with disabilities, availability of appropriate accommodations), and/or examined course content or curricular validity when determining inclusion of students with disabilities in their assessment programs (Thurlow et al.). By 2002, research indicated that students with disabilities were being included in statewide assessment programs; however, it was not clear if test scores for students with disabilities were part of state accountability calculations (Bolt, Krentz, & Thurlow, 2002).

By 2001, assessment systems were evolving and all 50 states had state-level participation policies for students with disabilities in place for state or district testing (Thurlow et al., 2005). Additionally, English language learners and students with 504 plans were included in state policies and, thus in the research conducted by NCEO. Policies for participation, as well as accommodations, were becoming more specific for each of these groups. More assessment options were added to state repertoires including general assessment without accommodations, general assessment with accommodations, alternative assessment (available, albeit not always used, in all states), and two procedures not used in state-wide assessment before: (i) out-of-level testing and (ii) partial participation. As well, there were still two states that indicated that they might use the performance of students with disabilities to decide which assessment option was most appropriate. Some of the most notable changes in state participation policies included the rise in the number of state policies that prohibited use of nature or category of disability in assessment participation decision (from 11 to 22 states), looking at whether or not students with disabilities were being instructed in the content being assessed (from 15 to 28 states), and parental involvement in the assessment decision (from 9 to 25 states).

In 2006, Crawford and Tindal examined student assessment inclusion and participation rates in Oregon. They found that the assessment participation rate was part of the accountability structure and, as such, was designed to improve student achievement with the expectation that all students, including students with disabilities, participate in state assessment. To this end, the state was trying to extend the state assessment scale so *all* students would be assessed on a common set of academic standards across several forms of the state assessment. Students could take the state assessment with, or without, accommodations or with modifications (i.e., non-standard or unapproved test accommodations). Students with disabilities could also participate in (i) extended reading/writing/mathematics assessments if they had academic goals in these areas and ‘significant’ disabilities or (ii) extended career and life role assessment. Student assessment scores were aggregated for students who participated, with or without accommodations, in the Oregon general state assessment. However, the scores for students participating in the other assessments were not included as part of the aggregation.

The most recent analysis of inclusion, participation, and accommodations available was conducted by Christensen et al. (2008) and sponsored by NCEO. Christensen et al. (2008) examined 2007 data and found that state policies were still evolving – becoming more detailed and specific at this point in their development. Some states, including Washington D.C., now had policies posted on their websites. Again, though not to the same extent as in previous analyses, participation policies extended testing options for students with disabilities, as well as English language learners and students with 504 plans. Testing options found included:

- state testing without accommodations
- state testing with accommodations
- alternate assessments
- selective participation
- combination participation

- out-of-level assessment
- locally selected assessment
- state testing with modifications or non-standard accommodations

  and
- testing with unique aggregated accommodations.

Christensen et al. (2008) found that there were 27 states, down from 30 states from the previous analyses, providing some type of testing option for every student as well as prohibiting the exclusion of students from their state assessment programs. However, it should be noted that only two of these 27 states explicitly declared "exclusion prohibited." Eight states permitted exclusion and provided waivers based on exemptions such as parental exemption, emotional distress experienced by the student, student medical condition or illness, student refusal, student absence, or other. The "other" category encompassed a wide variety of reasons. For example, in Colorado, other could mean incarceration or the student was a foreign exchange student, and in Alaska, other could mean the student arrived late in school system or the student had a sudden and traumatic experience close to testing time.

Christensen et al. (2008) noted that inclusion of students with disabilities and participation decisions were determined by students' IEPs in all 50 states. Additionally, consideration was given to instructional relevance and instructional goals for the student, the student's current performance and level of functioning, and the student's level of independence when deciding whether the student with disabilities would be included and participate in the statewide assessment program. They noted that, for this group of students, there were many policy changes between 2005 and 2007, with many states citing level of independence, nature or category of disability, and instructional relevance/instructional goals when deciding whether or not to include students with disabilities in the their state-wide assessment program. As well, they found fewer states cited consideration of student needs and characteristics, content/nature/purpose of assessment, and “other” when deciding whether or not to include students with disabilities in the their assessment program. Christensen et al. also explored frequently cited participation decision-making criteria that *were not allowed.* These criteria, relatively unchanged since NCEO’s 2005 data analysis, included presence or category of disability, cultural/social/linguistic/environmental factors, excessive absences, and low expectations/anticipated low scores (with the latter cited by 28% of states).

Guidelines for inclusion in statewide assessment programs have changed very little since McGrew et al. (1992) and Ysseldyke, Thurlow, McGrew, and Shriner (1994) looked into issues of inclusion and exclusion of students with disabilities. By 2000, Elliott et al. found some states implementing some of the previously mentioned guidelines and piloting inclusive testing programs. By 2008, all states had adopted more sophisticated policies, with defined criteria regarding the inclusion and/or exclusion of students with disabilities in their state testing programs (Christensen et al., 2008). However, ideological differences still abound when it comes to inclusion of students with disabilities in assessment programs. Debate still, more often than not, centers on

> …. whether it is more psychometrically sound to base decision making on smaller numbers of students (e.g., general education students) who participate fully in a nonaccommodated test or to base decisions on *all* students, some of whom have had some changes to the test (Thurlow et al., 2000, p. 163).

With the focus now on inclusion for students with disabilities, and with many researchers, educators, and policy-makers looking at participation rates and aggregated data for students with disabilities, there has been a search for new or refined assessment protocols that are more inclusive and attentive to an individual's accessibility needs and preferences (Hansen & Mislevy, 2008). Such protocols have been associated with the universal design of assessments that, from inception, have been designed to be both accessible and valid for the widest range of students possible, including students with disabilities and English language learners. Universal design principles often include formatting changes such as adding bullets or adding white space (Baker, 2008), with "… universal design … mak[ing] … assessment[s] more amenable to accommodations a student may need in order to access the content of the items in the assessment" (p. 20). Though not intimately tied with universal design of assessments, it is hoped that, with the focus on analyses aiming to find some of the most effective accommodations to allow students with disabilities to demonstrate content knowledge rather than disability in federal, statewide, and district-wide assessment programs, this research will aid in the efforts made by those exploring universal test design. To this end, focus is now turned to the types of assessment accommodations provided for students with disabilities in district-wide, statewide, and federal assessment programs.

Accommodations for students with disabilities.

"An assessment accommodation is an alteration in the way a test is administered" (Elliott, Thurlow, Ysseldyke, and Erickson, 1997, p. 1) with the accommodation provided based on student need. Accommodations should not provide a student with an advantage on the content, or construct, being measured. Typically, there are two parts to the definition of assessment accommodation. Accommodations change the way tests are administered, given or taken, under standardized conditions (Bolt & Thurlow, 2004;

Fuchs et al., 2000a) and are intended to facilitate the measurement goals of the assessment (Bolt & Thurlow, 2004). Tindal and Fuchs (2000) reaffirm this definition and add that the construct being measured is not altered and changes are referenced to individual need and differential benefit, not overall improvement.

Assessment accommodations allow students with disabilities to participate in the assessment process in a meaningful way, providing a way to accommodate for a student's disability. Accommodations have been part of the effort to curtail unwanted exclusion of students with disabilities in assessment programs. With assessment accommodations, it is expected that students with disabilities be tested on the content they are expected to have competency in based on their educational experiences, usually noted in their IEPs. While not the only way to ensure all students have access to assessments, accommodations are one of the most frequently used methods of ensuring students, particularly students with disabilities, have access to assessment programs. Additionally, federal laws such as NCLB (2001) and IDEA (2004) require reasonable and valid accommodations to measure the academic achievement of students with disabilities. Even the popular media, in their quest to edify the general public on educational issues, have added to the lexicon of assessment accommodations. For example, Lewin (2002) in the New York Times looked at the question of "how far to accommodate students with learning disabilities on college entrance tests like the SAT" in terms of the "clash between disability rights and educational standards" noting that "requests for special accommodations proliferate, especially from affluent white families."

Variously considered as a way to *level the playing field* (Tindal & Fuchs, 2000), a *corrective lens to decrease distortion* (Chiu & Pearson, 1999), or *tools to help in the* *assessment process* (Enriquez, 2008), assessment accommodations attempt to remove *construct-irrelevant variance* due to the disabilities of students with disabilities. As such, accommodations may remove barriers to assessment access, increasing the probability that the construct, or content, is accurately measured (Baker, 2008).

> [W]ith appropriate accommodations, a student disability…, if unrelated to the constructs being measured, will no longer be a source hindering the true demonstration of their competence. Without accommodations, [students with disabilities] may score lower than they should (Chiu and Pearson, 1999, p. 4).

Thus, when a student with disabilities is not provided with appropriate accommodation[s] they cannot access the test content and are not able to demonstrate their knowledge, making it difficult to accurately measure the student with disabilities' understanding of the content under consideration on the assessment.

In his discussions of test validity, interpretation, and use, Messick (1990, 1995) defines construct-irrelevant variance as a type of systematic error that is introduced into the assessment process. Such error reduces the likelihood that test scores on the assessment adequately reflect the knowledge, or true achievement level, of the test-taker. Of particular interest, *construct-irrelevant difficulty* (Messick, 1995) is some aspect of the task, extraneous to the construct being assessed, that makes the task unduly difficult for some individuals or groups. Construct-irrelevant variance is considered a major source of bias in test scoring, test interpretation, and unfairness in test use.

> [L]ow scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected persons to display their competence, ... [nor should they occur] because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence (Messick, 1995, p. 746).

Low scores, as presented by Messick (1990, 1995); confer an inaccurate representation and a systematic underestimate of the abilities of students with disabilities. It should be noted that assessment accommodations are not considered assessment, or test, modifications as assessment accommodations do not change the construct being assessed.

Additionally, assessment accommodations have been viewed as a method to increase participation in national, state, and/or district assessment programs. Accommodations enhance the perceptions of fairness and credibility for these assessment programs when the same assessment accommodations are used in the same way (Reschly, 1993).

Specific legislation related to assessment accommodations is provided in both NCLB (2001) and IDEA (2004). IDEA (2004) requires participation of students with disabilities in state and district-wide assessments “with appropriate accommodations where necessary” ((SEC. 612) (a) (16) (A)) based on the IEP team and IEP information of the student with disabilities (see (SEC. 614) (d) (1) (A) (V) and (VI)). NCLB (2001) complements IDEA (1997, 2004) with its emphasis on stronger accountability for results. As such, NCLB (2001) requires the participation of all students on state accountability assessments, with provisions for reasonable adaptations or accommodations allowing students with disabilities access to assessment content as defined under section 612(a)(17)(A) of IDEA (2004) (see NCLB (2001): TITLE I A(1111) (b)(2)(I)(ii)).

Types of accommodations.

Assessment accommodations have typically been categorized in four (Thurlow Seyfarth, Scott, & Ysseldyke, 1997; Tindal & Fuchs, 2000; Ysseldyke et al., 1994), five (Christensen et al., 2008; Clapper, Morse, Lazarus, Thompson, & Thurlow, 2003;

Lazarus, Thurlow, Lail, Eisenbraun, & Kato, 2005; Thurlow et al., 2005), or six different categories (Elliott, 1997; Thurlow et al., 2000). Typical categories used to classify assessment accommodations are setting, presentation, timing, response, scheduling, and other. The 'other' category is generally used as a catchall for accommodations that do not fit neatly into the other classification areas. Most frequent categorization schemas place scheduling and timing in the same category as well as including a new category, equipment and materials accommodations, not found in earlier documentation on classification categories (Christensen et al.; Clapper et al.; Lazarus et al.; Thurlow et al., 2005). The number and types of assessment accommodations cited in the literature have varied little over the years research on assessment accommodations for students with disabilities has been conducted.

An example of typical assessment accommodations falling under the various categories follows (Table 1).

Table 1: *Types of Assessment Accommodations*

Setting	**Presentation**
• Administer the test to a small group in a separate location • Administer the test individually in a a separate location • Provide special lighting • Provide adaptive or special furniture • Provide special acoustics • Administer the test in a location with minimal distractions • Administer the test in a small group, study carrel, or individually	• Provide on audio tape • Increase spacing between items or reduce items per page or line • Increase size of answer bubbles • Provide reading passages with one complete sentence per line • Highlight key words or phrases in directions • Provide cues (e.g., arrows and stop signs) on answer form • Secure papers to work area with tape/magnets
Timing	**Response**
• Allow a flexible schedule • Extend the time allotted to complete the test • Allow frequent breaks during the test • Provide frequent breaks on one subtest but not another	• Allow marking of answers in booklet • Tape record responses for later verbatim translation • Allow use of scribe • Provide copying assistance between drafts
Scheduling	**Other**
• Administer the test in several sessions, specifying the duration of each session • Administer the test over several days, specifying the duration of each days' session • Allow subtests be taken in a different order • Administer the test in the afternoon rather than in the morning, or vice versa	• Special test preparation • On-task/focusing prompts • Any accommodation that a student needs that does not fit under the existing categories

Elliot et al., 1997, p. 2

It is generally recommended that

> [a]ccommodations… be provided for the assessment when they are routinely provided during classroom instruction. In other words, when classroom accommodations are made so that learning is not impeded by a student's disability, such accommodations generally should be provided during assessment (Elliott et al., 1997, p. 3).

Research, such as that conducted by NCEO, shows that state lists of approved standard accommodations which are considered not to be a threat to the validity of the assessment or the comparability of test items, vary from state to state and there is limited consensus regarding acceptable, allowable accommodations for students with disabilities (Bolt &

Thurlow, 2004). Perhaps, as a result of legislative requirements for students with disabilities participation in state and district-wide assessment programs, practices in allowing assessment accommodations are quite variable with differences in availability of state guidelines and, when provided, differences in the content of state guidelines on test accommodations. Additionally, "[s]tate accommodation policies are continually changing reflecting uncertainty of educational agencies" (Bolt & Thurlow, p. 142). Thurlow et al. (2000) noted that this lack of agreement across states poses problems, particularly for students with disabilities moving from one state to another.

One of the most frequently allowed accommodations is "[p]roviding extended time or unlimited time to [students with disabilities]" (Chiu & Pearson, 1999, p. 2). More recent research (Bolt & Thurlow, 2004) indicated the five most frequently allowed accommodations for statewide assessment programs are dictated response, large print, Braille, extended-time, and sign language interpreter.

Primary studies of the effectiveness of accommodations.

Many primary studies examining the effectiveness of testing accommodations for students with disabilities can be found in the literature. Primary research in this area usually falls under one of three research designs: experimental where the test administration condition was manipulated and there was random assignment to condition, quasi-experimental where the test administration condition was manipulated but students weren't randomly assigned to condition, and non-experimental often using an *ex post facto* comparison of students taking a standard version and an accommodated version of the same test. An example of primary research using each one of these designs follows.

Calhoon, Fuchs, and Hamlett (2000) provide an example of a primary study on the effectiveness of testing accommodations using an experimental design. Calhoon et al. compared the effects of computer-based test accommodation, non-computer-based test accommodation; i.e., teacher oral presentation, and no accommodation conditions on a constructed-response mathematics performance assessment. Four different testing conditions were examined (i) standard administration, (ii) teacher-read administration, (iii) computer-read administration, and (iv) computer-read administration accompanied by video. Over the course of four weeks 81 ninth- through twelfth-grade students with disabilities who were receiving mathematics and reading instruction in special education resource rooms, based on IEPs, were assessed under each of the different, counterbalanced testing conditions. The researchers found that students with disabilities performed better when the assessment was read aloud than when a standard paper and pencil administration was used, with the effect sizes ranging from approximately one-quarter to one-third of a standard deviation. There were no significant differences between the oral presentation, teacher versus computer, conditions. However, a survey of the students with disabilities indicated that they preferred the computer oral presentation as it afforded them anonymity when taking the test. A major limitation of this research relates to only using students with disabilities. The authors suggested that future research includes both students with disabilities and typically developing students in the analyses.

Helwig and Tindal (2003) provide an example of a primary study on the effectiveness of testing accommodations using a quasi-experimental design. Helwig and Tindal investigated the accuracy with which special education teachers were able to recommend oral accommodations for students. Using a 5-point Likert scale, teachers were asked to judge a student’s proficiency in reading and mathematics and then rate how important an oral accommodation would be to the student’s success on one of two forms (A and B) of a thirty-item, multiple-choice mathematics assessment. Students with disabilities (n = 245) and typically developing students (n = 973) in fourth through eighth grades in eight states then took an accommodated, items read aloud via a video presentation, and a non-accommodated form of the mathematics test. Research results were contraindicative of research in the area, whereby, in most of the comparisons, both students with disabilities and typical developing students performed better in the non-accommodated condition than in the accommodated condition. It was even more surprising that students considered to be “low readers” followed this trend. There was no connection between performance on reading and basic math skills tests and the need for oral administration accommodations. As well, teachers were not able to predict which students would benefit from the oral administration accommodation as teacher ratings of student need for assessment accommodations only coincided with actual student performance approximately one-half of the time. The authors recognized that one of the major limitations of their study was the elimination of students who did not experience at least one-half a standard deviation change in assessment score between the assessment conditions. This effectively reduced, by one-half, the total number of students accounted for in the analyses of the assessment accommodation condition. It also reduced the number of teacher ratings by one-half, potentially eliminating many correct recommendations. Helwig and Tindal also noted that it might have been beneficial for the students participating in the study to have practice in using the accommodation prior to the testing situation.

Zurcher and Bryant (2001) provide an example of a primary study on the effectiveness of testing accommodations using a non-experimental design, albeit not an *ex post facto* design. Zurcher and Bryant examined the comparability and criterion validity of test scores for college-aged students with disabilities, specifically learning disabilities, and typically developing college students serving as the control group, under accommodated and non-accommodated conditions. Thirty undergraduate volunteers from three different colleges in southwestern Texas, 15 students with disabilities and 15 students with typical development, were selected to participate in the study. Students with disabilities selected to participate had to be eligible to take, but had not yet taken, the Miller Analogies Test under accommodated conditions: extended-time or oral administration using an audiocassette, reader and/or scribe. Using a counter-balanced design, the test was split into two halves and each student, a student with disabilities matched with a typically developing student, took one-half of the assessment using a student-specific accommodation and the other half of the assessment without any accommodation. Although typically developing students did not display a significant test score gain under accommodated conditions, results did not support the test interaction hypothesis (Sireci et al., 2003; Sireci et al., 2005) as their matched counterparts, students with disabilities, also did not display a significant gain under accommodated conditions. The authors noted several methodological limitations including small sample size, relatively short half-tests that may not have captured the potency of the accommodation effect, and lack of random assignment and matching which made across group comparisons difficult. For example, the GPA for students with disabilities was 2.72, while the GPA for their typically developing peers was 3.27.

Syntheses of the literature on the effectiveness of accommodations.

Several syntheses of the literature on the effectiveness of test accommodations for students with disabilities exist, most looking at testing accommodations after the implementation of NCLB (2001). Starting in 2002, NCEO began a review of primary studies in this area, generally providing three-year snapshots, starting with 1999 to 2000, of research on the effects of test accommodations.

Tindal and Fuchs (2000) conducted one of the first synthesis of research literature on the effectiveness of testing accommodations. They were seeking to provide personnel in school districts and state departments of education with a "comprehensive synthesis of the research literature on the effects of test accommodations on students with disabilities" (p. 16). In an effort to summarize research on changes to test administration over the preceding decade they identified 114 studies on more than 20 different accommodations, including research on test accommodations, test modifications, and the use of alternate assessments. Tindal and Fuchs categorized the research they reviewed into the three approaches: descriptive, comparative, and experimental. Additionally, the research studies were synthesized and organized according to types of test changes, generally assessment accommodations, based on a taxonomy proposed by NCEO. The research reviewed was grouped according to changes in schedule, presentation, test directions, use of assistive devices/supports, and test setting.

While the authors concluded that research on assessment accommodations was in its infancy, as most research at that point was usually not generalizable and needed to be interpreted with caution, there were consistent significant effects for moderately to significantly disabled preschoolers taking tests in the presence of familiar examiners. As well, "…making changes in the way tests are presented had a positive impact on student performance although the results have not always been differential for students with disabilities versus those without disabilities" with the "most clear and positive finding … to be in the use of large print or Braille and in the use of read aloud of math problems both of which appear differentially effective" (Tindal & Fuchs, 2000, p. 58). Tindal and Fuchs further suggested research on assessment accommodations (i) use experimental rather than descriptive or comparative designs and (ii) be studied in the context of validity and not necessarily in the context of population, such as students with disabilities or English language learners.

Thompson, Blount, and Thurlow (2002), in an NCEO technical report, extended the work of Tindal and Fuchs (2000), reviewing 46 empirical studies published from 1999 through 2001, to provide evidence regarding whether the use of certain assessment accommodations (i) threatened test validity or score comparability and (ii) were useful for individual students as "[t]he enactment of the No Child Left Behind Act of 2001 [brought] urgency" (p. 5) to research questions focusing on assessment accommodations. The authors believe that "[o]ne of the most viable ways to increase the participation of [students with disabilities] in assessments is through the use of accommodations" (Thompson et al., p. 8), participation that was mandated in NCLB 2001. Components of research summarized in the technical report included

> type of assessment, content area assessed, number of research participants, types of disabilities included in the sample, grade-level of the participants, research design, research findings, limitations of the study, and recommendations for future research (p. 9).

Thompson et al. (2002) noted a dramatic increase in the number of research studies on test accommodations, with 58 published in the nine-year span from 1990 through 1998, as compared to 46 published from 1999 though 2001. The two most common purposes for studying assessment accommodations were the investigation of differential boost, or test interaction hypothesis, where students with disabilities had greater test score gains than their typically developing peers and the investigation of assessment accommodations on test score validity. Criterion-referenced tests used for state accountability were the most common types of tests examined, in 21 studies, with norm-referenced or other standardized tests following closely behind, in 17 studies. Almost one-half of all tests under investigation were mathematics tests, while approximately one-third were reading or language arts tests. The number of participants in the studies under investigation ranged from three to almost 21,000, with the majority of studies looking at elementary school students. Twenty-seven of the studies documented participants' disabilities, with the two most common types of disabilities being learning and cognitive disabilities. Researchers in 21 of the 46 research studies reviewed identified limitations for their studies with the three most common limitations cited being "unknown variations among students included in the study, sample sizes too small to provide adequate statistical support, and nonstandard administration of the accommodations across proctors and schools" (Thompson et al., p. 6).

With respect to assessment accommodations, Thompson et al. (2002) noted that

> three accommodations showed a positive effect on student test scores...: computer administration [four of seven studies], oral presentation [six of seven studies], and extended time [four of seven studies]. However, additional studies on each of these accommodations also found no significant effect on scores or alterations in item comparability (p. 23).

Thompson et al. (2002) suggested that research on assessment accommodations lacked clarity in the (i) definitions of the constructs tested and (ii) accommodations needed by individual students. They also suggested that researchers explore students perceptions of desirability and usefulness of the accommodations provided, as they are the primary consumers of assessment accommodations. Further, they believe "[m]ore rigorous research, using designs comparing scores and interactions between the presence and absence of a disability are needed in the future" (p. 23).

Bolt and Thurlow (2004) identified and reviewed 36 studies on five of the most frequently mentioned accommodations for research conducted between 1990 and 2002. They selected studies on dictated response (k = 16), large print (k = 4), Braille (k = 2), extended-time (k = 22), and use of a sign language interpreter (k = 2) based on the 1999 NCEO report on state accommodation policies. Studies were selected based on the following four criteria:

1. The study was conducted or published after 1990.
2. The study focused on the effects of accommodations for students with disabilities in kindergarten through 12th grade.
3. The study examined the effects of accommodations on achievement or college entrance tests.
4. The study design allowed for the analysis of the effects of single accommodations, as opposed to the effects of accommodation packages.

> Of all the studies investigated, 17 used traditional experimental methodologies, 4 of which involved individualized assignment of students to accommodation packages. Comparative methodologies were used in 13 studies; 5 studies were descriptive, … and the remaining study was a meta-analysis (Bolt & Thurlow, 2004, p. 145).

The authors also examined the different approaches used to examine assessment accommodations in the research studies; differential boost studies (interaction of the disability status and accommodation condition), boost studies (accommodation increased test scores), studies of measurement comparability of the test (examination of factor structure and/or DIF in accommodated and unaccommodated conditions), and comparative studies (comparison of students with disabilities' "accommodated" assessment scores to "non-accommodated" assessment scores of students with or without disabilities).

Bolt and Thurlow (2004) found mixed results for the three of the five accommodations under review. Studies looking at dictated response, large print, and extended time produced supportive and non-supportive results for each of these assessment accommodations. It should be noted that much of the research indicated that "dictated response" is an effective accommodation and boosts the test scores of students with disabilities, findings similar to Chui and Pearson (1999). However, some researchers point out that this may result in implausibly high scores for this population. As very little research was found for Braille and use of an interpreter for instructions, little could be concluded about the use of these assessment accommodations. The authors discussed several issues with the studies they reviewed including, providing test accommodations for students who have a clear need for a specific accommodation; poor student selection (e.g., selecting students with disabilities who do not need accommodations); more than adequate time for extended time studies such that the research condition is not mimicking the less-than-adequate time provided in the actual testing situation; examining alternative types of extended time such as more frequent breaks; and ensuring students with disabilities and typically developing peers participating in the research condition are comfortable with and have used the assessment accommodation under investigation.

Tindal and Ketterlin-Geller (2004) reviewed research examining the effects of assessment accommodations on large-scale tests of mathematics, expressly mathematics tests with specific relevance for the National Assessment of Educational Progress (NAEP). Specific accommodations reviewed included assessment in small group settings, extended-time, use of calculators, read-aloud, and multiple accommodations (also called administration accommodation packages). The authors noted that NAEP did not allow for the use of assessment accommodations until 2002, thus prior results did not include a representative sample of students with disabilities.

Tindal and Ketterlin-Geller (2004) identified all published literature on large-scale mathematics assessments, finding a total of 28 studies published prior to 2000 and 14 studies published between 2000 and 2002. Unlike other authors of syntheses in this area, they were not specifically interested in the different study approaches of boost, differential boost, measurement comparability, or comparison of accommodated and non-accommodated test scores. They found results of the research they reviewed, generally based on the different approaches, to be tentative with conflicting overall test results. They alleged that the "one consistent finding … beginning to emerge … is the interaction of the item with specific skills of individuals" (p. 13), leading them to state that "[c]onstruct-irrelevant variance (unintended influence of skills and knowledge that are not part of the construct being measured) is item specific" (p. 8) such that studies on assessment accommodations consider using (i) universal design in item development, (ii) organize tests into sections in an effort to quarantine construct-irrelevant variance by allowing accommodations on sections where it does not interfere with the measurement of the construct under consideration, and (iii) use computer adaptive testing as the presentation of items is based on item characteristic curves, distribution on an ability scale, and the "item's target construct relative to an access skill" (p. 13). The authors noted that the latter is still under development and was not available for general use in 2003.

Johnstone, Altman, Thurlow, and Thompson (2006), in a continuation of the work of Tindal and Fuchs (2000) and Thompson et al. (2002), reviewed recent research on the effects of assessment accommodations for students with disabilities on large-scale assessments. Such research and research syntheses are needed

> [a]s states and school districts strive to meet the goals for adequate yearly progress required by NCLB, [given that] the use of individual accommodations continues to be scrutinized for effectiveness, threats to test validity, and score comparability (Johnstone et al., 2006, p. iii).

Johnstone et al. (2006) summarized information and findings from 49 empirical studies conducted between 2002 and 2004. Research examined involved 1 – 100 participants, 100 – 1,000 participants, or over 1,000 participants from multiple age categories being tested, generally on norm-referenced or criterion-referenced mathematics or reading/language arts large-scaled assessments. Subjects targeted for the research under review fell under the learning disability category more often than any other disability category. As with the Thompson et al. (2002) synthesis, the components of research summarized included the type of assessment, content area assessed, number of research participants, types of disabilities included in the sample, the participant grade-level, research findings, limitations of the study, and recommendations for future research. The authors extended the components summarized to include research purpose, type of accommodation, and percentage of sample that were students with disabilities.

There were two primary purposes for the studies reviewed, that of examination of the effect of assessment accommodations on test scores (k = 23) and the effects of assessment accommodations on test score validity (k = 13). Researchers used a variety of research methods, with the two most common methods being experimental or quasi-experimental in nature (k = 21) and reviews of/research using extant data (k = 17). Two studies conducted during this timeframe were considered to be meta-analyses; however, upon further examination these studies would not be considered "formal" meta-analyses. Fifteen different types of accommodations found were grouped according to presentation (k = 21), timing/scheduling (k = 8), response (k = 2), technological aids (k = 2), and multiple accommodations (k = 11). When viewing the 49 studies the authors did not find any common themes. They cited this lack of consistency in research results as an indicator of the need for further research in this area.

Johnstone et al. (2006) found the limitations most frequently mentioned by the researchers were noting that studies were too narrow in scope, involved a small sample size, or had confounding factors. Echoing the research limitations found by Thompson et al. (2000), the authors pointed to the need for clearer definitions of the constructs tested and examination of student perception of the desirability and usefulness of the accommodations they were provided. Additionally, the authors pointed to the need to study the institutional factors affecting accommodations judgment; how schools, districts, and states decide which assessment accommodations are allowable and which are not.

Zenisky and Sireci (2007) provided a further secondary analysis of the research, reviewing 32 published studies on assessment accommodation research conducted between 2005 and 2006 with all but five of the studies published in refereed journals.

Research conducted with the most frequency during this timeframe focused on (i) the empirical evaluation of test score comparability for tests administered with and without accommodations and (ii) descriptive studies of current accommodations practices for students with disabilities and their typically developing peers. As well, the research examined generally looked at academic measures, criterion-referenced tests, miscellaneous cognitive and intelligence measures, and instruments developed for research purposes for content in mathematics and reading, with state criterion-referenced assessment often used for NCLB purposes as the most commonly used data collection instruments. Participants in these studies ranged from nine to 107,000 with most studies collecting data on 100 to 300 participants. As well, participants were from drawn from various grade levels, K – 12, and included college/university students. One study used participants in an adult education setting. As with other synthesis studies in this area, there was a wide range of disabilities included in the research; learning disabilities being the most commonly represented disability. However, it should be acknowledged that ten studies did not provide information on specific disability for participants. While most studies examined assessment accommodations that fell under presentation and timing/scheduling categories, a few studies looked at accommodations falling under setting categories. This narrowing of assessment accommodations to two primary categories is in contrast to the four categories reported in the summaries of accommodations by Johnstone et al. (2006) and Thompson et al. (2002). It should be noted that timing/scheduling accommodations, specifically extended time, was, again, one of the most-studied accommodations. Other frequently studied accommodations included oral accommodations and computerized administration. Most of the studies conducted used non-experimental (k = 14), followed by quasi-experimental (k = 11), and experimental (k = 7) research designs. Of the empirical research, over 50% used primary data collection rather than existing data sets for their analyses. Some of the research studies focused on assessing the need for accommodations as well as the selection and implementation of accommodations, frequently using surveys to collect this information.

Zenisky and Sireci (2007) noted that empirically tested oral presentation, timing (extended time), and accommodations for computerized assessment were often found to have positive effects on test scores, with some studies reporting no effects for assessment accommodations. By and large, timing accommodations yielded positive effects on test scores. No studies reported negative effects on test scores for testing accommodations.

Limitations most frequently noted by the investigators represented in this summary of research were small sample size, lack of diversity in the sample, and issues with operationalization and implementation of the assessment accommodations. As well, some researchers cited test or testing context; for example, number of items on the measure used; and unexpected results as study limitations.

Zenisky and Sireci (2007) cited a number of promising avenues for future research including "varying or improving on research methods with respect to testing for the effects of specific accommodations and improving test development practices to reduce the need for accommodations" (p. iv). Specific directions for future studies on assessment accommodations were "(1) further study of extended time, (2) computers and assistive technology as accommodations, (3) the role of teachers, and (4) the interaction hypothesis" (p. 15). The authors note that directions such as these are needed to further refine research in the area of assessment accommodations and expand our knowledge of how best to obtain valid measures of student performance since “variations across operational definitions, tests, populations, settings, and contexts still curb all but the most general policy implications” (p. 17). With the high-stakes consequences of decisions made based on test score interpretation, particularly in light of NCLB (2001), general policy implications are no longer adequate.

Thurlow (2007), in a paper presented at the American Education Research Association conference, summarized the findings of syntheses on the effectiveness of assessment accommodations by Tindal and Fuchs (1999), Thompson et al. (2002), Johnstone et al. (2006), and Zenisky and Sireci (2007, in press at the time of her presentation). Thurlow noted the increase in the amount of research conducted, beginning in 1990, in this area. Aggregating across the syntheses, Thurlow saw a significant amount of research conducted using oral administration and extended-time accommodations. The author found that the results from studies on oral administration to be

> complicated by the inclusion of different groups of students, the study of different content areas, the use of different media for presenting the accommodation (person vs. video vs. audio tape), and by other refinements (such as the length of the passage to be read) (p. 6),

with results showing positive effects for students with disabilities, positive effects for students with disabilities and typically developing peers, or no effects. Research focusing on extended time accommodations was more consistent, generally showing positive effects for students with disabilities. Thurlow found that the most commonly allowed assessment accommodations in assessment programs were not necessarily the most frequently studied accommodations; the most commonly allowed assessment accommodations being large print, individualized administration, small group administration, magnification, Braille, use of a separate room, writing directly in the test booklet, and extended time (time beneficial to the test taker).

Thurlow (2007) observed an expansion in the number of states providing assessment accommodation policies and guidelines, an increase in the complexity of the accommodations, and increased length in the documentation regarding accommodations. As well, Thurlow found that states were also becoming concerned with the "[c]larity about the effects of … test changes on the validity of test results" (p. 10). States were also trying to increase the validity of accommodations such as oral administration, scribe, and sign language interpretation, which include a human component, referred to as "access assistants" by NCEO, by providing written guidelines for most, albeit not all, access assistants.

Thurlow (2007) recommended aligning research with existing state policies on accommodations allowed *without* restrictions and accommodations allowed *with* restrictions, specifically those allowed with restrictions; oral administration, use of calculator, use of scribe, and extended time; as they are the most controversial of the testing accommodations. With a growing number of states implementing assessment accommodation policies and guidelines, Thurlow indicated that this type of alignment was especially relevant when considering how best to affect policy on testing accommodations, noting that most states do not have the resources to conduct research on assessment accommodations which have impact on specific state accommodation policies.

Cormier, Altman, Shyyan, & Thurlow (2010) summarized the results of 40 empirical studies conducted between 2007 and 2008. Most of the studies focused on either (i) the effects of accommodations on test scores of students with disabilities, k = 13, or (ii) a comparison of test scores for unaccommodated versus accommodated assessment conditions, k = 11; i.e. boost or differential boost studies. Most studies conducted during this time examined math or reading content and research participants were enrolled in the K – 12 educational system. A majority of studies had large, more than 300 participants, sample sizes. As with previous syntheses of the research in this area; e.g., research examining the effects of read-aloud or extended-time conditions, results from the aggregate research was mixed.

Cormier et al. (2010) found that research on extended time accommodations was declining, while research investigating accommodation packages was increasing. They noted that "[a]lthough this accommodation was studied frequently in the past, it has lost its place as an accommodation in many states because of a move to untimed tests" (p. 18). While investigation of accommodation packages is valuable, others have expressed concern that empirically effective accommodation packages may include extraneous accommodations that do not add to the efficacy of the package (Elliott, Kratochwill, & McKevitt, 2001).

Synthesis studies of the effectiveness of accommodations.

The most frequently cited large-scale secondary analysis of the effectiveness of assessment accommodations was conducted by Chiu and Pearson in 1999. Using meta-analytic techniques, Chui and Pearson examined 30 research studies searching for empirical evidence to support the hypothesis that test accommodations would increase the test scores of students with disabilities and English language learners relative to a situation where no accommodations were provided and relative to typically developing peers. Additionally,

> … to determine if the accommodations under investigation ‘matched’ the needs of the target students, [they] checked to ensure that the included research studies had explicitly described the nature of the target students and had provided narrative descriptions for the accommodations used (p. 6).

For the studies they examined, Chui and Pearson found the most frequently studied accommodation was timing of the test, or extended time (47%), with test setting (2%) and response format (2%) were being the least frequently studied. Students with learning disabilities (61%) were the most commonly studied subgroup, with timing of the test being the most frequently studied accommodation for this subgroup.

Chui and Pearson (1999) noted that

> … the significant *Q* test for homogeneity of variance revealed that the variations among the accommodation effects were large, implying that using the mean effect alone could be misleading because it would fail to portray the diversity of accommodation effects (p. 15).

To counter this issue Chiu and Pearson only used effect sizes where both the target groups, students with disabilities and English language learners, and general education populations were included; i.e., equivalent groups or test-retest designs. The recomputed mean effect size was 0.11 using Hedges and Olkin’s (1985) procedure to “examine the relationship between the characteristics of the studies and outcome measures” (p. 15).

They found test accommodations have a small, positive effect on the target students under analysis. Evidence pointed to an overall weighted mean effect of 0.16 for students with disabilities and English language learners, providing them with a slight advantage over their typically developing “peers,” with an overall weighted mean effect of 0.06 (Chui & Pearson, 1999). They noted that, for the types of accommodations examined, presentation format was the only accommodation with a homogenous mean relative effect, while all other accommodations exhibited heterogeneous effects. However, they suggested that their results be interpreted with caution, as there were a variety of accommodations, statuses for students, and implementations of accommodations. Further, some confidence intervals for effect sizes were extremely wide and could envelop the mean effect and the relative mean effect for the type of accommodation, thus leading them to state that there was no difference in the efficacy of the accommodation for the target population relative to the general education population. Chui and Pearson concluded that students with disabilities and English language learners could increase their test scores on standardized tests with appropriate test accommodations.

Specific issues with this meta-analysis are related to combining English language learners and students with disabilities populations to study accommodation effects. While many studies provide information on the use of test accommodations with these groups, recent considerations in the field indicate that effective accommodations for students with disabilities, for the most part, are different from those found to be efficacious for English language learners (Enriquez, 2008). As well, this meta-analysis is over ten years old and was conducted prior to NCLB, which mandated testing for AYP and school accountability. There has been rapid growth in the testing industry, with much more research into testing accommodations, since Chiu and Pearson (1999) conducted their meta-analysis, the only meta-analysis to date, on this particular topic.

It must be noted that two further meta-analyses examining the effects of assessment accommodations on students with disabilities were conducted within the past five years, but were limited in their scope. Elbaum (2007), as part of a larger study on the efficacy of oral test accommodations for students with disabilities on math assessments, used meta-analysis to examine existing research on read-aloud accommodations for students with disabilities. Gregg and Nelson (2012) used meta-analysis to examine the use of extra time for students with learning disabilities transitioning from high school to college.

Elbaum (2007) focused on studies using read-aloud accommodations on math assessments that may, or may not, have been considered high-stakes assessments. Elbaum calculated separate mean effect size differences, *d*, for studies examining (i) elementary school students and (ii) secondary school students. Findings indicated that there was a small effect for elementary school students, $d = 0.20$, and a very small effect, $d = 0.12$ for secondary school students. Elbaum concluded that there was "… a statistically significant association of students' school level with the difference in effect sizes for students with and without [learning disabilities]" (p. 225). Further, Elbaum found

> … the accommodation boost for elementary students is clearly of greater magnitude for students with [learning disabilities]than it is for students without [learning disabilities], the impact on secondary students shows greater benefits for students without disabilities (p. 227).

Gregg and Nelson (2012) examined the use of extra time for students with learning disabilities, specifically those students transitioning from high school to college. Using the results from nine studies, their meta-analyses focused on three comparisons: scores of students with learning disabilities in accommodated conditions to typically developing peers in non-accommodated conditions, scores of students with learning disabilities to typically developing peers in accommodated conditions, and scores of students with learning disabilities to typically developing peers in non-accommodated conditions. Using *Comprehensive Meta-Analysis V.2*, they estimated Cohen's *d* effect sizes. They found that typically achieving students in unaccommodated conditions outperform students with disabilities using an extended time accommodation (*d* = -0.41). They were unable to provide similar information for their other two comparisons as "[t]he results … underscore the lack of research available to make conclusions about the comparability of scores for transitioning students with [learning disabilities] taking tests with extended time to their normally achieving peers" (p. 136).

Test accommodation interaction hypothesis and differential boost.

Considered a well-controlled research approach, the test interaction hypothesis involves testing the interaction between testing condition (accommodated and unaccommodated conditions) and disability status (students with and without disabilities). The test interaction hypothesis postulates that appropriate accommodations will boost the scores of students with disabilities more than their typically developing peers (Bolt & Thurlow, 2004; Sireci et al., 2003; Sireci et al., 2005). This "[d]ifferential impact on students with and without disabilities provides evidence that the accommodation removes a barrier based on disability" (Macarthur & Cavalier, 2004, p. 55) and effectively removes construct-irrelevant variability (Messick, 1995). "Boost studies;" employing a within-subjects or a random-independent-groups (across subjects) design and having a control group that does not receive accommodations to determine whether or not students with disabilities score significantly higher under accommodated conditions (Bolt & Thurlow, 2004); do not test the significance of an interaction between disability status and testing condition as is found with research work using the test accommodation interaction hypothesis. Research studies exploring how test scores for accommodated students with disabilities compare to test scores of other students with disabilities or those of typically developing students, called “comparative studies”, also do not test the significance of an interaction between disability and testing condition (Bolt & Thurlow, 2004).

The interaction hypothesis also referred to as the “maximum potential thesis,” posited by Zuriff (2000), states that “students without disabilities would not benefit from extra examination time because they are already operating at their maximum potential under timed conditions” (p. 101). A similar theory, differential boost (Fuchs & Fuchs, 1999) posits that both students with disabilities and their typically developing peers will benefit from testing accommodations. However, students with disabilities are expected to benefit differentially more than their typically developing peers. The test accommodation interaction hypothesis, maximum potential thesis, and differential boost theory are used to justify the use of test accommodations for students with disabilities as (i) test scores of students with disabilities are improved relative to the score they would receive under standard administrative conditions, (ii) typically developing students’ test scores will not improve if they take the test using the same test accommodations, and (iii) students with disabilities and typically developing peers, the student factor, interacts with the administration condition (standard or accommodated administration).

In 2000, Zuriff examined five studies that utilized the maximum potential thesis in their design, testing the interaction between assessment condition and disability status. These studies investigated the use of extra examination time for college students with learning disabilities versus their typically developing peers. All studies cited used a common measure, the Nelson-Denny Reading Test, considered reliable, related to scholastic achievement, and normed through the fourth year of college. The author found support, albeit very weak empirical support, for the maximum potential thesis. Contradictory evidence for the maximum potential thesis came from typically developing students seeing test score gains, albeit not as large as students with disabilities, in untimed assessment conditions. Zuriff recommended examining individual differences under timed and untimed conditions for all students participating in research studies looking at the maximum potential thesis, as this would allow for a better understanding of patterns in the data that is not afforded when only using group means.

Sireci et al. (2003) reviewed 150 studies concerned with the effects of test accommodations, critiquing all studies in light of the “interaction hypothesis [such] that test accommodations should improve the test scores for targeted groups, but should not improve the scores of examinees for whom the accommodations are not intended” (p. 2). Of the 150 research studies, 46 examined the effects of test accommodations for students with disabilities and English language learners. Of the 46 studies, only 38 studies empirically looked at data from accommodated tests with 21 using an experimental design: 12 for students with disabilities and 8 for English language learners. Less than one-half of the research studies examined were found in peer-reviewed journals. The authors’ critique was structured using three primary criteria: (i) group that was to be helped by the assessment accommodation, that is students with disabilities or English language learners, (ii) type of accommodation examined; for example, presentation accommodations, timing/scheduling accommodations, and response accommodations, and (iii) type of research design, that is literature reviews, experimental studies, and non-experimental studies. The 38 studies reviewed spanned several subject areas and multiple grades. At the time of publication, 26 studies relating to assessment accommodations had been critically reviewed.

Sireci et al. (2003) concluded that the vast majority of studies showed improvements for all students taking accommodated tests, with the "accommodation of extended time improv[ing] the performance of students with disabilities more than it improved the performance of students without disabilities" (p. 2). They noted that "there are no unequivocal conclusions that can be drawn regarding the effects, in general, of accommodations on students' test performance" (p. 48). Sireci et al. felt that the interaction hypothesis as typically stated was on "shaky ground" (p. 48) and proposed a revision to the hypothesis, namely differential boost (Fuchs & Fuchs, 1999). Differential boost allows that typically developing students may benefit from assessment accommodations, though not to the same extent as their peers with disabilities. With respect to extended time, Sireci et al. (2003) found "gains for students without disabilities, although the gains for students with disabilities were significantly greater" (p. 63). Research exploring the use of oral presentation accommodations was unclear, with half of the studies finding positive effects, while the remaining studies saw either no effects or similar effects for students with disabilities and their typically developing peers.

Issues with the studies reviewed included the heterogeneous nature of both the students (large within-group diversity) and the assessment accommodations, and diversity in the creation and implementation of accommodations. Although students with disabilities were heterogeneous with respect to type of disability, they were generally ethnically homogeneous groups of students, thus results from the studies under consideration cannot be generalized to minority students. As well, much of the research was undertaken in Los Angeles, California, making generalizability to other locales contentious. Additionally, virtually all of the research was conducted on elementary school students, making generalization to other levels impossible. Further, effect sizes were not reported in most studies. While effect sizes could be estimated for some of the studies, this was not possible for all studies under review.

Sireci et al. (2005), in a later secondary study of the test accommodation interaction hypothesis were, again, seeking empirical support for the interaction hypothesis, whereby "…test accommodations lead to improved test scores for students with disabilities relative to their non-disabled peers" (p. 459). The authors reviewed several recent empirical studies that focused on the effects of accommodations on test performance, particularly the test performance of students with disabilities. Of the studies they reviewed, they selected 28 and categorized them based on the type of test accommodation; extended time, oral (read-aloud) presentation, or multiple accommodations; and research design; experimental, quasi-experimental, and non-experimental using an *ex post facto* comparison of students taking a standard version of the test and students taking an accommodated version of the same test.

Of the studies they reviewed, Sireci et al. (2005) found that the most common accommodations examined were oral administration, at 39%, and extra time, at 24%. Studies investigating oral administration were often accompanied by extra time as a second accommodation, thus making it difficult, if not impossible, to decouple the effects of the accommodations. As well, a variety of different accommodations was analyzed within a single study for some of the studies being reviewed. Most of the studies focused on students in third through eighth grades taking tests in mathematics, reading, and science.

For research relating to extended time, Sireci et al. (2005) found that five of eight studies provided qualified support for the interaction hypothesis. For the most part, the results indicated that students with disabilities exhibit greater score gains than typically developing peers. However, results from two of the eight studies did not display any gains. Five of the ten studies concentrating on oral accommodations provided partial support for the interaction hypothesis. The research literature substantiated findings that a more valid interpretation of mathematics achievement was possible when students with disabilities received oral; e.g., read-aloud, accommodations. This could not be said for other subject areas. For studies relating to multiple accommodations, all seven of the studies reviewed provided support, at some level, for the interaction hypothesis. Four of the seven studies using experimental designs also demonstrated results that were consistent with the interaction hypothesis.

While two fairly consistent findings were discussed, those of extended time tending to improve the performance of all students, albeit students with disabilities showing the greatest gains, supporting a differential boost interpretation, and oral accommodations on mathematics tests improving performance for some students with disabilities, consistent conclusions could not be drawn across the studies. With the wide variety of accommodations, the differences between accommodation implementation, and the heterogeneity of students receiving accommodations, heterogeneity being found even within the students with disabilities groups, it was not surprising that there were a lack of consistent inferences.

Sireci et al. (2005) concluded that the vast majority of research explored showed that all student groups had test score gains under accommodated conditions, with students with disabilities displaying the largest test score gains. As with the Sireci et al. (2003) research review, the authors felt that qualification of the interaction hypothesis, with greater gains experienced by students with disabilities implying that the standardized testing conditions are too stringent for all students and not that the test accommodations are unfair, better explained their findings, particularly their findings regarding the use of extended time. Additionally, their findings were consistent with the concept of differential boost put forth by Fuchs and Fuchs (1999), whereby "an accommodation …. increases the performance of students with disabilities more than it increases the scores of students without disabilities" (p. 24). Further, Sireci et al. (2003) concluded (i) most educational tests are speeded, (ii) oral accommodations on math tests produce gains for students with disabilities, however, the same cannot be said for tests in other content areas, and (iii) students with disabilities need extra time to demonstrate their true knowledge, skills, and abilities.

Sireci et al. (2005) noted several issues with the studies they reviewed. These issues included the use of small, ethnically homogenous groups of students with disabilities whose results could not be generalized to minority students with disabilities and almost all the studies focused on elementary grades. They noted that only one of the experimental studies looked at test accommodations for secondary school students. They believed this was a tremendous issue, as there are a growing number of states implementing high school graduation examinations. The growing number of graduation examinations, coupled with a dearth of information on the potential usefulness of assessment accommodations and/or the interaction effect of accommodations on such examinations for this group, was seen as a major limitation.

Issues with this review that could not be controlled for were the great diversity (i) within the students with disabilities group, (ii) in the way the test accommodations were created, and (iii) in the way the test accommodations were implemented. Such diversity makes it very difficult to make unequivocal statements about the research findings.

Gaps in the literature.

Concerns that students with disabilities are tested fairly when examinations are used for promotion and high-stakes decisions abound and are discussed in non-academic and academic circles alike, with discussion on this topic commonly found in mainstream newspapers such as the New York Times.

> [Q]uestion[s] of how far to accommodate students with learning disabilities on college entrance tests like the SAT has become a familiar one [in mainstream society], as requests for special accommodations proliferate, especially from affluent white families (Lewin, 2002).

Information that had been the sole purview of educational policymakers and researchers is becoming part of the mainstream ethos. Delineation of educational legal issues, particularly those relating to issues of equity and access, have become commonplace in the news. Articles with information such as the following have become part of the mainstream lexicon:

> Judge Charles R. Breyer of Federal District Court [of California] ruled that students with learning disabilities had the right to special treatment, through different assessment methods or accommodations like the use of a calculator or the chance to have test questions read aloud (Lewin, 2002).

With such judgments coming to the fore, it is imperative we become better able to make sound decisions based on strong evidence.

With existing educational legislation regarding students with disabilities and assessment accommodations, states are tasked with creating and implementing assessment accommodations. However, there is an "… amazing lack of agreement across states in how to go about making participation and accommodation decisions, and which accommodations are acceptable" (Thurlow et al., 2000, p. 162). Many researchers have noted that states continue to make changes to their assessment accommodations policies despite the lack of a solid research base on accommodations (Bolt & Thurlow, 2004; Sireci et al., 2003; Sireci et al., 2005; Thompson et al., 2002; Thurlow & Bolt, 2001). Secondary studies point to a lack of definitive findings, providing suggestions on how this might be remedied (Bolt & Thurlow, 2004; Johnstone et al., 2006; Thompson et al., 2002; Tindal & Fuchs, 2000; Tindal & Ketterlin-Geller, 2004; Zenisky & Sireci, 2007). Educators and policy-makers need more information regarding the effectiveness of testing accommodations for students with disabilities and whether they remove or reduce presentation, response, setting, and timing/scheduling barriers in assessment. It has also been noted that much of the research does not directly address the use of accommodations that are frequently allowed under state policy (Bolt & Thurlow, 2004; Tindal & Fuchs, 1999).

There appears to be a lack of experimental research and empirical evidence when it comes to understanding which assessment accommodations are efficacious. Researchers and those examining the existing literature have noticed that very few studies examining assessment accommodations use experimental designs (Bolt & Thurlow, 2004; Tindal & Fuchs, 1999). Ysseldyke et al. (1998) noted

> … research on accommodations needs to be *experimental* in nature, and designed to address the perception that the use of accommodations may invalidate a test. Experimental research goes beyond simply examining the performance of students who use accommodations and comparing it to the performance of students who do not use accommodations by providing appropriate controls (p. 31).

Additionally, several researchers indicated that the empirical research base regarding the effects of specific testing accommodations is very limited (Bolt & Thurlow, 2004; Fuchs et al, 2000a). Such research helps us answer questions about which accommodations would be beneficial for specific groups of students with disabilities, and for which situations these accommodations would be the most beneficial, thus providing more accurate assessments of students with disabilities. As Ysseldyke et al. (1998) noted

> [s]pecific issues arise for each disability type, or combination of disabilities, and for each specific accommodation [with] considerably more rhetoric and opinion than sound empirical evidence about the validity of specific accommodations. The knowledge base about the effects of accommodations is not adequate to address many practical, everyday questions, nor is it in a form that is readily accessible to or easily understood by personnel in states and districts (p. 21).

The existing research on assessment accommodations is spotty, with some types of accommodations being glossed over and some groups of students with disabilities being skipped over. Chui and Pearson (1999) noted a dearth of research in the areas of accommodations such as "assistive devices, combinations of accommodations, presentation formats, response formats, setting of tests, and radical accommodations" (p. 33), with learning disabled students receiving the most attention in the research literature. While this has slowly been changing, with studies looking at a larger variety of accommodations and students with disabilities, syntheses of the literature in this area have only considered three- to four-year slices of research work. As educational research can be very cyclical in nature, with different studies occurring in the same time frame overlapping in areas examined, trends for the different types of assessment accommodations, and students with disabilities groupings may be hidden.

The existing research in the area of assessment accommodations for students with disabilities is far from conclusive. Much of the research in this area, at best, remains equivocal and open for debate. There is very little agreement on which accommodations, or combinations of accommodations, allow students with disabilities to demonstrate what they know without providing an unfair advantage for these students. Long recognized in research syntheses and secondary studies, research on assessment accommodations, provide ambiguous information, as these syntheses highlight the contradictory findings for the research which was reviewed (Johnstone et al., 2006; Sireci et al., 2003). As well, "variations across operational definitions, tests, populations, settings, and contexts still curb all but the most general policy implications" (Zenisky & Sireci, 2007), such that "… more empirical study is warranted to further investigate the effects of testing accommodations for students with disabilities" (Bolt & Thurlow, 2004, p. 151).

As noted in 1999 by Chiu and Pearson, there has been enough research in the field of assessment accommodations and students with disabilities to make meta-analysis useful. Although much primary and secondary research on students with disabilities and testing accommodations has been conducted, there have been no meta-analyses of students with disabilities across all categories of assessment accommodations conducted since Chiu and Pearson's research in 1999. In the intervening years, well over 100 primary studies have been conducted. With the capacity to examine the convergence across studies objectively and systematically, and the use of a common metric, meta-analysis has the potential to fill in the gaps in the assessment accommodation literature, providing more definitive empirical answers to the hypotheses posed by research in this area. Zenisky and Sireci (2007) found that

> [g]reat diversity exists both with respect to the individuals requiring assessment accommodations and the range of accommodations available [and that] such diversity does not easily lend itself to consensus on policy for valid testing practice. The completion of more well-constructed meta-analyses of specific accommodations is one strategy that researchers should consider, in addition to further empirical study of specific accommodations with different—both heterogeneous and homogeneous—student populations (p. 17).

As well, Sireci and Pitoniak (2007) believe that meta-analysis, potentially based on state practices, would be useful at this point in time. While not overcoming all of the pitfalls of existing primary research in this area, using meta-analysis to aggregate and quantitatively analyze existing research will provide a more rigorous examination of the data collected to date. With the addition of meta-regression, providing a statistical means to delve deeper into possible explanations for variance, together with effect size findings provided through a meta-analysis of existing research studies, it is hoped that this research will fill some of the gaps discussed by those in the field.

Meta-regression.

Meta-regression extends regression analyses by examining multiple studies to model, estimate, and explain the variation among reported empirical results (Stanley, 2001). Meta-regression is used when heterogeneity in effect sizes is found or is believed to exist and "… aims to relate the size of the effect to one or more characteristics of the study involved" (Thompson & Higgins, 2002, p. 1559). Increasingly, "[m]eta-regression has become a commonly used tool for investigating whether study characteristics may explain heterogeneity of results among studies in a systematic review" (Higgins & Thompson, 2004, p. 1663).

There are a variety of meta-regression approaches. The regression model used may be linear or logistic with a single study as the observation or unit of analysis. In a simulation study comparing and contrasting meta-regression approaches which model heterogeneity, Morton, Adams, Suttorp, and Shekelle (2004) identified four meta-regression approaches: fixed-effects utilizing logistic regression, random-effects meta-regression, control rate meta-regression, and Bayesian hierarchical modeling. Further, Morton et al. identified and evaluated five meta-regression methods: fixed-effects with and without moderators; random-effects with and without moderators; and control rate meta-regression. They used the results of their simulation to provide meta-regression practitioners with a set of guidelines. Specifically, Morton et al. noted that results can be biased if important moderators were not incorporated at the person or study level, moderators that are aggregates of person-level rather than study-level characteristics can produce biased results, control rate (in health and medical studies) needs to be incorporated if it affects treatment, and bias can be reduced using a larger number of studies and a larger number of subjects with proper modeling.

There are several statistical issues with meta-regression. These include, but are not limited to, a small number of degrees of freedom in research that reviews a small number of studies and the use of highly collinear moderators. While there are several issues with this technique and many researchers call for more study of meta-regression (Higgins & Thompson, 2004; Stanley, 2001; Thompson & Higgins, 2002) it has the potential to explain differences between studies and can aid in understanding the causes of heterogeneity, a truly handy instrument in the meta-analyst's tool box.

Delimitations

Delimitations for this study relate to both the unit of analysis and the analytic techniques proposed.

In standardized, and other, assessments we need an accurate and adequate measure of student knowledge. This means we must endeavor to minimize construct-irrelevant variance, as well as provide methods to increase access to these assessments for students with disabilities. One of the goals for standardized assessment is to ensure, in part by providing empirical evidence, that test scores for all students are valid and comparable, regardless of population subgroup. As such, this study will be limited by the adequacy of the assessments used in the primary research studies under examination.

Sireci et al. (2003) have noted several limitations of the extant research. Limitations included focus on a "relatively small, and ethnically homogenous groups of students" (p. 65), with "… most of the studies focused on elementary school grades…" (p. 66), and "… virtually no *experimental* studies involved secondary students…" (p. 66). It is hoped that expanding the bandwidth of the studies to include primary studies for a longer time period, mid-1999 through mid-2011, will help circumvent these particular limitations.

Research design limitations for primary research in this area include poor and inconsistent classification of students with disabilities and their typically developing peers, absent or poor control groups, insufficient time for accommodations that require additional materials, and validity concerns due to a poor match between test content and curriculum.

Another potential limitation for this research relates to one of the subgroups of students with disabilities; students with learning disabilities. Students with learning disabilities comprise almost one-half of the population of students with disabilities (Tindal & Fuchs, 2000) and are a heterogeneous group. This makes logical analysis of assessment accommodations difficult. As well, it is difficult to conduct studies in the area of test accommodations as it is difficult to

> find... and recruit… sufficient numbers of students with disabilities and students without disabilities to participate in studies involving taking tests, particularly if the design requires them to take a test twice; under standard and accommodated conditions. The small numbers of students with disabilities in specific disability categories make it particularly hard to find sufficient numbers of different types of students with disabilities who are prepared to take a test in a specific subject area in a specific grade level (Scarpati, 2003 cited in Sireci et al., 2005, p. 487).

Several primary studies examining extended time used speeded tests; thus all students would be expected to show test score gains when given extra time. This makes results from these studies equivocal and a potential limitation for the present research.

A major limitation that cannot be overcome concerns lack of reporting of appropriate statistics; i.e., at a minimum means, standard deviations, and number of participants, thus studies that do not contain useable statistics cannot be included in the analyses. As well,

> [m]ost of the studies that focused on multiple accommodations were *ex post facto* studies that analyzed data from a large-scale assessment and broke out accommodated test administrations from non-accommodated administrations… [which] … typically do not use an experimental design… (Sireci et al., 2005, p. 475).

Additionally, there was incomplete reporting which resulted in low statistical power and questionable findings for some of the primary studies being considered for the meta-analysis. Due to the preceding issues there is the potential to lose a great number of research studies during the coding phase of this research.

Research is only useful insofar as we can generalize the findings from research on assessment accommodations to students in classrooms (Tindal & Fuchs, 2000). When coding the primary studies, appropriate sampling in the primary studies must be examined to ensure students are sampled appropriately. Primary studies that do not conformation to appropriate sampling procedures will not be included in the meta-analysis.

It must be noted that using meta-analysis does not allow us to examine *measurement comparability*; i.e., to see if internal characteristics are the same for accommodated and unaccommodated tests. This limitation cannot be avoided with meta-analytic techniques.

Definitions

A number of definitions specific to this study apply. Terms relating to students with disabilities and legislation regarding students with disabilities, assessments and accommodations, assessment of students with disabilities, organizations involved with students with disabilities and research regarding students with disabilities, as well as meta-analytic techniques are defined in the following section.

Terms specific to students with disabilities include the definition of student with disabilities, Individualized Education Plan, Least Restrictive Environment, and Free and Appropriate Public Education.

The thirteen legislative special education categories used to identify students with disabilities, delineated in IDEA (2004), are

> mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as 'emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities (Part A (SEC. 602) (3) (A) (i), 118 STAT.2652, 2004).

Individualized Education Plans (IEPs) are used to define an appropriate education, guide delivery of educational services and frame methods for evaluating outcomes for students with disabilities. IEPs "… must include a statement of the student's current levels of educational performance and a statement of measureable annual goals, including short-term objectives or benchmarks" (Tindal & Fuchs, 2000, p. 10).

The least restrictive environment (LRE) allows that,

> [t]o the maximum extent appropriate, children with disabilities... be educated with children who are not disabled, and... special classes, separate schooling, or other removal of children with disabilities from the regular educational environment should occur only when the nature or severity of the disability is such that education in regular classes with the use of supplementary aids and services cannot be achieved satisfactorily (Federal Register, 1999, (20 U.S.C. 1412(a)(5)(B))).

Section 504 of the Rehabilitation Act of 1973 defines a free and appropriate public education (FAPE) as school district provision of a "'free appropriate public education' … to each qualified person with a disability who is in the school district's jurisdiction, regardless of the nature or severity of the person's disability" (U.S. Department of Education, 2007, p. 1).

Terms specific to assessment and accommodation include the definition of test/assessment accommodation, high-stakes assessments, statewide assessment programs, partial participation, out-of-level testing, combination participation, assessment modification, and alternate assessments.

Test accommodation, or assessment accommodation, refer to accommodations providing support for students with disabilities involving adjustments to the assessment presentation, setting, timing or scheduling, or response and are generally dependant on the disability involved. Accommodations should not provide any advantages to individuals taking the test in question.

High-stakes assessments generally refer to assessment results tied to important decisions which may significantly impact the lives of students and educational professionals (Reschly, 1993). Statewide assessment programs, as part of the accountability structure for states since NCLB (2001), are considered to be high-stakes assessments.

Partial participation in assessment programs occurs when students take certain parts of the assessment, but are not required to take the entire assessment.

Out-of-level testing occurs when students take assessments designated for students in lower grades.

Combination participation occurs when students take different parts of different assessments from an entire assessment program. For example, students might take certain parts of state reading, writing, mathematics, and science assessments.

Test modification, assessment modification, or non-standard accommodations involve student use of modifications or accommodations that change the construct being measured, thus test scores for these students are considered invalid and student participation is not included in aggregated results for the assessment.

Alternate assessments are normally designed for a specific subgroup of students. These assessments are most frequently used to assess students having significant cognitive disabilities who would otherwise not be able to access the assessment, even with accommodations.

Terms specific to assessment of students with disabilities include access to assessment programs, inclusion in education, participation in assessment programs, and unwarranted exclusion.

Access to assessment programs; for example, state assessment programs, refers to the ability of all students to have an equal opportunity, or the right, to participate in the assessment program in order to demonstration their abilities in the area(s) being measured and receive benefits provided by the demonstration of their abilities (e.g., graduation from high school). It is expected that all students have access to assessment programs regardless of their social class, ethnicity, background or physical disabilities. Access to assessment programs for students with disabilities often requires bridging technologies such as accommodations, modifications, or alternate assessments and "deals specifically with removing barriers for student" (Baker, 2008, p. 24) and allows students with disabilities a way to demonstrate their skills and abilities.

Inclusion in education refers to the education of students with disabilities in the regular classroom for all, or nearly all, of the school day. Inclusion models do not allow for the education of students with disabilities in a separate school or classroom. Inclusion in assessment programs; for example, state assessment programs, refers to including students with disabilities in the assessment experience. Unlike access where students have the right to participate and be provided with the tools to participate, inclusion simply refers to being included in the process or program, including assessment programs.

Participation in assessment programs, such as statewide assessment programs, refers to students with disabilities taking part in the assessment process and having their results included in any reports generated from the assessment efforts; i.e., district accountability reports used as part of the AYP requirements for the federal government. Participation differs from access, as it is not mandated by law. Participation differs from inclusion in that, although students with disabilities may be included in programs, they may not be able to participate in the program and/or their results may not be included in the reports generated from the assessment program.

Unwarranted exclusion refers to the

> … directed or arranged non-participation in state or national assessment programs involving students for whom the assessment is appropriate to curriculum goals pursued in their educational programs and the receptive or expressive language demands of the assessment tasks are within the student's behavioral repertoire (Reschly, 1993, p. 46).

Organizations involved with students with disabilities, in legislative and/or research capacities include the National Center on Educational Outcomes, Council of Chief State School Officers, Council for Exceptional Children, and National Association of Directors of Special Education.

The National Center on Educational Outcomes (NCEO), founded in 1990, is tasked with working with federal and state agencies to assess educational results for students with disabilities (Elliot et al., 2000). This mandate includes investigation of access to, inclusion in, and participation on state and federal assessment programs for students with disabilities, as well as their participation in accountability systems. NCEO has been tracking and analyzing state policies on assessment participation and accommodations since 1992.

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization. This council consists of heads of departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO's mandate is to provide "leadership, advocacy, and technical assistance on major educational issues" (http://www.ccsso.org, retrieved May 23, 2009). The Council provides information on major educational issues to civic and professional organizations, federal agencies, Congress, and the general public.

A major organization,

> [t]he Council for Exceptional Children (CEC) is the largest international professional organization dedicated to improving the educational success of individuals with disabilities and/or gifts and talents. CEC advocates for appropriate governmental policies, sets professional standards, provides professional development, advocates for individuals with exceptionalities, and helps professionals obtain conditions and resources necessary for effective professional practice (http://www.cec.sped.org, retrieved May 23, 2009).

The National Association of Directors of Special Education (NASDSE), founded in the late 1930s, provides services to state agencies assisting in their efforts to improve educational outcomes for students with disabilities. NASDSE provides leadership throughout the United States, the federal territories and the Freely Associated States of Palau, Micronesia and the Marshall Islands. The association believes

> [a]ligning policies and practices to improve educational outcomes for [students with disabilities] is critical ensure full participation [of students with disabilities] in their education and transition to post-school employment (http://www.nasdse.org/AboutNASDSE/LetterFromOurPresident/tabid/404/Default.aspx, retrieved May 23, 2009).

Terms specific to meta-analytic techniques include mean effect, mean relative effect, Q-statistic, fixed-effects, random-effects, sensitivity analysis, and publication bias.

The mean effect, computed by weighting each effect size by the inverse of its variance (i.e., the effect size is multiplied by its weight), is used to find the *central tendency* for the aggregate of the effect sizes computed in the meta-analysis.

The mean relative effect, as it applies to research on students with disabilities and general education populations, is (i) the difference between the mean effect on students with disabilities (target population) and the mean effect on the general education population or (ii) the difference between the mean effect on students with disabilities in a non-accommodated assessment condition and the mean effect on students with disabilities in an accommodated assessment condition.

The Q-statistic is "…a measure of weighted squared deviations…" (Borenstein, Hedges, Higgins, & Rothstein, 1009, p. 105) and is used to assess heterogeneity in effect size estimates; i.e., the variability in true effect sizes. The Q-statistic helps determine whether effect size is consistent. If effect size is consistent we are able to focus on the summary effect size statistic, if not, we must focus on the dispersion of effect sizes.

The fixed-effects model is one of the two statistical models used in meta-analyses. Under the fixed-effects model, one true effect size is assumed to underlie all studies in the meta-analysis.

The random-effects model, the second of the two statistical models used in meta-analyses, allows for the possibility of different effect sizes underlying the studies included in the meta-analysis. That is, if we were able to select a random sample of primary studies from the infinite number of studies available, the true effect sizes would be distributed about a mean.

In meta-analytic studies, a sensitivity analysis focuses on “the extent to which the results are (or are not) robust to assumptions and decisions that were made when carrying out the synthesis” (Borenstein et al., 2009, p. 368).

Publication bias refers to the likelihood that certain types of research, specifically research conducted that did not find significant results, is not included in a meta-analysis. When meta-analyses do not include unpublished research work, an upward bias in effect size summary statistics will be found. Methods to examine publication bias include funnel plots, Rosenthal’s Fail-safe N, Orwin’s Fail-safe N, and Duvall and Tweedie’s Trim and Fill.

Summary

The purpose of the study was to: (a) determine whether there is empirical support for effects of testing accommodations, (b) provide an estimate of the mean effect size, and (c) contribute to the understanding of effective test accommodations for students with disabilities.

This study aims to add to the existing body of research and research syntheses on testing accommodations for students with disabilities by extending the original work of Chiu and Pearson (1999). This research narrowed the focus, from English language learners and students with disabilities on a variety of different assessments, to students with disabilities on high-stakes and/or large-scale, paper and pencil assessments only, focusing on participation on federal, state, and district tests with accommodations for students with disabilities. Further, meta-regression analyses and graphic representations, not available to Chiu and Pearson in 1999, provide a unique contribution to research in this area.

Sireci et al. (2005) stated that our "… challenge is to implement … accommodations appropriately and identify which accommodations are best for specific students" (p. 486). This cannot be accomplished solely through the use of primary and secondary analyses. Synthesis of research, that is meta-analysis, must be employed to provide more definitive answers to research questions posed in the area of assessment accommodations and students with disabilities. To that end, this study provides a quantitative, rather than a qualitative, view of the aggregate research on all researched testing accommodations for students with disabilities, something that has not been done since 1999 by Chiu and Pearson.

Chapter Two

Method

The present research proposed using two different statistical methods, meta-analysis and meta-regression, in an effort to examine research on the efficacy of assessment accommodations for students with disabilities. Use of these meta-methods allowed us to scrutinize the existing research literature for overall trends using quantitative methodologies in an effort to better understand findings across the breadth of the research literature in this area.

Purpose of the current study.

The purpose of the current study was threefold. The current study sought to establish if assessment accommodations provide a more effective assessment of students with disabilities than no accommodations; estimate the strength of this effect; and add to the knowledge base pertaining to effective assessment accommodations for students with disabilities. As such, results from this study were used to summarize previous research, estimate population parameters, and generalize findings from prior research.

Research Hypotheses

The current study addressed the following hypotheses for the meta-analytic portion of the research:

- Research Hypothesis 1: Is there empirical support for effects of test accommodations for the target group, students with disabilities, as opposed to their typically developing peers?
- Research Hypothesis 2: As measured by effect size, does each of the following constitute an effective accommodation for students with disabilities?
    - Presentation test accommodations?
    - Response test accommodations?
    - Setting test accommodations?
    - Timing/Scheduling test accommodations?

The current study addressed the following hypothesis for the meta-regression portion of the current research:

- Research Hypothesis 3: Which type of accommodation(s)–Presentation, Response, Setting, or Timing/Scheduling–more effectively remove construct-irrelevant variance from target students' test scores?

Meta-analysis

Meta-analysis, one type of research synthesis, was selected as a method to integrate research findings from multiple research studies, *vis-à-vis* assessment accommodations for students with disabilities. "Research syntheses attempt to integrate empirical research for the purpose of creating generalizations" (Cooper & Hedges, 1994, p. 5). Meta-analysis provides a statistical method to integrate information from primary studies on assessment accommodations for students with disabilities selected for further scrutiny and analysis, something which could not be accomplished using syntheses of the research literature; i.e., integrative narrative reviews.

The research design for the present study was based on Cooper and Hedges' (1994) stages of research synthesis found in their "definitive *vade mecum*" (p. 7). These stages include: (i) problem formulation, (ii) data collection/literature search methods, (iii) data evaluation/coding and evaluating research reports, (iv) analysis and interpretation/meta-analytic calculations of effect size(s), and (v) public presentation/meaningful interpretation and effective presentation of the synthesis results. The problem formulation was addressed via the purpose for this study and the research hypotheses posed. The purpose and research hypotheses form the basis for the selection of studies for the meta-analysis. Reports selected for the present meta-analysis were based upon the following selection and exclusion criteria.

Criteria for selection of studies.

Studies selected had to meet several criteria in order to be considered for the meta-analysis. Explicit inclusion and exclusion criteria aid in the selection of relevant studies, as well as limiting researcher bias (Lipsey & Wilson, 2001). General categories guiding selection criteria were "(a) the distinguishing features of a qualifying study, (b) the research respondents, (c) key variables, (d) research design, (e) cultural and linguistic range, (f) time frame, and (g) publication type" (pp. 16 - 17). Although an exhaustive search of the literature is not required when defining inclusion criteria (White, 1994), it is recommended that researchers do not use criteria that are too strict as useful reports may be overlooked (Lam & Kennedy, 2005).

Inclusion criteria were separated into two non-overlapping groups: (i) substantive domain of inquiry and (ii) methodological characteristics. This allowed for a more granular look at existing research prior to creating a meaningful common metric across the studies under consideration.

Studies that did not fully meet both substantive and methodological inclusion criteria were included in some cases. The rationale for including these studies is provided in the analyses section. Further, coding was created to explicate inclusion of these studies.

Substantive inclusion criteria.

Initial substantive inclusion criteria focused on four different areas: (i) types of students included in the analyses, (ii) type of assessment accommodation used, (iii) type of assessment under investigation, and (iv) year of publication.

Substantive inclusion criteria were as follows:

(i) Experimental or quasi-experimental studies that quantitatively examined the effects of assessment accommodations for students with disabilities in the regular educational system from kindergarten through college. Definition of *students with disabilities* followed categories of disability outlined in IDEA (2004) legislation.

(ii) Studies examining assessment accommodations falling under the categories of presentation, response, setting, and timing/scheduling as defined by Sireci et al. (2003).

(iii) Studies examining large-scale, high-stakes, or commonly-used published assessments of achievement or college entrance.

(iv) Studies conducted and/or published on or after 1999 through June, 2011. This was purposefully done in order to ensure that studies included did not overlap with the previous meta-analysis conducted by Chiu and Pearson (1999).

Substantive characteristics were coded and accounted for in the statistical analyses conducted.

Demographic variables were also recorded as such variables were seen as a potential source of covariate and/or mediator information.

Methodological inclusion criteria.

Initial methodological inclusion criteria also guided the selection of studies for the meta-analysis. Methodological inclusion criteria focused on four different areas: (i) available data, (ii) examination of single assessment accommodation, (iii) assessment accommodation validity, and (iv) research examining boost, differential boost, and/or the interaction hypothesis.

Methodological inclusion criteria were as follows:

(i) Experimental and quasi-experimental studies with statistical data such as means and standard deviations, or significance test results necessary to calculate an estimated effect size of the impact of the testing accommodation under study.

(ii) Study designs focusing on the effects of single accommodations as opposed to effects of accommodation packages, that is, multiple accommodations for individual students. Note that more than one assessment accommodation may be analyzed in a single study with results for each accommodation reported separately. However, analysis needed to focus on one accommodation at a time for inclusion in the meta-analysis.

(iii) Assessment accommodation which did not alter the construct being assessed; i.e., studies examining assessment accommodations and not assessment modifications were included in the meta-analysis.

(iv) Research examining boost, differential boost (Fuchs & Fuchs, 1999), and/or the interaction hypotheses (Sireci et al., 2005) for students with disabilities and/or typically developing students.

Study quality was not explicitly coded. Research by Ahn and Becker (2011) showed that the use of quality weights in meta-analysis does not add to the analysis nor does it significantly change results found, thus they recommend against the use of quality weights. However, for the present meta-analysis, type of publication was noted; i.e., article, dissertation, report, and conference proceeding, in lieu of study quality.

Methodological characteristics were coded and accounted for in the statistical analyses conducted.

Categorization of test accommodation research.

Methodological inclusion criteria are intimately linked with the type of methodological approach used by researchers in this field. Tindal (1998, cited in Bolt & Thurlow, 2004) categorized primary research on assessment accommodations into three approaches. A fourth approach, or category, was added by Fuchs et al. in 2000a. The four approaches are descriptive, comparative, experimental, and individual diagnosis.

The descriptive approach provides a logical analysis of difficulties associated with disability, conducted to determine which accommodations are considered to be helpful and allow students with disabilities to demonstrate their knowledge and skills on an assessment (e.g., surveys of perceived integrity and effectiveness of accommodations). Such research is generally relevant to policy presentations, policy interpretations, or implementation analysis.

The comparative approach examines test scores, generally existing test scores, to see how accommodations affect scores of different groups of students. Research employing this type of approach helps articulate how accommodations function in an applied setting. Such research has issues with confounding factors, such as decisions to provide accommodations and how accommodations are administered, limiting any conclusions reached. Post hoc comparisons are primary examples of studies employing a comparative approach.

The experimental approach isolates effects of accommodations by manipulation of presence and/or absence of accommodations among different groups. This is generally the preferred approach for research in this area. Examples of research employing the experimental approach are group experiments and single subject experiments.

The individual diagnostic approach examines the set of procedures used to determine which accommodations an individual student with disabilities should receive. "Because accommodated students frequently receive multiple accommodations that are based on their individual needs, the individual approach seems to exemplify how accommodations are used in real testing situations" (Bolt & Thurlow, 2004, p. 143), thus, are more likely to provide information on real-world assessment conditions.

While Bolt and Thurlow (2004) suggest that accommodations should only be considered valid if they are supported by each one of these four approaches this meta-analysis endeavored to provide information based on research guided by experimental approaches, focusing on research that looked at boost, differential boost, or the interaction hypothesis.

Exclusion criteria.

Studies which were not included in the meta-analyses of testing accommodations for students with disabilities were excluded based on the following criteria:

(i) Studies did not report means and standard deviations and/or significance test results. Such research did not provide enough information to create an aggregate metric for an effect size.

(ii) Studies did not use large-scale assessments, high-stakes assessments, commonly used/published achievement or college entrance assessments, or proxies for these types of assessments (e.g., researcher-developed assessments using items from state assessment item banks). Aggregating multiple types of tests was thought to provide an apples-to-oranges rather than an apples-to-apples type of comparison.

(iii) Studies looked at assessment accommodation packages. Unless information from such studies could be disentangled, these studies were excluded from the meta-analyses.

(iv) Studies examined assessment modifications. Including such studies was beyond the scope of the present analyses. Further, these studies were thought to cloud interpretations which could be made as assessment validity would be altered in such a way that results from the assessment would no longer be comparable to results from a more standardized type of testing condition.

(v) Studies did not report primary research findings for students; i.e., secondary studies.

(vi) Studies published before 1999.

(vii) Studies found in multiple sources, such as dissertations, papers, and publications. For studies located in multiple sources, the study with the most information which could be coded and/or was thought to be easier to retrieve was selected.

(viii) Qualitative studies.

(ix) Research, not reported in English, or for which English translations were not available.

Of the 81 studies located, 47 studies were excluded from the meta-analyses. These studies were excluded from the meta-analyses as the purpose for the research conducted did not match that of the current study, data did not include information that could be used to calculate an effect size, some of the data necessary to calculate an effect size were missing, or the study was eliminated after performing an outlier analysis. Citations and reasons for the studies' exclusion may be found in Appendix H. A further eight studies could not be located (see Appendix I).

Selection criteria were tested and refined by applying these criteria to five randomly selected studies. One of the studies, Burch (2004), was rejected as the students used computers to answer test questions. This was not apparent when reviewing the title, abstract, and research questions for the article. The four articles which were coded were:

(i) Abedi, J., Kao, J. C., Leon, S., Mastergeorge, A. M., Sullivan, L., Herman, J., & Pope, R. (2010)

(ii) Helwig, R., Rozek-Tedesco, M.A., Tindal, G. (2002)

(iii) Kosciolek, S. & Ysseldyke, J. E. (2000)

(iv) Ofiesh, N., Mather, N., & Russell, A. (2005)

Final selection criteria, both substantive and methodological, were integrated into the Coding Manual (Appendix D), providing a method of labeling all studies reviewed. This was done to assist in potential future analyses, whereby excluded studies, solely and in combination with studies selected for the present research, could be analyzed using similar methods.

Overview of the selection process.

The selection process started with a review of citations found in secondary studies, located on the NCEO website, involving the summary of the research on the effects of tests accommodations. Secondary studies included both narratives and syntheses of the research literature. As well, titles and keywords found through a comprehensive database search were screened. Additionally, bibliographies from located studies were examined for research work that might potentially be included. Studies thought to be of interest were marked for retrieval. Inclusion and exclusion criteria guided the identification of studies thought to be relevant to the population of studies to be used in this meta-analysis, with exclusion of studies that did not meet the substantive and methodological inclusion criteria. While this was a guiding principle, exceptions were made in certain cases where studies found met some, but not all, of the inclusion criteria. The rationale for including these studies was provided in the coding database accompanying each study. Note that a coding form was developed (see Appendix F).

This form was used to structure the coding database, as well as for training an additional coder for the inter-rater reliability study.

Unpublished reports were also considered for retrieval during the selection process. It was thought these studies were necessary to provide a methodologically sound meta-analysis. Glass et al. (1981) noted that there was reporting bias, whereby research with significant results or results with a high *surprise* factor were more likely to be published while results from research where there were non-significant findings or findings that are contrary to mainstream theory were less likely to be published. As well, for journals where blind review was not conducted there may be issues of editorial bias; reputation of author, affiliation of author, novelty of research affecting editorial selection; and/or reviewer bias; author prestige, author nationality; affecting reviewer selection. Several reports and conference proceedings were located. Of these, five reports were included in the final analyses. While it was expected that not all journal articles located would be peer-reviewed, this was not the case. Of the final 19 journal articles included in the analyses, all were peer-reviewed.

Of concern when identifying potentially relevant studies was the differentiation between test accommodation and test modification. Studies where test modifications were used, or where it was not clear whether a test modification or a test accommodation was used, were removed from the pool of studies used in the meta-analysis.

The screening of potentially relevant research was an iterative process, whereby selection criteria and guidelines for selection were further refined and clarified. The initial screening of these articles included examination of the title of the study, study abstract, and research purpose/questions for the study. As this process was not wholly reliant on identification of studies through citations provided by the electronic databases searched, it was expected that fewer studies were missed due to insufficient or misleading information found in these citations. Moreover, the general rule for inclusion of studies identified through electronic database citations was to err on the side of over-inclusion rather than exclusion of prospectively applicable studies. Studies not meeting inclusion criteria were winnowed from the meta-analysis and were not included in final counts of studies found.

Search strategy.

The search strategy employed for the meta-analysis was guided by the selection criteria as well as an extensive search strategy designed to be congruent with the meta-analytic research hypotheses posed. Hedges (1994) stated that "[t]he sampling procedure must be designed so as to yield studies that are representative of the intended universe of studies" (p. 35). While the notion of *exhaustive sampling* is meant to garner a representative and sufficient sample of studies of assessment accommodations and students with disabilities, it must be noted that representativeness of the variability of studies in the potential universe of studies in the field may not be achieved due to issues of publication bias, including both editorial and reviewer bias. A combination of Lipsey and Wilson's (2001) and White's (1994) suggestions for finding research reports were used to identify relevant research. Lipsey and Wilson's approach utilizes the following sources:

> (a) review articles, (b) references in studies, (c) computerized bibliographic databases, (d) bibliographic references volumes, (e) relevant journals, (f)

> conference programs and proceedings, (g) authors or experts in the areas of interest, and (h) government agencies (2001, p. 25).

White's approach includes "(a) footnote chasing or review of bibliographies of selected articles, (b) consultation, (c) searches in subject indexes, such as electronic database searches, (d) browsing, and (e) citation searches of electronic databases" (1995, p. 46). It should be noted that there is overlap between these approaches. It must also be noted that electronic database searching is more prevalent with the introduction of personal computing, greater personal computing power, and the push to store as much information online as possible. As well, many online databases now include search and retrieval functionality.

Computerized database searches.

Computerized database searches were conducted to find potentially eligible studies for the meta-analysis. Articles, reports, papers, or dissertations will be referred to as *research studies* in this section. As most current online searches yield both the bibliographic reference and the research study in question it was not generally necessary to locate the research study once the bibliographic reference was located. Location of some of the research studies did require a two-step process, whereby the bibliographic reference was found using one database but the study itself was located in a different database. For example, a citation for a research study would be found using ERIC but the copy of the article was available through the PsycINFO database.

Computerized database searches were conducted using natural language and controlled vocabulary keyword searches (White, 1994). Natural language refers to terms that "emerge naturally from the vocabularies of authors" (p. 49) while controlled vocabulary keywords refer to the "terms … added to the bibliographic record by the employees of A&I services or large research libraries" (p. 50). Generally, controlled vocabulary keywords are found in a thesaurus produced specifically for the database being used. Keywords are typically associated with the title, abstract and/or standardized descriptors for the study in question.

Lipsey and Wilson (2001) recommend using keywords that broadly cover the domain of interest by

> (a) identifying all those standardized descriptors in a given database that may be associated with the studies of interest and (b) identifying the range of terms that different researchers might include in their study titles or abstracts that give a clue that the study might deal with the topic of interest (p. 26).

They further recommend using appropriate Boolean connectors; for example, and, or, not, to limit or expand the search as necessary. Further, they recommend caution when trying to narrow the size of the search as many eligible research studies may be missed. As there is often a fine line between a search which is too expansive and one which is too restrictive, there was much trial and error in finding the appropriate search terms and Boolean connectors. Some of the trial and error in creating appropriate search phrases was reduced through examining the titles and abstracts of research studies which were identified during the review of the literature.

Based on the recommendations of Lipsey and Wilson (2001) and White (1994) a list of search criteria, keywords, and connectors was developed. Search criteria included, but were not limited to, combinations of following terms: accommodation, test, standardized assessment, large-scale assessment, high-stakes assessment, and disability. A complete list of search criteria used for searching databases, databases searched, and number of eligible studies found is located in Appendix G. It should be noted that once studies were located, they were reviewed for eligibility as not all studies located were considered relevant for the purposes of the present meta-analysis.

While the current meta-analysis does not involve multiple disciplines, it does involve many different facets of educational research; for example, research on state assessment programs, validity of assessment accommodations, and policies developed for effective use of assessment accommodations. As such, multiple divergent databases were used to locate eligible studies. These databases were Academic Search Complete, Applied Social Sciences Index and Abstracts (ASSIA), British Periodicals, Dissertations & Theses @ University of Denver, ERIC, Google Scholar, JSTOR, ProQuest Dissertations & Theses (PQDT), ProQuest Education Journals, PsycINFO, PsycARTICLES, and Sociological Abstracts.

An effort to retrieve unpublished studies was made by searching Dissertations & Theses @ University of Denver and ProQuest Dissertations & Theses (PQDT). As it was suspected that the number of unpublished studies found was not representative of the number of unpublished studies in this area, publication bias was explored using *Comprehensive Meta-Analysis V.2.2.050. Comprehensive Meta-Analysis V.2.2.050* provides a method, similar to the calculation of a fail-safe number, to represent the number of unpublished studies with a negligible, or zero, effect size. This was deemed necessary to examine the overall effect of publication bias. Funnel plots and calculations for several types of 'fail-safe' numbers are provided by the program.

Overview and results of the search process.

A comprehensive search strategy, based on a number of different approaches, was used to locate eligible research studies for the current meta-analysis. Reference lists found in syntheses, searches of electronic databases, conference proceedings, web sites, and hand searches of journals such as the American Educational Research Journal, Educational Measurement: Issues and Practice, and Educational Researcher were used to identify likely studies for the meta-analysis. As there is generally a lag between publication and listing in electronic databases, hand searches of nine journals, focusing on large-scale assessment, assessment/test accommodations and special education, were also conducted. As well, in an effort to ensure the most recent studies were included, papers presented at conferences sponsored by the American Educational Research Association, the Council for Chief State School Officers Large-Scale Assessment Conference, National Council on Measurement in Education, and National Association of School Psychologists in 2010 and 2011 were examined. Further, web sites for organizations such as NCEO (with a searchable database), Wisconsin Center for Education Research, Center for Research on Evaluation, Standards, and Student Testing, College Board, and Behavioral Research and Teaching at the University of Oregon were explored for prospective research studies. Additionally, secondary studies identified as a part of the review of the literature provided summaries of the research on testing accommodations for students with disabilities, supplying useful search terms for types of accommodations being used, as well as additional direction regarding research findings *vis-à-vis* accommodation use. Moreover, research studies needed to be published or conducted between January 1999 and July 2011.

The initial search was broadened in an effort to locate studies on the interaction hypothesis (Sireci et al., 2005) and included the terms differential boost (Fuchs & Fuchs, 1999), boost studies, and comparative studies.

Database searches were conducted for substantive and methodological terms. In addition, using database indices, citations, and abstracts several subject headings which were of potential interest were identified. A combined search of pertinent substantive and methodological terms yielded a single meta-analysis by Chiu and Pearson (1999). For purposes of this meta-analysis, the Chiu and Pearson study was used to frame the timeline for study eligibility.

Titles, keywords, abstracts, and research questions/hypotheses/purposes for each research study found were reviewed for inclusion in the meta-analysis. All studies were reviewed by the primary researcher and were selected for inclusion or exclusion. As well, eligible research studies were reviewed for prospective keywords for additional database searches. Furthermore, reference lists for these studies were used to identify additional studies. Research studies considered ineligible, based on exclusion criteria, were cited (see Appendix H).

Several attempts were made to locate studies which appeared to meet the substantive and methodological criteria. Efforts to collect as many unpublished studies as possible were also made. Several online databases were searched for the missing studies. When the researcher was unable to find the research studies, online library resources were used. A total of seven studies were not retrievable. See Appendix I for a complete listing of citations for irretrievable studies.

Comprehensive searches of online databases yielded 226 studies, not including duplicates. Eighty studies; comprising 33 research articles, 11 research reports, 34 published dissertations, and 2 papers; i.e., unpublished research studies; were initially identified as eligible research studies. After reviewing these studies, all 80 research studies were found to focus on the effects of test accommodations for students with disabilities and were empirical. These 80 eligible studies were then reviewed (i) for serious methodological flaws such that designs posed threats to external validity or did not use random assignment when possible (Bangert Drowns, 1993), (ii) to determine if there was sufficient statistical information to calculate effect sizes, and (iii) to determine if they matched the substantive research hypotheses posed by this study. While none of the studies were considered to have serious methodological flaws, 27 were eliminated as they did not match the substantive research hypothesis; e.g., the primary study examined multiple accommodations for individual research participants or did not disaggregate students with disabilities from English language learners. Results indicated that 44 of the remaining 53 studies appeared to contain the information necessary to calculate effect sizes. However, 5 studies did not contain information necessary; e.g., mean and standard deviations, to calculate effect sizes, and a further 3 research studies were eliminated as they used a comparative research design. Not included in this total were 20 duplicates, and of these, 10 were duplicates for rejected studies. The work that was easiest to locate, generally journal articles, was coded while the duplicate, generally a report or dissertation, was used to locate and code information that was not included in the primary work. Based on these analyses, 36 studies were retained for inclusion in the meta-analysis. It should be noted that when selecting a research study for inclusion, it was thought that journal articles and dissertations were the most accessible sources of information, thus they were more likely to be included in analyses than reports or conference proceedings.

The 36 eligible studies were further evaluated to ensure that explicit information regarding the nature of the disabilities of the target group and, where necessary, comparison groups was provided. As well, the research studies were reviewed looking for unambiguous descriptions of assessment accommodations used in the research and details regarding implementation of the accommodations.

Coding and classifying study variables.

As part of the meta-analysis, variables identified in the research studies were coded according to a codebook (Appendix D) used to collect data for the present meta-analysis. Coding forms were developed based on the codebook. Both the codebook and coding forms developed were adapted from Lipsey and Wilson (2001), Stock (1994), and Van Horn, Green, and Martinussen (2009), with coding formulated to allow for statistical analysis of the eligible research studies. Due to the complexity encountered during initial coding, a coding manual was also developed. The coding manual contains instructions on how to enter information on the coding form, study inclusion and rejection rules, and glossaries for useful keywords (see Appendix G).

Coding was based upon both substantive and methodological concerns (Glass, McGraw, & Smith, 1981; Stock, 1994). As well, coding information was based on “two rather different parts” (Lipsey & Wilson, 2001, p. 73): information regarding (i) research study characteristics and (ii) empirical findings. While some variables used in the codebook were decided upon *a priori*; for example, publication type and research study type, many of the variables were established at a later stage, thus capitalizing on the iterative nature of the coding process.

Development of a codebook was an iterative process, progressing through the data collection phase of the study as this researcher became more knowledgeable about the domain of inquiry and the statistical demands and biases which needed to be addressed in the meta-analysis. Steps in coding and classifying study variables included the following: (i) creating the codebook with initial set of codes (Lipsey & Williams, 2001; Van Horn et al., 2009); (ii) reading five articles with the initial codebook and revising as new information came to light; (iii) coding one article during the coder training session and revising with the aid of the second coder; (iv) coding three more articles with the revised codebook and revising again; (v) create coding forms (Appendix F) and a coding manual (Appendix D) to accompany the codebook (Appendix E); (vi) coding all remaining studies; (vii) using a second coder to code 15% of the studies using the coding manual, codebook and coding forms; (viii) calculating inter-rater reliability for completed coding for a 15% random sample of eligible studies.

The codebook consisted of the following broad categories: report identification, study retrieval information, study citation, research participant information, assessment citation and demographic information, research methodology, research design, research results, and a proxy for quality of study. Each category was defined in terms of the variables it contained with different levels or options associated with each variable described in the codebook. For example, report identification contained data regarding the year of publication, type of publication (dissertation, article, report, paper), and name of publication. Assessment citation and demographic information contained data related to the kind of scales used; names of tests or diagnostic systems, reliability, test item format, and construct or content assessed (see Appendix D). Test item format was included as

> … Koretz and Hamilton (2000) found differences between the performance of students with disabilities' performance on multiple choice and constructed response items, [thus] future research should further evaluate potential differential impact of accommodations on these different item formats (Zenisky & Sireci, 2007, p. 17).

It should be noted that students with ADHD were classified as 'other health impairment' in one study as

> after the passage of IDEA in 1990 and a subsequent 1991 memorandum, that the U.S. Department of Education and its Office of Special Education chose to reinterpret these regulations, thereby allowing children with ADHD to receive special educational services for ADHD per se under the 'Other Health Impaired' category of IDEA (Barkley, 2006, p. 16-17).

The coding form reflected each of these broad categories with the different levels or options provided.

A proxy for study quality was used as there is much disagreement in the field regarding classification of study quality. Ahn and Becker (2011) found that using "quality weighting adds uncertainty to average effect sizes but does not eliminate serious bias related to study quality… [and] adds bias in many cases" (p. 579-580). Therefore, a pseudo-measure of quality, grouping primary studies by (i) published journal articles and conference proceedings that are peer reviewed, (ii) published reports which may or may not undergo a peer review process, and (iii) unpublished dissertations which are reviewed by dissertation committee members, was used. The ‘quality’ for journals, conference papers, and dissertations was, arguably, considered ‘equivalent,’ while research reports were viewed as being of ‘lesser quality.’

It must be pointed out that some variables found in the research studies were very difficult to classify; for example, participant disability classification; thus room was left for qualitative descriptions. These descriptions were later analyzed, identifying commonalities and differences that were then coded so they could potentially be included in the statistical analysis (Lipsey & Wilson, 2001). Per Lipsey and Wilson’s recommendation, such qualitative descriptions were “only used for critical issues and when absolutely necessary” (p. 74). As well, there were instances where variables could not be coded based on the data included in the study being analyzed. In these instances, an explicit option to indicate that it was not possible to “tell what the status of the study [was] on that item” (p. 88) was provided in the codebook and the accompanying coding form *via* a *missing* option, and coded as not reported. It was also necessary to distinguish between missing and not applicable (Lipsey & Wilson, 2001), thus a *not applicable* category was also provided.

As several different research designs are found in research studies involving assessment accommodations, coding for research design was implemented. This allowed for the inclusion of studies with diverse research designs, whereby different effect sizes were calculated to reflect the differences in research design (Lipsey & Wilson, 2001). It should be noted that this was not a factor in the Chui and Pearson (1999) meta-analysis as most studies conducted prior to 2000 used boost research designs.

Dependent and non-independent effect sizes.

It is recommended that the same data set should only be used once in an analysis (Lipsey & Wilson, 2001). For example, the results of a research study may be presented at a conference and then later reported in a journal. In such instances, for the present study, the unit of analysis was the research study containing the most information that could be readily coded.

Some eligible research studies provided dependent and non-independent effect sizes; that is, there were multiple samples with multiple results reported within a single research study. When this occurred, it was necessary to distinguish between the types of effect sizes as only effect sizes that are independent are suitable for the calculation of the overall mean effect size in a meta-analysis (Lipsey & Wilson, 2001). Issues calculating mean effect when multiple effect sizes are present include problems estimating the variance across the studies, issues when conducting significance testing, problems looking for moderators, providing inaccurate sample size(s), and giving too much weight to a few studies. When using the Hunter and Schmidt (1990) method to calculate effect sizes, multiple effect sizes in a single study appear to be less of an issue, with some data indicating that these estimates may in fact be better (Martinussen & Bjørnstad, 1999). Suggestions to resolve this issue include (i) picking one of the results randomly, (ii) using the most common effect size, and (iii) computing the mean effect size and the mean sample size, which is the mean of the subjects per effect size and not the mean of all the subjects involved (Martinussen[1], 2007). Martinussen (2007) recommended using the third method and, in the cases where the samples in the research study were dependent, the third method was employed. In the cases where multiple samples in a single research study were independent, the information was captured twice; once to analyze the data while accounting for the independent samples, using the substudy as the unit of analysis, and once when not accounting for the independent samples; i.e., to examine the aggregate, using the study as the unit of analysis. It should be noted that in the instances where substudy was the unit of analysis, and there were dependencies, there was a reduction in the effect size estimation.

Coding characteristics of operational definitions.

Consideration of certain constructs central to the meta-analysis needed to be taken into account. Specific operational and conceptual criteria for assessment, assessment accommodation, and student with disabilities were used to guide coding information for their associated variables; for example, type of assessment, category of accommodation, and sampling method.

A range of large-scale assessments was used in the research studies collected. To account for the variety of assessments, each assessment was coded in relation to the assessment category measured (achievement, aptitude, performance, placement, selection, screening, diagnosis, other), construct and/or content measured (mathematics, reading/language arts, science, writing, social studies, physical education, multiple content areas, other), method of standardization (norm-referenced, criterion-referenced,

[1] Personal communication with Dr. Monica Martinussen, (May, 2007).

domain-referenced, standards-based), and assessment format (multiple-choice, fill-in-the-blanks, short answer questions, open-ended questions). Assessment citation information was entered as qualitative information.

To account for the diversity of assessment accommodations included in the analysis, accommodation operational definitions were coded in relation to predetermined categories based on the NCEO criteria of presentation, response, setting, and timing and scheduling. These categories were further broken down into specific accommodation; i.e., oral administration as a sub-category for presentation. Every effort was made to determine the mode students used to answer the assessment questions; i.e., paper and pencil or computer. If students used a computer to read or hear assessment directions, questions, response options, etc. and used a paper and pencil form to answer the questions on the assessment, then the assessment was included in the meta-analysis. It was rejected if the students used a computer to answer the assessment questions.

To accurately report on the students with disabilities category, each research study was coded according to explicitly stated information on type of disability. Disabilities were coded according to the 13 special education categories listed in federal special education law (Individuals with Disabilities Act reauthorization of 2004, PUBLIC LAW 108–446, 2004):

> mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as 'emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities (Part A (SEC. 602) (3) (A) (i), 118 STAT.2652, 2004).

With the iterative nature of coding, some adjustments were made to the coding process. It was originally hoped that there would be viable number of studies using original versions of high-stakes, large-scale, or standardized tests. However, the majority of studies used researcher-developed assessments, drawing from large-scale and/or high-stakes assessment item banks; using such data was believed to be appropriate. Additionally, both achievement and ability measures were included in the meta-analysis as achievement and ability are highly correlated (Tindal & Fuchs, 2000). Comparative research designs; i.e., post hoc analyses, were dropped from the meta-analysis as they lacked use of random assignment or counterbalancing thus did not appear to adequately address either meta-analytic research hypothesis posed. It was felt that empirical research; i.e., experimental or quasi-experimental research, was a better match to the research purpose for this study as it is a way of gaining knowledge through direct observation or experience. As well, although type of assessment, that is, norm-referenced, criterion-referenced, domain-referenced, standards-based, curriculum-based, was coded it was not included in any of the analyses as there were much missing data.

Issues of reliability throughout the coding process.

Another area of consideration during the coding process was the avoidance of errors and biases introduced when coding the data. By providing explicit, unambiguous descriptions of each coded variable in the codebook, “coding errors” associated with judgments were, for the most part, avoided. Additionally, use of electronic coding forms, with data entered directly on a computer, were used to avoid commonplace coding errors associated with data entry, thus avoiding reentry or copying of data from one database to another. Although these preventive measures were implemented, a statistical analysis of coding errors and bias was conducted, as the introduction of coding error cannot be entirely avoided. After the coding manual (Appendix D), codebook (Appendix E), and coding form (Appendix F) were developed, two different coders reviewed and coded 15% of eligible studies. A measure of inter-rater reliability, percentage agreement, for a random sample of 15% of all studies was calculated. In the event there was disagreement between the two raters, the rationale for the difference was discussed and eventual consensus on coding was reached; and, when needed, the coding form reflected changes. The inter-rater reliability by category, the categories being study citation, participant information, assessment information, accommodation information, statistical analysis, and results (i.e., means and standard deviations), and 'additional' results (i.e., significance tests and correlation coefficients between the non-accommodated and accommodated conditions), ranged from 77% to 100%, and was 92% overall. The percentage agreement for continuous participant and results data, used to calculate effect sizes for the primary studies, was 98.9%. Additionally, the reliability coefficient calculated for these data, reached 1.00 and was statistically significant. The inter-rater reliability was considered adequate for purposes of this study. While final coding was consensual, calculation of reliability did not include coding which changed; i.e., it was computed before the original codes were changed.

To minimize other possible issues of reliability, joint training sessions for the coders were conducted. During the training sessions the coding manual, codebook, and coding form were reviewed, followed by a discussion regarding code entry using an Excel spreadsheet. Once the review was completed, the two coders examined and coded a previously coded study together. Additionally, all coding decisions were recorded, together with the rationale for these decisions, and the information was saved to an Excel spreadsheet. Further, the same ID number was used for the same research study even when the study was found in multiple sources such as papers, research reports, and journal articles. An alpha character, beginning with *A,* was appended to the ID number when multiple instances of the same study were found. For studies with multiple samples and multiple results, a lower case roman numeral following the ID number and the alpha character, beginning with *i*, was appended to the ID number.

If both multiple independent sections and a summative section with information to estimate an effect size were present in a study, the information from the summative section was not included in the meta-analysis.

In an effort to ensure comparisons made were *apples to apples* and <u>not</u> *apples to oranges*, eligible studies had to focus on (i) students with disabilities and groups compared to students with disabilities; not English language learner or other group comparisons, (ii) testing accommodations which could be categorized under presentation, response, setting, and/or timing/scheduling, (iii) studies examining a single accommodation, and (iv) large-scale, high-stakes, published assessments, or researcher-developed assessments using items banks from large-scale and/or high-stakes assessments. It was expected that these assessments would present fewer issues with reliability and validity.

Statistical methods of analysis.

Following the coding of eligible studies, a suitable effect size statistic and appropriate statistical methods to combine effect sizes across studies were selected. Meta-analytic experts have devised statistical procedures for calculating a variety of effect sizes, weighting the mean effect sizes, estimating the effect of other potential moderators, correcting effect sizes for attenuation, and combining effect sizes from studies employing different designs. In texts authored by Borenstein et al. (2009), Hunter and Schmidt (1990), and Lipsey and Wilson (2001), information on meta-analytic statistical procedures is presented. These texts, together with coursework in meta-analysis taken at the University of Denver, provide primary references for the statistical methods used in the present meta-analysis.

Comprehensive Meta-Analysis V.2.2.050 (Borenstein, Hedges, Higgins, & Rothstein, 2009) (http://www.metaanalysis.com/index.html) was used to compute the necessary meta-analytic statistics.

Methods for calculating independent effect sizes.

“A critical step in meta-analysis is to encode or ‘measure’ selected research findings on a numeric scale, such that the resulting values can be meaningfully compared to each other and analyzed much like any other set of values on a variable” (Lipsey & Wilson, 2001, p. 34). Effect size statistics, previously referred to; provide the “index used to represent study findings in a meta-analysis” (Lipsey & Wilson, 2001, p. 34). In order to meaningfully aggregate findings from primary studies it is generally necessary to determine a standardized scale appropriate to the types of research designs seen in the eligible research studies. As the unit of analysis; i.e., the research report, research article, conference paper, or dissertation; consistently examined differences between means for (i) students with disabilities, (ii) students with disabilities compared to other students with disabilities, or (iii) students with disabilities compared to typically developing peers, effect sizes based on the standardized difference between means formed the basis of the analysis.

For primary studies Hedges' *g*, an unbiased estimator of δ, the standardized mean difference, based on Cohen's *d*, was used to calculate the effect size for differences between means.

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{2.1}$$

$$d = \frac{\overline{Y}_e - \overline{Y}_c}{s_p} \tag{2.2}$$

where $\overline{Y}_e$ is the mean of the experimental group, in this case students with disabilities, $\overline{Y}_c$ is the mean of the control group, in this case typically developing students, and s_p is the pooled sample standard deviation.

$$g = d\left(1 - \frac{3}{4m - 1}\right) = d\left(1 - \frac{3}{4(n_e + n_c) - 9}\right) \tag{2.3}$$

For these calculations, means and standard deviations needed to be available for each unit of analysis. In some cases means and standard deviations were not available, so effect sizes were calculated from reported test statistics, such as a t-tests or tests of significance, when these data were available. Note that use of the pooled standard deviation for the groups under study is generally recommended. However, if the standard deviations for the groups under study are very different it is recommended that the standard deviation for the control group be used instead (Lipsey & Wilson, 2001).

While the control group standard deviation is the recommended standard deviation for the groups under study, the standard deviation used was pooled within groups; for example, pooled within the students with disabilities subgroup separately from the typically developing students subgroup. Pooling within groups does not assume the study-to-study variance (τ^2) is the same for all subgroups. As it was "anticipate[d] that the true between-studies dispersion [was] actually different from one subgroup to the next … tau-squared [was estimated] for each subgroup" (Borenstein et al., 2009, p. 163). With several studies within each subgroup, these estimates were not considered imprecise (Borenstein et al., 2009). In an effort to ensure these assumptions were appropriate, a sensitivity analysis was performed comparing pooled within-group standard deviation and pooled across-group standard deviation results.

The random-effects model was employed, as there was variation beyond sampling error from differences among studies' effect sizes. The random-effects model does not produce the substantial Type I bias for mean effects significance tests and moderator variables; i.e., interactions, seen with fixed-effects models. As well, confidence intervals generated using the random-effects model do not overstate the degree of precision for the meta-analytic findings (Hunter & Schmidt, 2000). Statistical significance of effect sizes were calculated using 95% confidence intervals. Effect sizes with confidence intervals that did not include zero were considered statistically significant.

While one effect size was provided per independent study, or independent section of a research study; i.e., substudy, a correction to the observed standard deviation was used to account for sampling error. Additionally, before combining the effect size data for the difference between the means into a mean effect size, Lipsey and Wilson (2001) recommend assessing the effect of outliers and adjusting individual effect sizes based on the consideration of common sources of error. All corrections were performed prior to running the final analyses.

The steps followed for calculating independent effect sizes included estimating the mean effect size, tests of significance for the test statistics and the size of the effect, and estimating and testing the variation between the units of analysis.

All effect sizes were interpreted using Cohen's (1992) labels for "mean" effect sizes where 0.8 is considered a large effect size, 0.5 is considered a medium effect size and 0.2 is considered a small effect size. At present, in the testing accommodation literature for students with disabilities, there are no clearly defined demarcations between small, medium, and large effects. Therefore, the values cited by Cohen were used as lower-bound estimates for calculated mean effect sizes as using this more conservative estimate was considered to be the more prudent course of action rather than possibly providing an overestimate with respect to the efficacy of testing accommodations.

Accounting for variance in the distribution of effect sizes.

After calculating independent variance estimates, variance in the distribution of effect sizes was accounted for. Mean effect size is difficult to interpret without examining the variance in the distribution of effect sizes and ensuring that parametric statistical test assumptions have been addressed.

Outlier analysis.

As the "purpose of meta-analysis is to arrive at a reasonable summary of the quantitative findings of a body of research studies" (Lipsey & Wilson, 2001, p. 107), the presence of extreme values for effects may be unrepresentative of the research area of interest. Such outliers may produce spurious results; disproportionately affecting means, variances, and other statistics used in the meta-analysis; hence the need for outlier analysis. The distribution of effect sizes was analyzed and outliers were identified (Hedges & Olkin, 1985 cited in Lipsey & Wilson, 2001). Once the degree of dispersion for existing outliers was determined and their effect on the summary statistics assessed, appropriate procedures for handling the outliers was addressed on a case-by-case basis. In general, the outlier was removed from further analysis. Potential reasons for the existence of outliers in a meta-analysis include methodological error and poor validity of operational definitions (Lipsey & Wilson, 2001).

Outlier analyses, examining standardized effect sizes, were conducted prior to the meta-analysis and meta-regression analyses. To start, science and social studies results were removed from analyses. As assessments in some studies were run across multiple years and multiple subjects, it was felt that keeping results for a single subject—math—across multiple years was a more appropriate match to the present research purpose.

Once the remaining studies were deemed an appropriate match to research purpose for the current study, incremental outlier analyses using study as the unit of analysis was conducted, followed by the same analyses using substudy as the unit of analysis.

Table 2 provides results of the incremental outlier analysis with study as the unit of analysis. For accompanying histograms, see Appendix J.

Results from the Bouck and Yadav (2008) study had extreme values for both students with disabilities and their typically developing peers. Tests of normality were statistically significant (p <0.001) which indicated non-normality. These values were removed from the data and the analysis was repeated. Results from the second iteration showed extreme values for students with disabilities and typically developing students for the Lewandowski and Lovett (2008) study. Again, tests of normality were statistically significant (p <0.001), thus the data from this study were removed and the analysis was repeated. A final test showed extreme values for students with disabilities for the Lesaux, Pearson, and Siegel (2006) study. With statistically significant tests of normality (p = 0.005), results from this study for both students with disabilities and students with typical development were removed and a final analysis was completed. While tests for normality were not significant (p > 0.005), and the assumption of normality was not rejected, it was felt that it was not necessary to remove this study as only students with disabilities, and not their typically developing peers displayed extreme values.

Table 2: *Outlier Analysis for Effect Size Estimates - Study as the Unit of Analysis*

Study	Group	ES[a]	Issues	Result
		Analysis 1		
Bouck &Yadav (2008)	students w/o disabilities[b]	11.63	skewness, kurtosis, & normality	removed
Bouck &Yadav (2008)	students w/ disabilities	3.30	skewness, kurtosis, & normality	removed
Lewandowski & Lovett (2008)	students w/o disabilities[b]	1.87		
Lesaux, Pearson, & Siegel (2006)	students w/ disabilities	1.43		
		Analysis 2		
Lewandowski & Lovett (2008)	students w/o disabilities[b]	1.87	skewness, kurtosis, & normality	removed
Lesaux, Pearson, & Siegel (2006)	students w/ disabilities	1.43		
Lewandowski & Lovett (2008)	students w/ disabilities	1.30	skewness, kurtosis, & normality	removed
		Analysis 3		
Lesaux, Pearson, & Siegel (2006)	students w/ disabilities	1.43	skewness, kurtosis, & normality	retained

[a] ES is Hedges' g effect size estimate
[b] students w/o disabilities refers to typically developing students

The incremental outlier analysis, with substudy as the unit of analysis, is provided in Table 3. For accompanying histograms, see Appendix J.

While it was expected that, given the addition of substudy, there would be a different set of outliers, this was not the case. The same iterative analyses were run, with the same results.

Studies with extreme values (Bouck &Yadav, 2008; Lewandowski & Lovett, 2008); i.e., those not in line with information from other primary studies listed in Table 2 and Table 3, were removed from further analyses.

Table 3: *Outlier Analysis for Effect Size Estimates - Substudy as the Unit of Analysis*

Study	**Group**	**ES**[a]	**Issues**	**Result**
	Analysis 1			
Bouck &Yadav (2008)	students w/o disabilities[b]	11.63	skewness, kurtosis, & normality	removed
Bouck &Yadav (2008)	students w/ disabilities	3.30	skewness, kurtosis, & normality	removed
	Analysis 2			
Lewandowski & Lovett (2008)	students w/o disabilities[b]	1.87	skewness, kurtosis, & normality	removed
Lesaux, Pearson, & Siegel (2006)	students w/ disabilities	1.43		
Lewandowski & Lovett (2008)	students w/ disabilities	1.30	skewness, kurtosis, & normality	removed
	Analysis 3			
Lesaux, Pearson, & Siegel (2006)	students w/ disabilities	1.43	skewness, kurtosis, & normality	retained
Meloy, Deville, & Frisbie (2002)	students w/ disabilities	1.20		

[a] ES is Hedges' g effect size estimate
[b] students w/o disabilities refers to typically developing students

While it might be argued that the larger effect sizes seen for the outlier studies were the result of a good match between the study participants and the accommodation under investigation, this did not appear to be the case as there were no discernable differences between these studies and those that were included in the meta-analyses. It was felt that removing these specific studies, particularly as no relevant differences between ‘outlier’ and ‘included’ studies were seen, provided a more conservative estimate of the mean effect for testing accommodations. Thus, in the event statistically significant mean effects were found, the use of a more conservative estimate was thought to provide a better approximation of the mean effects than potentially overestimating these effects.

Analysis of the homogeneity of variance and the distribution of effect size.

Examination of the homogeneity of the effect size distribution; i.e., the distribution of primary effect sizes around the mean effect size, is one of the next steps in meta-analytic research. With a homogenous distribution, the amount by which the effect size distribution differs from that of the population is equal to that expected by sampling error. Rejection of homogeneity of variance suggests the variability of the effect sizes is larger than sampling error and, therefore, "each effect size does not estimate a common population mean" (Lipsey & Wilson, 2001, p. 115). The Q statistic was employed to test the homogeneity of the distribution of primary effect sizes.

The Q statistic is distributed as a chi-square with *k* – 1 degrees of freedom where *k* is equal to the number of effect sizes used in the meta-analysis, *ES* is the individual effect size for *i* = 1 through *k* effect sizes, and $\overline{ES}$ is the weighted effect size over the *k* effects.

$$Q = \sum \omega_i \left(ES_i - \overline{ES}_i \right)^2 \tag{2.4}$$

where ω_i is the individual weight for ES_i, ES_i is the individual effect size for *i* = 1,…, *k* effect sizes, and $\overline{ES}_i$ is the weighted mean effect size over *k* effect sizes.

From a statistical perspective, the Q statistic examines the assumption of a fixed-effects model, with a significant Q indicating a heterogeneous distribution, challenging the fixed-effects model. Conversely, a non-significant Q may not be indicative of a fixed-effects model. For example, if there is a small number of primary studies and each examines a small number of subjects, there may not be enough statistical power to be able to reject the homogeneity of variance assumption (Lipsey & Wilson, 2001; Morton et al., 2004).

Sources of variance associated with the distribution of the primary study effect sizes were expected to be randomly distributed. This led to the adoption of the random-effects, or unconditional, model. The random-effects model differs along two dimensions; study characteristics and the effect size parameter. That is, effect size variation is explained by a random component as well as by subject-level sampling error. Hedges (1994) explained that “studies in the study sample … differ from those in the universe as a consequence of the sampling of people into the groups of the study” (p. 31) with “the study sample (and their effect size parameters) differ[ing] from those in the universe by as much as might be expected as a consequence of drawing a sample from a population” (p. 31) such that there is “variation of observed effect sizes about their respective effect size parameters” (p. 31), referred to as study-level and subject-level random variability by Lipsey and Wilson (2001).

The assumptions of the fixed-effects model, whereby random error found in the primary studies was due to subject-level sampling error alone and effect sizes were presumed to estimate the consequent population effect, was considered untenable on theoretical grounds. The primary analyses forming the basis of the present meta-analysis were considered to be part of a larger universe of primary analyses that do not have a common effect size for the population of potential eligible studies. That is, the observed effects sizes were expected to have both study-level and subject-level sampling error variability. As well, the assumptions necessary for the fixed-effects model were difficult to meet.

While potentially tenable, the mixed-effects model, which assumes that variance not explainable by sampling error can be attributed to both random and systematic sources of variance, was not employed. It was believed that regardless of how much attention was devoted to the design of the coding tools, allowing for the quantification of potential moderator variables, the coding conducted would not be able to capture the information in enough detail to meet the assumptions necessary to conclude differences were truly systematic sources of variance. Additionally, the mixed-effect model allows for the use of a random-effects model to combine the studies within each subgroup; i.e., students with disabilities and typically developing students, and a fixed-effects model to combine the subgroups to yield the overall mean effect size. As the research purpose was to compare subgroups, and not aggregate these two groups, use of the mixed-effects model was not warranted.

Due to the nature of the design of the present meta-analysis, effect sizes found for the primary studies examined were derived from a non-uniform set of sample characteristics; i.e., assessment accommodations for students with disabilities. Therefore, homogeneity of variance of the primary effect sizes was not expected due to the degree of differences between both assessment accommodations and students with disabilities. This led to the use of the random-effects model in the final analysis examining the efficacy of assessment accommodations and their delivery to students with disabilities as opposed to their typically developing peers.

When coding the data, several studies using a repeated measures design did not contain test score correlation, necessary for effect size estimation, between the non-accommodated and accommodated conditions. For studies missing these correlations, the correlations were estimated using information from test websites, searching the online version of the Mental Measurements Yearbook, and other research studies with similar tests (i.e., for the same age group assessing the same test content), frequently using test-retest reliability as an approximation of this value for the measures in question. Both Borenstein et al. (2009) and Lipsey and Wilson (2001) have mentioned this issue noting that using estimates, particularly test-retest reliability scores, "affects the confidence interval around the mean effect size thus caution should be used in interpreting the confidence interval" (Lipsey & Wilson , 2001, p. 43). Sensitivity analyses were performed, see 'Sensitivity analyses,' examining differences between studies using a repeated measures design and those using an independent groups design to ensure that the using these estimates were not drastically different.

Some studies using counterbalancing provided different results for test and/or order of condition results. In these cases, all data provided in the study were included in the analyses. While it was expected that there may be issues with some of the study variables; particularly as tests used in counterbalanced designs might not be parallel or the order of administration of the condition might affect the results; the data were included in the meta-analysis as they were still thought to provide legitimate evidence with respect to the research hypotheses posed.

Both boost and differential boost/interaction study data were combined in the analyses used to answer the hypotheses posed by the current research. Borenstein et al. (2009) point to issues of combining data from studies using different designs, as there may be substantive differentiation as well; this was not suspected to be an issue for the present study. There were several instances in the primary research (see Abedi et al., 2010; Johnson, 2000; Kosciolek & Ysseldyke, 2000; Schnirman, 2005; and Walz, Albus, Thompson, & Thurlow, 2000) where the same data set was used to answer questions regarding the efficacy of accommodations for students with disabilities and whether or not these accommodations were differentially effective for students with disabilities as compared to their typically developing peers. Similarly, meta-analyses conducted by Elbaum (2006) and Gregg and Nelson (2012) included results from primary research for both boost and differential boost/interaction research approaches.

Data from primary studies using repeated measures and independent group designs were combined in the analyses conducted. While this is not an issue "from a statistical perspective [as] the effect size … has the same meaning regardless of the study design" (Borenstein et al., 2009, p. 25), there may be issues regarding the focus of the studies and the effect sizes. Morris and DeShon (2002) note that the

> …IG [independent groups] focus of research [is] on differences across alternative treatments using raw score metric while RM [repeated measures] focus of research [is] on individual change using change score metric (p. 110)

and that "[t]he use of change score metric will often produce larger effect sizes than raw score metric" (p. 110). Still Borenstein et al. (2009) point out that "we need to assume that the studies are functionally similar in all other important respects" (p. 361).

With respect to the current research work, it was felt that the benefits of combining the different designs based on substantive grounds, and use of *Comprehensive Meta-Analysis V.2.2.050* to calculate and appropriately weight the different studies included, provided information that would not be fully addressed examining the results based on the two different research designs. Sensitivity analyses examining the differences between the results for the aggregate versus the disaggregated studies provided useful information to make certain that there were not drastic differences between estimates for the repeated measures, independent groups, and aggregated analyses (see 'Sensitivity analysis').

Sensitivity analysis.

Table 4 provides a comparison of the mean effect size estimates for the random-effects model for the two different research designs, repeated measures and independent groups, to the mean effect size estimates when combining both research designs.

The mean effect size estimates comparing students with disabilities to their typically developing peers for primary studies, using a repeated measures design ($\overline{ES}$ = 0.31 for students with disabilities; $\overline{ES}$ = 0.17 for typically developing students) or an independent groups design ($\overline{ES}$ = 0.26 for students with disabilities; $\overline{ES}$ = 0.15 for typically developing students), as compared to the combination of both repeated measures and independent groups primary studies ($\overline{ES}$ = 0.30 for students with disabilities; $\overline{ES}$ = 0.17 for typically developing students), are extremely similar. Further, standard errors and confidence intervals were not considered very different. However, there was a non-significant mean effect size estimate for typically developing students for the independent groups research design. This is, most likely, to be expected given the smaller number of primary studies constituting the mean effect size estimate.

This sensitivity analysis provided evidence for combining primary study information for both repeated measures and independent groups research designs when answering the first research hypothesis posed by the current study.

Table 4: *Sensitivity Analysis for Research Hypothesis 1 - $\overline{ES}$ Estimates, Confidence Intervals, & Significance*

		Mean effect size & 95% confidence interval for Hedges' g				
Comparison group	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p*(ES)**
		Combined Studies (random-effects model)				
students w/ disabilities	62	0.30	0.04	0.21	0.38	< 0.001
students w/o disabilities[b]	57	0.17	0.03	0.11	0.22	< 0.001
		Repeated Measures Designs (random-effects model)				
students w/ disabilities	48	0.31	0.05	0.22	0.41	< 0.001
students w/o disabilities[b]	46	0.17	0.03	0.11	0.23	< 0.001
		Independent Groups Designs (random-effects model)				
students w/ disabilities	14	0.26	0.12	0.02	0.50	0.033
students w/o disabilities[b]	11	0.15	0.12	-0.08	0.38	0.193

[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

[b] students w/o disabilities refers to typically developing students

Sensitivity analyses for research hypothesis 2 are displayed in Table 5. The mean effect size estimates for the random-effects model, when combining both research designs, are compared to the mean effect size estimates for the two different research designs; repeated measures and independent groups.

As can be seen, the mean effect size estimates comparing the four different categories of accommodations–presentation, response, setting, and timing/scheduling–are similar for presentation and timing/scheduling accommodations for the repeated measures research design ($\overline{ES}$ = 0.19 for presentation; $\overline{ES}$ = 0.47 for timing/scheduling) as compared with the combination of repeated measures and independent groups research designs ($\overline{ES}$ = 0.22 for presentation; $\overline{ES}$ = 0.47 for timing/scheduling). The same cannot be said for the independent groups research design as the mean effect size for presentation, $\overline{ES}$ = 0.39 is larger, albeit still within the small range (Cohen, 1992), and timing/scheduling, $\overline{ES}$ = -0.04 is smaller. It must be noted that there is only one timing/scheduling study for the independent groups research design, rendering sensitivity analyses for this comparison moot. As there are so few primary studies for either response or setting accommodation categories, sensitivity analysis was not considered relevant. Additionally, these two accommodation categories were not subject to intensive meta-analytic scrutiny or closely examined in the meta-regression analyses.

Again, evidence for combining primary study information to answer the second research hypothesis under investigation, for both repeated measures and independent groups research designs, albeit only for presentation and response assessment accommodations, is supported by the sensitivity analysis.

Table 5: *Sensitivity Analysis for Research Hypothesis 2 - $\overline{ES}$ Estimates, Confidence Intervals, & Significance*

		Mean effect size & 95% confidence interval for Hedges' g				
Type of Accommodation	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p*(ES)**
		Combined Studies (random-effects model)				
Presentation	41	0.22	0.06	0.12	0.33	< 0.001
Response	3	0.24	0.38	-0.50	0.98	0.525
Setting	1	0.32	0.17	-0.02	0.66	0.061
Timing-Scheduling	17	0.47	0.09	0.30	0.64	< 0.001
		Repeated Measures Designs (random-effects model)				
Presentation	30	0.19	0.06	0.07	0.31	0.002
Response	1	1.14	0.17	0.80	1.48	< 0.001
Setting	1	0.32	0.17	-0.02	0.66	0.061
Timing-Scheduling	16	0.48	0.09	0.31	0.65	< 0.001
		Independent Groups Designs (random-effects model)				
Presentation	11	0.39	0.15	0.09	0.70	0.011
Response	2	-0.19	0.08	-0.35	-0.03	0.021
Timing-Scheduling	1	-0.04	0.41	-0.84	0.77	0.931

[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

[b] students w/o disabilities refers to typically developing students

Publication bias analysis.

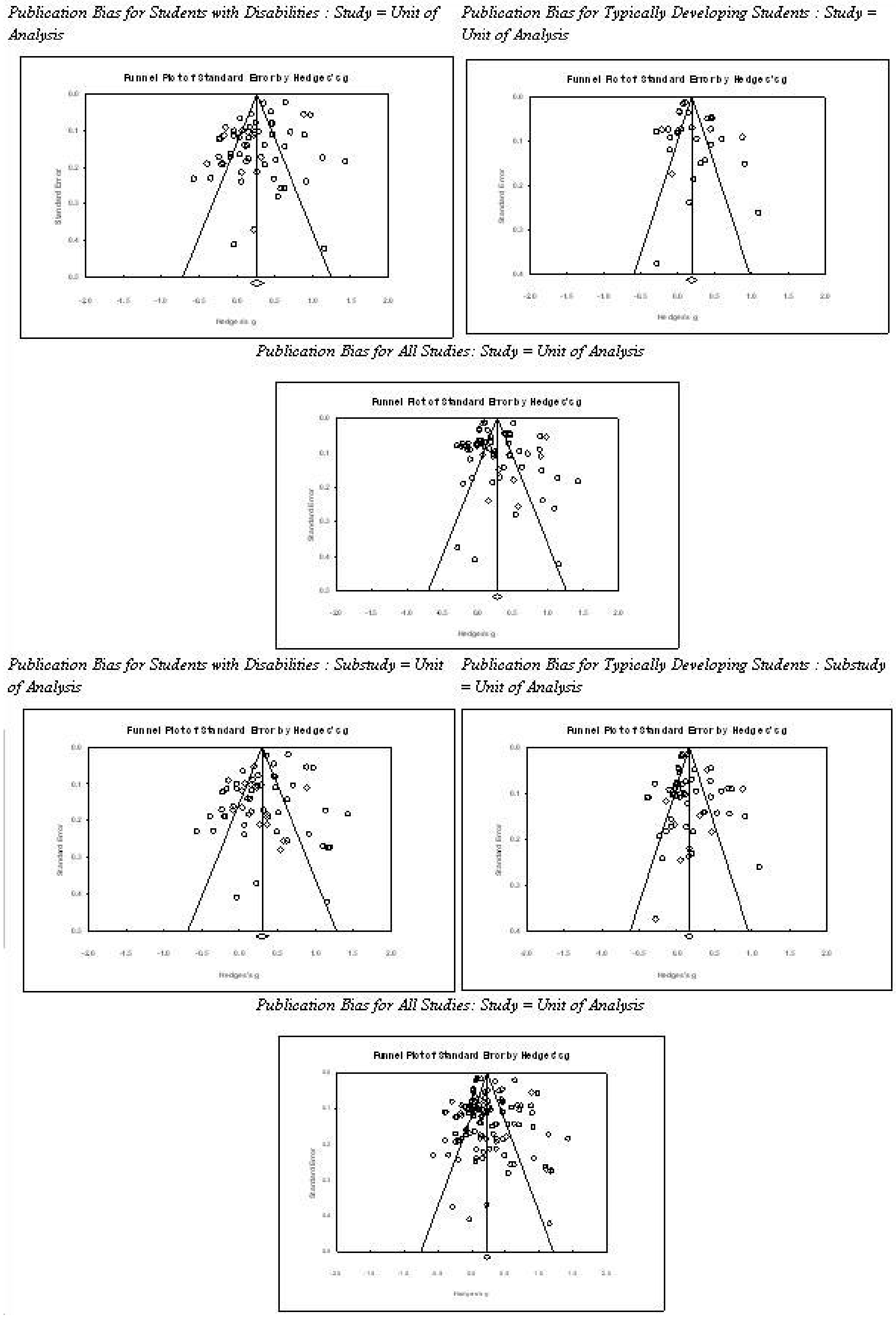
Publication bias was investigated, as it is generally held that non-significant research results are more likely to go unreported than those for studies with significant research results.

To obtain a sense of the data, weights used for the random-effects model were plotted against effect size estimates. Data with study and substudy as the unit of analysis for three different groupings were plotted; all studies included in the meta-analysis, studies with information for students with disabilities, and studies with information for students with typical development (see Appendix K). A visual examination of the weights indicated that there were no obvious patterns, or shifts to the right of the mean effect size estimates, indicating bias (Borenstein et al., 2009).

Funnel plots for the same groupings, study and substudy as the unit of analysis for all studies, studies with information for students with disabilities, and studies with information for typically developing students are displayed in Figure 2.

With effect sizes plotted against the x-axis and standard errors plotted against the y-axis we expect to see larger studies at the top, clustered about the mean effect size, with smaller studies at the bottom of the graph spread across a wider set of values. When publication bias is present, we expect to see symmetry at the top of the graph, some studies missing in the middle of the graph, and an even larger amount of studies missing at the bottom of the graph (Borenstein et al., 2009). Inspection of the graphs in Figure 2 does not reveal shapes that would be expected in the absence of substantial publication bias. Rather, the graphs in Figure 2 appear to indicate some amount of publication bias for each of the different groupings; data with study and substudy as the unit of analysis for all studies, studies with information for students with disabilities, and studies with information for typically developing students; albeit the appearance of publication bias is somewhat less for groupings for studies with information for students with disabilities than those for typically developing students. This is a bit perplexing, as there were more smaller studies (n = 10 through 100) than there were larger studies. It is expected that, with the absence of much larger studies due to the nature of the population under investigation, students with disabilities, studies which would generally be considered 'small' are being considered 'large' in these plots.

With the examination of the study weights plotted against effect size estimates, knowledge gained from prior research syntheses indicating publication of several studies that do not have significant results, and knowledge of the total number of research participants in each study the possibility of an issue with publication bias, while worrisome, was not considered an impediment to the current research study. Additionally, results for the Classic fail-safe N (Rosenthal, in Borenstein et al., 2009) suggest that, using study as the unit of analysis, 7517 studies would be required to nullify any effects found for students with disabilities and 1740 studies would be required to nullify any effects found for typically developing students. With substudy as the unit of analysis 8788 studies for students with disabilities and 2984 studies for typically developing students would be required to nullify any effects found. Further Duvall and Tweedie's (2000, cited in Borenstein et al. (2009)) Trim and Fill, an iterative method for imputing values to determine where missing studies are likely to fall, adding the values to the analysis, then re-computing the combined effect to fill the funnel plot for the left side and/or the right side, suggest that no studies are missing for students with disabilities and their typically developing peers for both study and substudy as the unit of analysis.


Publication Bias for Students with Disabilities : Study = Unit of Analysis

Publication Bias for Typically Developing Students : Study = Unit of Analysis

Publication Bias for All Studies: Study = Unit of Analysis

Publication Bias for Students with Disabilities : Substudy = Unit of Analysis

Publication Bias for Typically Developing Students : Substudy = Unit of Analysis

Publication Bias for All Studies: Study = Unit of Analysis

Figure 2: Publication Bias for the Random-Effects Model

Meta-regression

Rationale for meta-regression.

Thompson and Higgins (2002) assert that, in contrast to meta-analysis, "meta-regression aims to relate the size of the effect to one or more characteristics of the studies involved" (p. 1559). With respect to the present research; for example, certain assessment accommodations may improve test scores of students with disabilities, a single meta-regression analysis can be used to scrutinize research relating to construct-irrelevant variance for multiple assessment accommodations across multiple primary research studies. These explorations into the sources of heterogeneity provide potential scientific value such that meta-regression is becoming a more widely used statistical technique (Morton et al., 2004; Thompson & Higgins, 2002).

In an effort to understand which types of assessment accommodations remove construct-irrelevant variance from the test scores of students with disabilities, the present research study employed meta-regression analyses. Additionally, it was hoped that meta-regression analyses would aid in understanding how much assessment scores for students with disabilities would improve once the construct-irrelevant variance was removed. Meta-regression was selected as it addresses a common problem seen in meta-analysis: the lack of integration of effects of multiple related predictors to yield a summary of overall prediction. As such, meta-regression allowed for the integration of effects of multiple, and possibly related, predictors to provide an overall estimation of the most effective assessment accommodations. With the ability to specify fixed-effects, random-effects, and mixed-effects models, this researcher was able to estimate the likelihood of effect generalization.

Statistical methods of analysis.

Meta-regression allowed for the examination of the extent to which a particular covariate (moderator, effect modifier), with defined values for each primary study under consideration, explained heterogeneity between the primary studies under investigation (Thompson & Sharp, 1999). Thompson and Higgins (2002) suggest that it is "easiest to think of meta-regression in the context of a continuous covariate" although they note that "[h]eterogeneity is … often addressed in practice by subgrouping [studies] with different characteristics" (p. 1563), where the subgroup analysis corresponds to the use of a categorical study-level covariate in the meta-regression.

A variety of meta-regression analytic techniques exist, with "methods differ[ing] in a number of respects, including how they allow for residual heterogeneity, that is, heterogeneity which remains unexplained by the covariate" (Thompson & Sharp, 1999, p. 2693). Common meta-regression methods include fixed-effects models, random-effects models, control rate models, and Bayesian and/or hierarchical models. It should be noted that estimation of the residual between-study variations is generally seen as problematic, with different researchers advocating a variety of different estimates such as empirical Bayes estimation and restricted maximum likelihood estimation (REML). Meta-regression models may be employed with or without the inclusion of moderators.

"The outcome (or dependent) variable in a meta-regression analysis is usually a summary statistic … [which is] assumed to be the true variance" (Thompson & Higgins, 2002, p. 1563). This assumption is not valid when there are a *small* number of studies included in the meta-regression analyses. The outcome variable in the present study was the effect size estimate, or test accommodation, for each included study. Effect sizes with positive values were interpreted as showing that the assessment accommodation had a positive impact, while those with negative values were seen as indicating a negative impact.

The fixed-effects meta-regression model uses logistic regression, often weighted, with moderators at the study or study group level (i.e., students with disabilities and typically developing peers represent two study groups). Random-effects meta-regression models generally regress the log odds ratio on the regression intercept and study-level moderators. Random-effects meta-regression models include a random study effect to take between-study variation into account. Control rate meta-regression uses the outcome for the control group(s) from studies as the single covariate for the model. The control rate is used as a proxy for covariate differences between the studies. The Bayesian hierarchical model may also be used as a meta-regression model, where Bayesian estimation approaches–prior probability and likelihood–are used to compute a posterior probability and then used to assess heterogeneity.

A meta-regression can use either a linear or logistic regression model where the unit of analysis, similar to meta-analysis, is an individual study. Two common questions answered by meta-regression relate to “estimating the treatment effect controlling for differences across studies and determining which study-level moderators account for the heterogeneity” (Morton et al., 2004, p. 10).

Morton et al. (2004) suggest that heterogeneity can be broken into two components: (i) study incomparability, where the differences among the studies relate to the variables being studied, and (ii) design incomparability, where the differences seen are due to the designs of the studies not the study variables. Study incomparability is beyond the control of the researcher, who must then decide whether to focus on a particular variable (e.g., a particular treatment may work differently for a specific population or subpopulation) or the "group" of variables (e.g., focus on a specific subgroup to reduce incomparability). Design incomparability is under the control of the researcher. Morton et al. (2004) recommend that "[r]esearchers may actually plan differences across studies to induce heterogeneity and increase generalizability, [as] assessing and understanding such differences is a strength of systematic reviews" (p. 9).

Thompson and Higgins (2002) note that using meta-regression techniques is appropriate even when initial tests of heterogeneity for effect sizes is not significant. Non-significant results do not reliably indicate that there is a lack of heterogeneity, as the tests used generally have low statistical power. Further, Thompson and Higgins state that it is "not reasonable to assume that all of the heterogeneity is explained" (p. 1562) and that "'residual heterogeneity' must be acknowledged in the statistical analysis" (p. 1562), generally using a random-effects rather than a fixed-effects meta-regression. "Ignoring residual heterogeneity … underestimate[s] the [standard errors], SEs, of the regression coefficients, … overstat[ing] the importance of the covariate" (Thompson & Sharp, 1999, p. 2705). It is important to use appropriate standard errors to calculate a prediction interval around the estimated regression line (Thompson & Sharp, 1999).

For random-effects meta-regression analyses, Thompson and Higgins (2002) suggest weighting the regression such that "more precise studies have more influence in the analysis" (p. 1562) with each study weight "equal[ing] … the inverse of the sum of the [within-study] variance and the residual [between-study] variance" (p. 1562). Specification of whether the weights were taken *equal* to the inverse variances (for a fixed-effects model) or *proportional* to the inverse variance (for a multiplicative, not additive, adjustment for residual heterogeneity) is a necessary component in random-effects meta-regression.

As random-effects meta-regression "… estimates the mean of a distribution of effects across studies" (Thompson & Higgins, 2002, p. 1562) and generates wider confidence intervals for the regressions coefficients, it was considered the appropriate model to use for the present research study. Wilson's meta-regression macro (2005: metareg.sps), obtained as a free download, was used. This macro allows for both fixed- and random-effect model estimation, using an inverse variance weighted generalized least squares regression with full-information maximum likelihood estimation. The current study employed the random-effects model estimated *via* iterative maximum likelihood as the random-effects model provides a more conservative estimate of the variance accounted for than the fixed-effects model.

Disability classification and assessment accommodation were re-categorized for the meta-regression analyses. The disability classification was aggregated to form two groups, students requiring special education services and students with learning disabilities. The learning disabilities category included students with learning disabilities in reading or learning disabilities in reading and math. Assessment accommodation was aggregated forming two categories: presentation and timing/scheduling. Segmented text and read aloud constituted the presentation category and extended time constituted the timing/scheduling category. When there were more than two levels for categorical variables, indicator variables, based on the codebook, were created. Separate meta-regression analyses were performed for four different conceptual groupings. The conceptual groupings were based on categories that were (i) thought to represent different substantive areas and (ii) either controllable or less controllable by the primary researcher(s). The groupings, or variable sets, represented different areas of potential residual variance that was not explained by sampling error alone. It was felt that the variables, collectively, could help provide a more interpretable evaluation of the results for this potential residual variance. Separate meta-regression analyses were run for the following conceptual groupings:

- Researcher-manipulated variable directed towards reducing construct-irrelevant variance for students with disabilities; i.e., assessment accommodation
- Population description; i.e., descriptions for students with disabilities including grade level
- Assessment description; i.e., assessment content and assessment format
- Dissemination; i.e., type of publication and publication year

It was hypothesized that assessment characteristics; i.e., test content and test format, population description; i.e., disability type and grade level, and dissemination; i.e., type of publication and publication year, have effects on test score change. The researcher-manipulated variable, assessment accommodations, was also expected to have effects on test score change.

Performing separate meta-regressions at this stage allowed for assessment of differential predictions of the effectiveness of assessment accommodations, represented by the effect sizes for each included study. Type of test accommodation was expected to have an effect on efficacy of test accommodations. A meta-regression using predictors from all the variable groupings was conducted following the initial set of analyses.

Meta-regression limitations.

Researchers in the area of meta-regression have noted a number of limitations impacting the results for studies using this analytic technique. These limitations include bias by confounding, aggregation bias, low within-study variance as opposed to across-study variance, clear separation of whether the data delineate (i) within-study, (ii) across-study, or (iii) a mixture of between- and across-study information, dependencies in measurement errors and measurement errors in the covariate, over-inclusion of study characteristics limiting the available degrees of freedom for the meta-regression, and the collinearity of the meta-regression moderators.

In a meta-regression observational associations, or differing characteristics, across the studies under examination can be highly correlated as "… meta-regression is across [studies] and does not have the benefit of randomization to underpin a causal interpretation" (Thompson & Higgins, 2002, pp. 1563 – 1564) thus displaying "bias by confounding" (p. 1564). As studies used in the present meta-regression analysis do not use observational associations, this was not a limitation. It should be noted that many of the primary studies included in the meta-regression did not have the benefit of randomization due to the nature of the subjects under investigation; specifically, participants could not be randomly assigned to disability type; i.e., hearing impaired or autistic. However, participants were randomly assigned to accommodation conditions; i.e., not accommodated or accommodated, for the independent groups research design and there was counterbalancing for condition for the repeated measures research design. At the same time, it should be noted that this issue is not the same as *bias by confounding*.

When few subjects are included in the primary studies that are included in the meta-regression, and their averages are used to describe demographics of some of the study variables, such as age, attenuation by measurement error becomes a limiting factor. If the averages across studies are not the same as the averages found within individual studies there may be *aggregation bias* which is “confounding at either the [study] level (biasing the relationship across [studies]) or at the individual level (biasing the relationship within [studies])” (Thompson & Higgins, pp. 1564–1565). Simply put, aggregation bias generally occurs as meta-regression does not include underlying subject-level variation because primary studies are the units of analysis. Aggregation bias is variously known as ecological bias, ecological confounding, or the ecological fallacy. As most of the primary studies being analyzed for this research work did not include subject-level variation, careful attention needed to be paid to aggregation bias, and appropriate correction factors for attenuation by measurement error employed wherever possible.

Meta-regression results can be used to examine any measureable study characteristic. When study characteristics do not exhibit high variability across the studies as compared to the variability of results found within each study, meta-regression outcomes are more difficult to interpret, as there is little ability to discriminate between the studies under scrutiny. It should be noted that such statistically non-significant relationships "should not be equated to the absence of true relationships" (Thompson & Higgins, 2002, p. 1565) for effects or differences between effects. Care was taken, in the present meta-regression analysis, when examining within-study and across-study variability, noting instances where there were issues of low variability within studies.

Different outcomes from meta-regression analysis may be obtained if researchers are able to use information within studies, provide more precise data than aggregate information, and remove issues of aggregation bias. This potential confounding across studies stresses the necessity of clearly separating whether the data represent within-study, across-study or a mixture of between- and across-study information. For example, if one of the study characteristics examined relates to the gender of the subjects under study; with some studies reporting results for males and females, some studies reporting results for males only, and some studies reporting results for females only; it would be very difficult to interpret the results if the meta-regression did not clearly detail how the data were entered into the analysis/analyses. Coding for the present study included details as to whether the data delineated within-study, across-study, or a mixture of between- and across-study information.

Conventional meta-regression analysis is also flawed by issues with dependencies in measurement errors (regression to the mean for students with learning disabilities) and issues of measurement error in the covariate appearing in the treatment effect (dependent variable) “causing an artifactual negative association” (Thompson & Higgins, 2002, p. 1566). Thompson and Higgins recommended handling these limitations by using more complex meta-regression models to address measurement error dependencies.

“Meta-regression requires the estimated treatment effect, its variance, and covariate values for each [study] in the systematic review” (Thompson & Higgins, 2002, p. 1566) such that when one or more of these data points are unavailable for a study it cannot be included in the analysis, limiting the number of studies, and potentially biasing the results. This also results in issues with the degrees of freedom available for the meta-regression analyses (Morton et al., 2004). Morton et al. (2004) suggest that, as is common with most statistical methods, a larger number of studies and larger number of subjects per study can reduce bias with proper modeling. They note that failure to incorporate important moderators at either the study, or person, level can bias the results of a meta-analysis. Inasmuch as possible, this researcher strived to achieve a balance between under-inclusion and over-inclusion of study characteristics, especially those deemed potential moderators, in the coding in an effort to include as many studies as possible in the analysis.

There may be issues with collinearity of the moderators included in a meta-analysis. This makes it impossible to disentangle the effects of individual moderators. For example, in the present study each state used its own protocol for implementing assessment accommodations. While primary researchers were not bound by states' assessments accommodation protocols, these protocols may have influenced the assessment condition in the primary study. Careful coding of the moderators was used to minimize this issue. As well, the present meta-regression employed the results from the meta-analysis to inform decisions regarding prespecification of moderators included in the meta-regression analysis, *a priori*, with an eye to potential moderators, during the construction of the coding manual, codebook, and coding form.

Despite the importance of including moderators, models that include moderators that are aggregates of person-level characteristics rather than study characteristics can produce biased results. Morton et al. (2004) suggest further exploration of the underlying data to examine potential trade-offs between the biases of incorporating versus excluding an aggregated covariate.

Meta-regression methodological issues.

There are several methodological issues with meta-regression techniques that make meta-regression analyses prone to difficult interpretive problems (Thompson & Higgins, 2002). While some of these issues are interconnected with limitations to meta-regression analyses previously outlined, the issues presented in this section relate specifically to *interpretation* of meta-regression results.

According to Thompson and Higgins (2002) "[d]ata dredging is the main pitfall in reaching reliable conclusions form meta-regression" (p. 1559) and may result in false positive findings. Data dredging occurs when there are few studies with many possible study or subject characteristics which might explain heterogeneity, with multiple analyses undertaken in a *post hoc* manner, using each of the available characteristics such that "… any set of (k-1) non-linearly dependent [study]-level moderators will 'explain' all the heterogeneity between the results of k trials" (Thompson & Sharp, 1999, p. 2706). "[Data dredging] can only be avoided by prespecification of covariates[, or moderators,] that will be investigated as potential sources of heterogeneity" (Thompson & Higgins, 2002, p. 1559). All potential moderators were specified during the initial proposal for the current study and refined during coding of the primary studies. Additional moderators were not specified after these processes.

Thompson and Sharp (1999) note that "near-collinearity of categorical variables describing trial characteristics can … be a problem" (p. 2706) which is often seen in practice.

Thompson and Higgins (2002) point out that it is

> … necessary to limit the number of moderators proposed for investigation again to protect against false positive conclusions. If multiple covariates [; i.e., moderators,] are of real scientific interest, false positive conclusions can be limited to a desired level by using a Bonferroni adjustment to the significance level for each covariate (p. 1567).

The researcher, while expecting that such protection might be required, did not find it necessary to use an adjustment, such as the Hochberg adjustment, to protect against false positive research conclusions as there were relatively few moderators under investigation.

In a random-effects meta-regression analysis both the within-study effect variances and residual between-study variances–heterogeneity not explained by moderators selected for the meta-regression analysis–need to be weighted (Thompson & Higgins, 2002). Issues of interpretation arise if neither or only one of these variances is weighted.

The results of observational studies are seen as harder to interpret than randomized trials [*true* experimental studies] for several reasons (Thompson & Higgins, 2002). For example, “[o]bservational studies are more variable in design than randomized trials thus heterogeneity in their results may reflect design differences rather than *true* diversity” (p. 1571). As well, issues of selection and other biases associated with observational studies will generally hinder the interpretation of meta-regressions. Thompson and Higgins also note that “[t]he variables adjusted for in statistical analyses to reduce confounding within studies are almost always different (or differently handled) in each study” (p. 1571). As well, the effects of publication bias in the available literature may be more extreme for observational studies.

Chapter Three

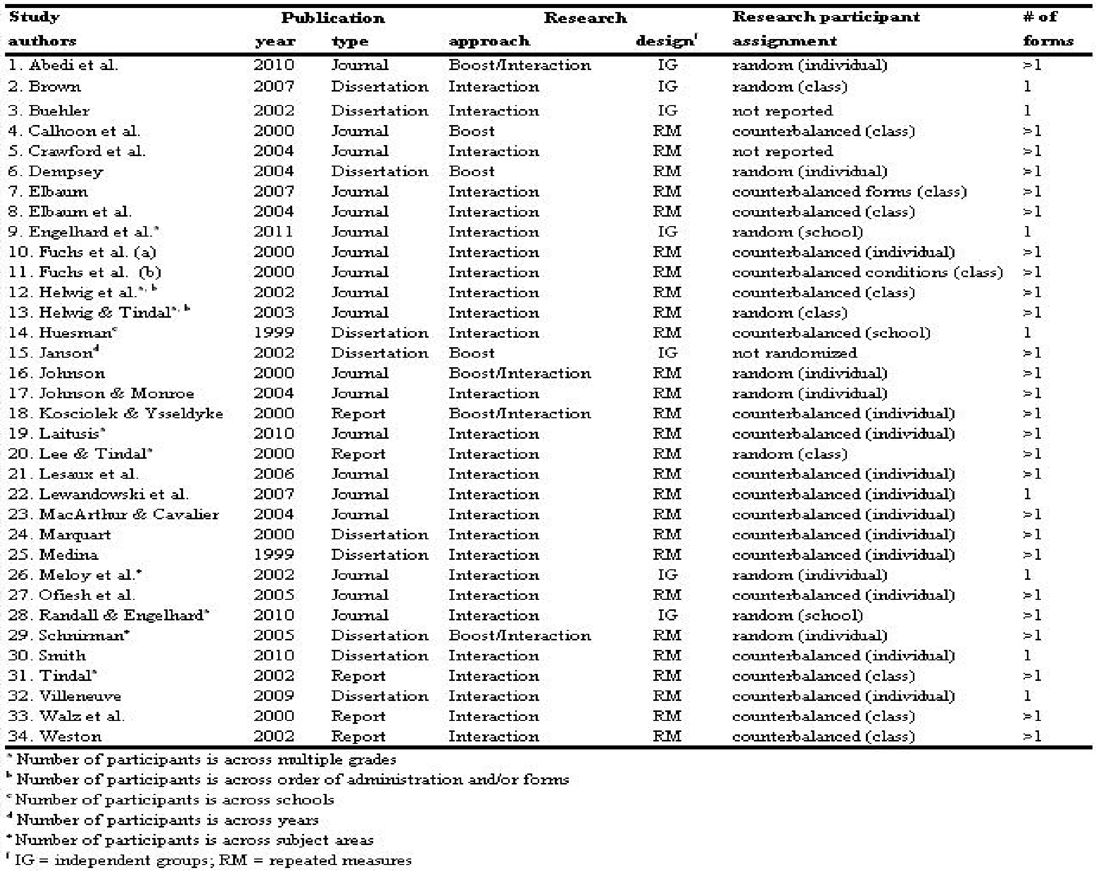
Results

Demographics for studies as the unit of analysis.

Thirty-four studies examining test accommodations for students with disabilities, from mid-1999 through mid-2011, were included in the meta-analyses and meta-regression analyses. Of the 34 studies, 3 were identified as boost studies, 27 were identified as differential boost studies, and 4 primary studies answered both boost and differential boost research questions. The four primary studies answering both a boost and differential boost research question employed the same dataset, using data collected for individuals with disabilities and their typically developing peers to answer the differential boost question, and then selecting the students with disabilities data subset to answer the boost question. The following five tables provide information on demographics for the primary studies included in the present analysis.

Table 6 provides information regarding publication, research approach, and research design. Study demographics included in Table 6 provide general descriptive variables regarding study design for the 34 studies included in the meta-analyses and meta-regressions. As can be seen in Table 6, the publication date for studies collected ranged from 1999 to 2011, with the bulk of the studies (8) being published in 2002. No studies with useable data were published in 2001, the year that NCLB was enacted, nor in 2008. Year of study publication does not appear to be related to the type of study conducted; that is, the year studies were published does not appear to be related to the type of publication, research approach, or research design used. The studies include journal articles (19), research reports (5), and dissertations (10). As was discussed in the preceding chapter, seven of the studies included in the analyses were published in alternate venues (see Appendix H for a complete list of duplicate studies). When selecting studies for inclusion it was thought that journal articles and dissertations were the most accessible sources of information, thus were more likely to be included in analyses than reports or conference proceedings. Note that primary studies either are prefaced with or followed by a numeral; e.g., 1, which is used as an identifier when referring to specific studies in the text, tables, or figures presented in this document.

Table 6: *Study Demographics - Publication and Research Information*



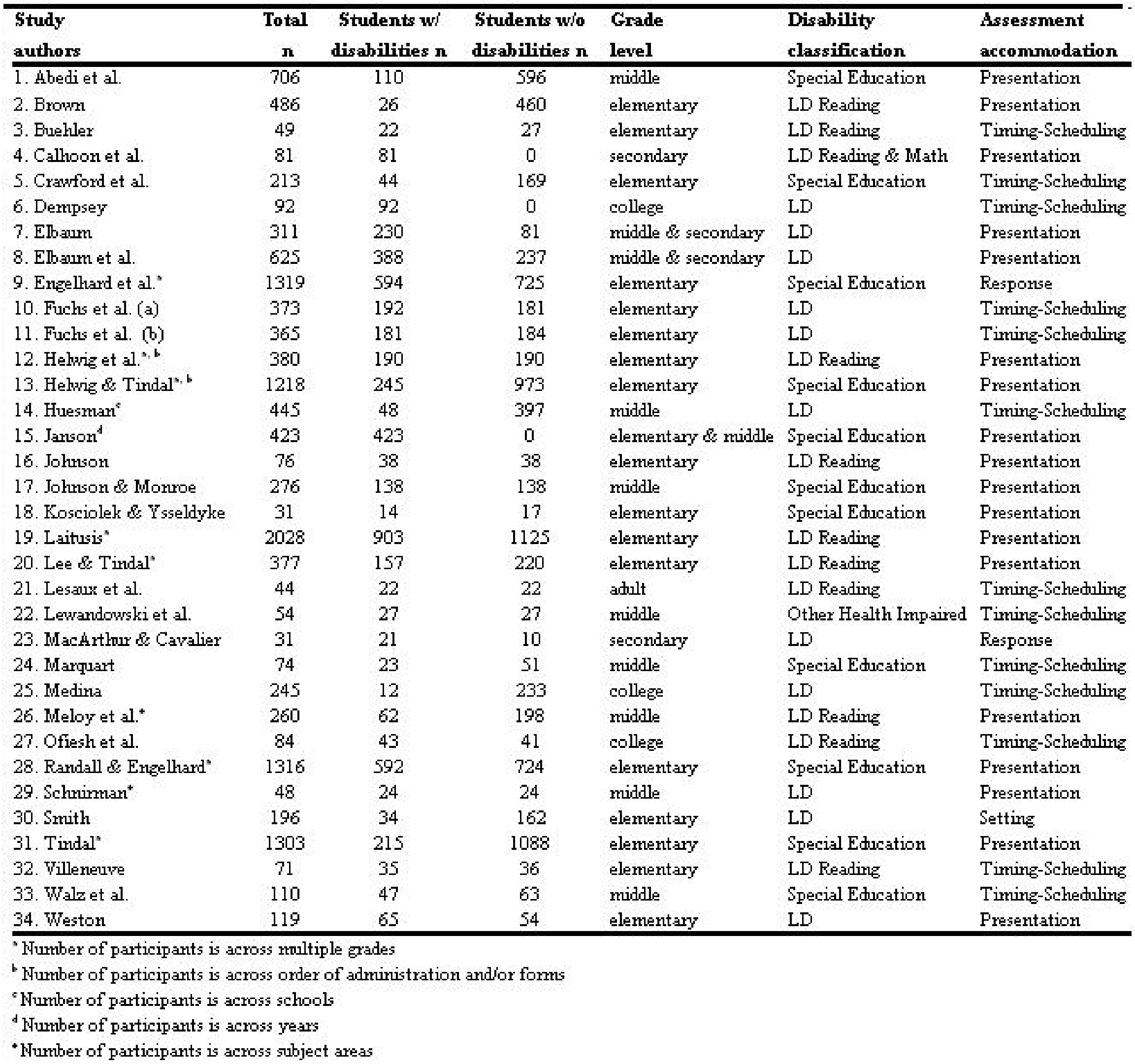
Study	**Publication**		**Research**		**Research participant**	**# of**
authors	**year**	**type**	**approach**	**design**[f]	**assignment**	**forms**
1. Abedi et al.	2010	Journal	Boost/Interaction	IG	random (individual)	>1
2. Brown	2007	Dissertation	Interaction	IG	random (class)	1
3. Buehler	2002	Dissertation	Interaction	IG	not reported	1
4. Calhoon et al.	2000	Journal	Boost	RM	counterbalanced (class)	>1
5. Crawford et al.	2004	Journal	Interaction	RM	not reported	>1
6. Dempsey	2004	Dissertation	Boost	RM	random (individual)	>1
7. Elbaum	2007	Journal	Interaction	RM	counterbalanced forms (class)	>1
8. Elbaum et al.	2004	Journal	Interaction	RM	counterbalanced (class)	>1
9. Engelhard et al.[a]	2011	Journal	Interaction	IG	random (school)	1
10. Fuchs et al. (a)	2000	Journal	Interaction	RM	counterbalanced (individual)	>1
11. Fuchs et al. (b)	2000	Journal	Interaction	RM	counterbalanced conditions (class)	>1
12. Helwig et al.[a, b]	2002	Journal	Interaction	RM	counterbalanced (class)	>1
13. Helwig & Tindal[a, b]	2003	Journal	Interaction	RM	random (class)	>1
14. Huesman[c]	1999	Dissertation	Interaction	RM	counterbalanced (school)	1
15. Janson[d]	2002	Dissertation	Boost	IG	not randomized	>1
16. Johnson	2000	Journal	Boost/Interaction	RM	random (individual)	>1
17. Johnson & Monroe	2004	Journal	Interaction	RM	random (individual)	>1
18. Kosciolek & Ysseldyke	2000	Report	Boost/Interaction	RM	counterbalanced (individual)	>1
19. Laitusis[a]	2010	Journal	Interaction	RM	counterbalanced (individual)	>1
20. Lee & Tindal[a]	2000	Report	Interaction	RM	random (class)	>1
21. Lesaux et al.	2006	Journal	Interaction	RM	counterbalanced (individual)	>1
22. Lewandowski et al.	2007	Journal	Interaction	RM	counterbalanced (individual)	1
23. MacArthur & Cavalier	2004	Journal	Interaction	RM	counterbalanced (individual)	>1
24. Marquart	2000	Dissertation	Interaction	RM	counterbalanced (individual)	>1
25. Medina	1999	Dissertation	Interaction	RM	counterbalanced (individual)	>1
26. Meloy et al.[e]	2002	Journal	Interaction	IG	random (individual)	1
27. Ofiesh et al.	2005	Journal	Interaction	RM	counterbalanced (individual)	>1
28. Randall & Engelhard[a]	2010	Journal	Interaction	IG	random (school)	>1
29. Schnirman[e]	2005	Dissertation	Boost/Interaction	RM	random (individual)	>1
30. Smith	2010	Dissertation	Interaction	RM	counterbalanced (individual)	1
31. Tindal[a]	2002	Report	Interaction	RM	counterbalanced (class)	>1
32. Villeneuve	2009	Dissertation	Interaction	RM	counterbalanced (individual)	1
33. Walz et al.	2000	Report	Interaction	RM	counterbalanced (class)	>1
34. Weston	2002	Report	Interaction	RM	counterbalanced (class)	>1

[a] Number of participants is across multiple grades
[b] Number of participants is across order of administration and/or forms
[c] Number of participants is across schools
[d] Number of participants is across years
[e] Number of participants is across subject areas
[f] IG = independent groups; RM = repeated measures

As can be seen in Table 6, 59% of the studies, 20 of 34, employed a repeated measures design with counterbalancing. Of the total number of studies, 23 used multiple forms of the assessment to measure differences between, or gains from, the non-accommodated and accommodated conditions. While six of the repeated measures studies used random assignment at the classroom or individual level, five of the seven studies employing an independent groups research design used random assignment at the school, classroom, or individual level. Three of the 34 studies did not report or use either counterbalancing or random assignment in their assignment of research participants. Two of the studies, one repeated measures and one independent groups study, did not include information regarding assignment of participants. Janson's (2002) dissertation work did not allow for randomization or counterbalancing of study participants. She was constrained by the data, an existing database with results from the Tennessee Comprehensive Assessment Program Achievement Test, used to answer her research question.

Table 7 provides information on primary study participant demographics and assessment accommodation. Study demographics included in Table 7 list information regarding participants in the primary research studies. The total number of participants included in the studies ranged from 31 (Kosciolek & Ysseldyke, 2000; MacArthur & Cavalier, 2004) to 2,028 (Laitusis, 2010). All studies with 377 participants or more (Lee & Tindal, 2000) were studies which involved a large group comparison between students with disabilities and their typically developing peers, or data were collected across multiple sites, grades, or years. Forty-one percent of all studies included in the analyses contained fewer than 120 total participants. The number of participants with disabilities ranged from 12 (Medina, 1999) to 903 (Laitusis, 2010), with 62% of these studies comprising fewer than 120 such participants. The number of typically developing peers included in interaction (differential boost) studies ranged from 10 (MacArthur & Cavalier, 2004) to 1,125 (Laitusis, 2010), with 47% of studies comprising fewer than 120 such participants.

Table 7: *Study Demographics - Participant Information*



Study authors	Total n	Students w/ disabilities n	Students w/o disabilities n	Grade level	Disability classification	Assessment accommodation
1. Abedi et al.	706	110	596	middle	Special Education	Presentation
2. Brown	486	26	460	elementary	LD Reading	Presentation
3. Buehler	49	22	27	elementary	LD Reading	Timing-Scheduling
4. Calhoon et al.	81	81	0	secondary	LD Reading & Math	Presentation
5. Crawford et al.	213	44	169	elementary	Special Education	Timing-Scheduling
6. Dempsey	92	92	0	college	LD	Timing-Scheduling
7. Elbaum	311	230	81	middle & secondary	LD	Presentation
8. Elbaum et al.	625	388	237	middle & secondary	LD	Presentation
9. Engelhard et al.[a]	1319	594	725	elementary	Special Education	Response
10. Fuchs et al. (a)	373	192	181	elementary	LD	Timing-Scheduling
11. Fuchs et al. (b)	365	181	184	elementary	LD	Timing-Scheduling
12. Helwig et al.[a, b]	380	190	190	elementary	LD Reading	Presentation
13. Helwig & Tindal[a, b]	1218	245	973	elementary	Special Education	Presentation
14. Huesman[c]	445	48	397	middle	LD	Timing-Scheduling
15. Janson[d]	423	423	0	elementary & middle	Special Education	Presentation
16. Johnson	76	38	38	elementary	LD Reading	Presentation
17. Johnson & Monroe	276	138	138	middle	Special Education	Presentation
18. Kosciolek & Ysseldyke	31	14	17	elementary	Special Education	Presentation
19. Laitusis[a]	2028	903	1125	elementary	LD Reading	Presentation
20. Lee & Tindal[a]	377	157	220	elementary	LD Reading	Presentation
21. Lesaux et al.	44	22	22	adult	LD Reading	Timing-Scheduling
22. Lewandowski et al.	54	27	27	middle	Other Health Impaired	Timing-Scheduling
23. MacArthur & Cavalier	31	21	10	secondary	LD	Response
24. Marquart	74	23	51	middle	Special Education	Timing-Scheduling
25. Medina	245	12	233	college	LD	Timing-Scheduling
26. Meloy et al.[*]	260	62	198	middle	LD Reading	Presentation
27. Ofiesh et al.	84	43	41	college	LD Reading	Timing-Scheduling
28. Randall & Engelhard[a]	1316	592	724	elementary	Special Education	Presentation
29. Schnirman[*]	48	24	24	middle	LD	Presentation
30. Smith	196	34	162	elementary	LD	Setting
31. Tindal[a]	1303	215	1088	elementary	Special Education	Presentation
32. Villeneuve	71	35	36	elementary	LD Reading	Timing-Scheduling
33. Walz et al.	110	47	63	middle	Special Education	Timing-Scheduling
34. Weston	119	65	54	elementary	LD	Presentation

[a] Number of participants is across multiple grades

[b] Number of participants is across order of administration and/or forms

[c] Number of participants is across schools

[d] Number of participants is across years

[*] Number of participants is across subject areas

Participant levels of education ranged from elementary school (grade 3) through college, and included one study that used adults. The majority of studies used elementary or middle school-aged students, k = 17 and k = 8, respectively, in their investigations. Research from three studies included in the analyses (Elbaum, 2007; Elbaum et al., 2004; Janson, 2002) focused on cross-level grades, elementary & middle, and middle & secondary, respectively. In 65% of the studies included, most study participants were classified as ‘learning disabled’ with one-half of these studies limiting participants to learning disabilities in reading or reading and math. Studies included all four assessment accommodation types. However, only two of the four accommodation types (presentation and timing/scheduling) were represented in multiple studies, k = 18 and k = 13, respectively. Setting (k = 1) and response (k = 2) accommodations were represented by only three studies. Thus the studies included in the present analyses effectively included only two (presentation and timing/scheduling) of the four (presentation, setting, timing/scheduling, response) accommodation types. Of the various types of accommodations which fall under each category (see Appendix D**,** section on coding accommodations for further information), extended time (which ranged from 20 minutes to 3 days) was used in all timing/scheduling accommodations and read aloud (for example, computer, audio-cassette, assessment proctor) was used in 89% of the presentation studies.

Table 8 relates participant grade level and disability type to type of accommodation examined in the primary study and provides a breakdown of accommodation information by level of education and type of disability evaluated in the primary study. Presentation accommodation research was conducted more frequently for participants with learning disabilities or in special education in the earlier grades while timing/scheduling accommodation research was conducted across all levels of education and focused on participants with learning disabilities. While some accommodations are used more frequently with certain disability groups, such as using an extended time accommodation with learning disabled individuals, it is apparent that there are many gaps in coverage of various accommodations and disability groups in the research literature.

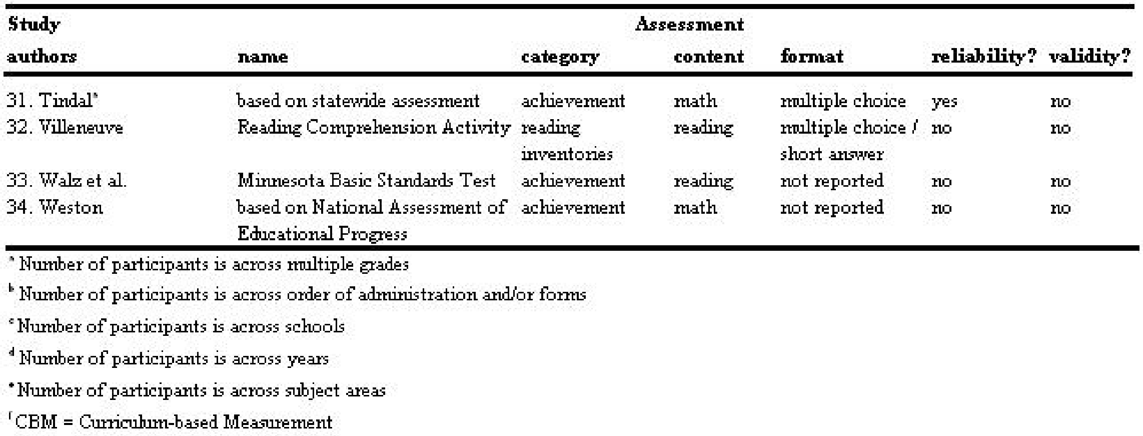
Table 8: *Study Demographics - Accommodation Type x Grade Level and Disability Classification*

	Presentation		**Timing-Scheduling**		**Response**		**Setting**	
	LD	*Special Ed*	*LD*	*Special Ed*	*Other Health*	*LD*	*Special Ed*	*LD*
Elementary	34. Weston	18. Kosciolek & Ysseldyke	10. Fuchs et al. (a)	5. Crawford et al.			9. Engelhard et al.	30. Smith
	16. Johnson	28. Randall & Engelhard	11. Fuchs et al. (b)					
	12. Helwig et al.	13. Helwig & Tindal	3. Buehler					
	20. Lee & Tindal	31. Tindal	32. Villeneuve					
	2. Brown							
	19. Laitusis							
Elementary / Middle		15. Janson						
Middle	29. Schnirman	17. Johnson & Monroe	14. Huesman	24. Marquart	22. Lewandowski et al.			
	26. Meloy et al.	1. Abedi et al.		33. Walz et al.				
Middle / Secondary	7. Elbaum							
	8. Elbaum et al.							
Secondary	4. Calhoon et al.					23. MacArthur & Cavalier		
College			25. Medina					
			6. Dempsey					
			27. Ofiesh et al.					
Adult			21. Lesaux et al.					

Table 9 contains information regarding the assessments used in the primary research studies. The assessments listed in Table 9 include standardized assessments such as the Iowa Test of Basic Skills (ITBS) and the Law School Admission Test (LSAT), researcher-developed assessments using questions from state assessments and the National Assessment of Educational Progress (NAEP), as well as assessments based on state standards (Washington Assessment of Student Learning). Sixty-two percent of these assessments were categorized as measures of achievement. Other assessments were characterized as measures of performance, aptitude, reading improvement, and reading inventories. It should be noted that 18% (6) of the assessments could not be categorized based on information provided in the primary studies. Content areas assessed included math (44%), reading (38%), science (6%), writing (6%), law (3%), and psychology (3%). The majority of assessments used a multiple-choice format (65%) with short answer, open-ended and a combination of multiple-choice/short answer formats being used with much less frequency. Furthermore, 9% of the studies did not report the assessment format used. As might be expected, assessments across all content areas used the multiple-choice format. Alternate formats used to assess math content were either the short answer or the multiple-choice/short answer combination. Writing was assessed using the open-ended format. Forty-one percent (14) of the primary studies reported information on the reliability of the assessments used, with 43% of those studies providing reliability information (18% of total studies) also providing information on the validity of the assessment. While reliability and/or validity information for 29% (10) of the studies could be found online, for a fee in some cases, an equal percentage, 29% (10), did not provide reliability or validity information.

Table 9: *Study Demographics - Assessment Information*

Study	**Assessment**					
authors	**name**	**category**	**content**	**format**	**reliability?**	**validity?**
1. Abedi et al.	based on statewide assessment	not reported	reading	multiple choice	yes	no
2. Brown	based on national/statewide assessment	not reported	science	multiple choice	yes	yes
3. Buehler	California Achievement Tests (5th ed)	achievement	reading	multiple choice	yes	yes
4. Calhoon et al.	Mathematics Performance Assessment	performance	math	short answer	yes	yes
5. Crawford et al.	Oregon Statewide Assessment Test-Writing	performance	writing	open-ended	yes	no
6. Dempsey	Law School Admission Test	aptitude	law	multiple choice	online	online
7. Elbaum	based on statewide test preparation assessment	not reported	reading	multiple choice	yes	no
8. Elbaum et al.	based on statewide assessment	not reported	math	multiple choice	yes	no
9. Engelhard et al.[a]	Georgia Criterion-Referenced Competency Tests	achievement	math	not reported	yes	no
10. Fuchs et al. (a)	Curriculum-based Measurements	achievement	math	short answer	yes	yes
11. Fuchs et al. (b)	Monitoring Basic Skills Progress	achievement	reading	multiple choice	yes	yes
12. Helwig et al.[a, b]	based on statewide assessment	achievement	math	multiple choice	no	no
13. Helwig & Tindal[a, b]	based on statewide assessment	achievement	math	multiple choice	no	no
14. Huesman[c]	Iowa Test of Basic Skills: Reading Comprehension Test	achievement	reading	multiple choice	yes	no
15. Janson[d]	Tennessee Comprehensive Assessment Program Achievement Test	achievement	math	multiple choice	online	online
16. Johnson	Washington Assessment of Student Learning	achievement	math	multiple choice / short answer	no	no
17. Johnson & Monroe	based on statewide assessment	performance	math	multiple choice / short answer	no	no
18. Kosciolek & Ysseldyke	California Achievement Tests	achievement	reading	multiple choice	online	online
19. Laitusis[a]	Gates-McGinitie Reading Tests (4th ed)	achievement	reading	multiple choice	online	online
20. Lee & Tindal[a]	based on statewide assessment	achievement	math	multiple choice	no	no
21. Lesaux et al.	Nelson Denny Reading Test: Reading Comprehension	performance	reading	multiple choice	online	online
22. Lewandowski et al.	Mathematics Calculation Test (researcher)	not reported	math	short answer	no	no
23. MacArthur & Cavalier	based on statewide assessment	not reported	writing	open-ended	yes	no
24. Marquart	TerraNova Level 18 Mathematics test	achievement	math	multiple choice	online	online
25. Medina	General Psychology Test (CBM[e])	achievement	psychology	multiple choice	no	no
26. Meloy et al.[a]	Iowa Test of Basic Skills	achievement	science	multiple choice	online	online
27. Ofiesh et al.	Nelson Denny Reading Test	achievement	reading	multiple choice	online	online
28. Randall & Engelhard[a]	Georgia Criterion-Referenced Competency Tests	achievement	reading	multiple choice	online	online
29. Schnirman[a]	Iowa Test of Basic Skills	achievement	math	multiple choice	online	online
30. Smith	Qualitative Reading Inventory-5	reading improvement	reading	short answer	yes	yes



Study			Assessment			
authors	name	category	content	format	reliability?	validity?
31. Tindal[a]	based on statewide assessment	achievement	math	multiple choice	yes	no
32. Villeneuve	Reading Comprehension Activity	reading inventories	reading	multiple choice / short answer	no	no
33. Walz et al.	Minnesota Basic Standards Test	achievement	reading	not reported	no	no
34. Weston	based on National Assessment of Educational Progress	achievement	math	not reported	no	no

[a] Number of participants is across multiple grades

[b] Number of participants is across order of administration and/or forms

[c] Number of participants is across schools

[d] Number of participants is across years

[e] Number of participants is across subject areas

[f] CBM = Curriculum-based Measurement

Two specific groups of students with disabilities, students with learning disabilities (primary study k = 22) and students receiving special education services (primary study k = 11), were used in more granular meta-analyses, thus further examination of demographics related to these two groups was considered warranted. The ranges for the total number of research participants for these two groups were 31 to 2,028 participants for the studies focusing on individuals with learning disabilities and 31 to 1,317 participants for studies examining individuals receiving special education services. The total number of participants with disabilities and total number of typically developing peers per study was similar. There were 12 to 903 participants with disabilities and 10 to 1,125 typically developing peers in primary studies of students with learning disabilities, while there were 14 to 630 participants with disabilities and 17 to 1,088 typically developing participants in primary studies examining students receiving special education services.

Table 10 displays comparative demographic information for the two most frequently studied disability groups: students with learning disabilities and the more general category of students with disabilities. As can be seen in Table 10, ratios across type of disability group for research design, research approach, assessment content, and assessment format categories are fairly similar in their distributions. However, type of publication, assignment of research participants, level of education, type of accommodation, reporting reliability, and reporting validity are far less similar in their distributions. The largest discrepancy between these groups is found for level of education. The bulk of the studies examining students using special education services were for students in the elementary and middle grades (100%), while primary studies scrutinizing the effects of testing accommodations on individuals with learning disabilities spread across level of education, with only 64% of studies conducted at elementary and middle grades. There were proportionally more dissertations for those studies examining individuals with learning disabilities than there were for students using special education services, 8:22 (36%) versus 2:11 (18%). There was a preponderance of randomization used in the assignment of research participants, 6:22 versus 5:11, for the group with learning disabilities. Additionally, there were proportionally more studies examining extended time for this group, 9:22 versus 3:11. Further, more primary studies involving individuals with learning disabilities, as compared with students receiving special education services, provided reliability and validity information: 11:22 versus 2:11 and 5:22 versus 1:11, respectively.

Table 10: *Study Demographics - Individuals w/ Learning Disabilities & Individuals Receiving Special Education*

	Learning disabilities	**Special education**
Number of primary studies	22	11
Type of publication		
journal	12	6
dissertation	8	2
report	2	3
Research design		
boost	2	1
boost & interaction	2	2
interaction	18	8
Research approach		
independent groups	3	4
repeated measures	19	7
Assignment of research participants		
counterbalanced	15	4
randomized	6	5
not random		1
not reported	1	1
Level of education		
elementary	11	6
elementary / middle		1
middle	3	4
middle / secondary	2	
secondary	2	
college	3	
adult	1	
Accommodation		
extended time	9	3
read aloud	11	5
other	2[a]	3[a, b]
Assessment category		
achievement	13	9
aptitude	1	
performance	4	
reading		2
not reported	4	
Assessment content		
math	10	4
reading	8	5
science	1	1
writing	1	
social studies	1	
law	1	
psychology		1
Assessment format		
multiple-choice	15	7
multiple-choice/short answer	2	1
short answer	3	2
open-ended	1	1
not reported	1	
Reliability reported?		
yes	11	2
online	5	5
no	6	4

	Learning disabilities	Special education
	Validity reported?	
yes	5	1
online	6	5
no	11	5

[a] other = scribe, special acoustics
[b] other = calculator, segmented text, simplified language

Overall trends in the demographic data for the current study parallel the trends observed in syntheses examining the effects of test accommodations (for example, see Bolt & Thurlow, 2004; Cormier et al., 2010; Thompson et al., 2002). In the synthesis by Cormier et al. (2010), 40 studies were examined. Similar to the current research, the most common content areas assessed were math and reading. Also aligned with the current research, the most common accommodations were presentation (56%) and timing/scheduling (38%).

Research quality was not examined as "quality weighting adds uncertainty to average effect sizes but does not eliminate serious bias related to study quality… [and] adds bias in many cases" (Ahn & Becker, 2011, p. 579 – 580). While not specifically examined, a pseudo-measure of quality bucketed the primary studies into three main groupings; published journal articles and conference proceedings which are peer reviewed, published reports which may or may not undergo a peer review process, and unpublished dissertations which are reviewed by dissertation committee members. The 'quality' for journals, conference papers, and dissertations may arguably be considered 'equivalent,' while research reports may be viewed as being of 'lesser quality.'

Results for the Meta-analyses

Meta-analysis research hypotheses.

The current study addressed the following two hypotheses for the meta-analytic portion of the research:

Research Hypothesis 1: Is there empirical support for effects of test accommodations for the target group, students with disabilities, as opposed to their typically developing peers?

Research Hypothesis 2: As measured by effect size, does each of the following constitute an effective accommodation for students with disabilities?

- Presentation test accommodations?
- Response test accommodations?
- Setting test accommodations?
- Timing/Scheduling test accommodations?

Using *Comprehensive Meta-Analysis V.2.2.050,* two separate meta-analyses were performed to answer the first research hypothesis, one meta-analysis using study as the unit of analysis and the other using substudy, or subgroup, as the unit of analysis. Each analysis performed, study-level and substudy-level, provided comparative information for the two groups under investigation, students with disabilities and their typically developing peers. With substudy as the unit of analysis, multiple effect sizes were calculated for some primary studies. For example, if the primary study examined the effects of test accommodations for students with disabilities and their typically developing peers in grade 4 and in grade 7 separately, the data from each of these grades was used to calculate a separate effect size for each group by each grade. While analyzing meta-analytic data by combining substudy information is generally recommended (Borenstein et al., 2009), using independent subgroups within a study is also a valid approach to answering the research hypothesis under investigation. As Borenstein et al. (2009) assert, when independent subgroups are present in a study and each of these subgroups contributes independent information, these "independent subgroups are no different than independent studies" (p. 223), thus allowing the researcher to compute the effect within the subgroup separately.

The information used to answer the first research hypothesis is presented in the following order: results using combined studies; i.e., study as the unit of analysis, and results using substudy as the unit of analysis; i.e., separate effect sizes presented for each substudy. It must be noted that only math assessment results for the Meloy et al. (2002) study are used to calculate effect sizes for the study-level meta-analysis while all assessments, math, reading, science, and using expressions, are used to calculate effect sizes for the substudy-level meta-analysis.

Study as the unit of analysis: Description of effect size.

With study as the unit of analysis, the final 34 studies yielded 65 separate effect sizes. For studies pursuing differential boost, or interaction hypotheses, research purposes data for students with disabilities and their typically developing peers was used to calculate a separate effect size for each group. Thus, the final number of effect sizes was comprised of 34 separate effects for students with disabilities and 31 separate effects for their typically developing peers. There were 5,740 students with disabilities with 8,877 typically developing peers totaling 14,617 participants represented by these studies.

Table 11 provides the breakdown of studies by the research approach and design used in the primary studies. As can be seen in Table 11, two different research designs, independent groups and repeated measures, were combined with three research approaches, boost, a combination of boost and differential boost, and differential boost. Of these combinations, the majority of research conducted to examine efficacy of test accommodations, particularly when comparing students with disabilities to typically developing peers, favored a repeated measures design. In total, over 75% of primary research studies used the repeated measures design. With known difficulties in obtaining a suitable number of students with disabilities to participate in such research, this is to be expected.

Table 11: *Number of Effect Sizes by Research Approach & Design (Unit of Analysis = Study)*

	Independent Groups	**Repeated Measures**
	Boost	
students w/ disabilities	1	2
students w/o disabilities[a]		
	Boost / Differential boost	
students w/ disabilities	1	3
students w/o disabilities[a]	1	3
	Differential boost	
students w/ disabilities	5	22
students w/o disabilities[a]	5	22

[a] students w/o disabilities refers to typically developing students

Information regarding study sample size for students with disabilities and their typically developing peers is provided in Table 12. The median per study sample size for students with disabilities for the independent groups research design was 110 (mean = 261, range 22 to 594), with a median of 528 for typically developing students (mean = 455, range 27 to 725). For the repeated measures research design these totals were 48 (mean = 145, range 12 to 903) and 138 (mean = 246, range 10 to 1125), respectively. There are proportionally fewer students with disabilities, hence the smaller numbers of students with disabilities represented in the primary studies.

Table 12: *Substudy Sample Size Based on Total Number of Effect Sizes*[a]

	Students w/ disabilities	**Students w/o disabilities**[b]
	Independent groups	
Count	7	6
Mean	261.29	455.00
Median	110.00	528.00
Mode	none	none
Minimum	22	27
Maximum	594	725
	Repeated measures	
Count	27	25
Mean	144.85	245.88
Median	48.00	138.00
Mode	48.00	none
Minimum	12	10
Maximum	903	1125

[a] Data include boost, combination, and differential boost studies
[b] students w/o disabilities refers to typically developing students

Substudy as the unit of analysis: Description of effect size.

With substudy, or subgroup, as the unit of analysis, the final 34 studies yielded 119 separate effect sizes. Where applicable, data for students with disabilities and their typically developing peers were used to calculate a separate effect size for each group. A total of 12 studies provided multiple effect sizes, ranging from 3 to 18 additional effects per study for the first research hypothesis (when combining both students with disabilities and students with typical development subgroups) and 2 to 9 for the second research hypothesis (when examining effect sizes for students with disabilities). These effect sizes represent 5,338 students with disabilities and 8,491 typically developing peers for a total of 13,829 participants.

Table 13 provides information on the number of effect sizes by the research approach and design used. Boost studies and boost/differential boost studies produced equivalent numbers of effect sizes. However, for differential boost effect sizes there were four times as many effect sizes for repeated measures designs as compared to independent groups designs. This is due to both the total number of repeated measures differential boost studies (k = 22 studies) and the number of substudies per study. Seven of these differential boost repeated measures studies contained a substantial amount of substudy data (k = 32 for students with disabilities and k = 32 for typically developing students).

Table 13: *Number of Effect Size Estimates by Research Approach & Design (Unit of Analysis = Substudy)*

	Independent Groups	**Repeated Measures**
	Boost	
students w/ disabilities	3	2
students w/o disabilities[a]		
	Boost / Differential boost	
students w/ disabilities	1	3
students w/o disabilities[a]	1	3
	Differential boost	
students w/ disabilities	10	43
students w/o disabilities[a]	10	43

[a] students w/o disabilities refers to typically developing students

Table 14 provides information on substudy sample sizes for students with disabilities and typically developing students. The median per study sample size for students with disabilities for the independent groups research design was 121 (mean = 144, range 22 to 316), with a median of 347 for typically developing students (mean = 302, range 27 to 596). The totals for the repeated measures research design were 35 (mean = 80, range 6 to 527) and 86 (mean = 131, range 10 to 654), respectively. Again, as students participating in research employing the independent groups design generally only took one test, it is expected that the number of participants would be greater. As well, and as might be expected, the numbers of participants for both designs was smaller when substudy, rather than study, was the unit of analysis.

Table 14: *Substudy Sample Size Based on Total Number of Effect Sizes*[a]

	Students w/ disabilities	Students w/o disabilities[b]
	Independent groups	
Count	14	11
Mean	143.86	302.09
Median	120.50	347.00
Mode	62.00	198.00
Minimum	22	27
Maximum	316	596
	Repeated measures	
Count	49	47
Mean	80.47	131.47
Median	35.00	86.00
Mode	24.00	181.00
Minimum	6	10
Maximum	527	654

[a] Data include boost, combination, and differential boost studies
[b] students w/o disabilities refers to typically developing students

Research hypothesis 1.

Research hypothesis 1 asked if there is empirical support for providing test accommodations to students with disabilities as opposed to their typically developing peers. To answer this question Hedges' g was used to calculate effect size for differences between means for each unit of analysis. Use of Hedges' g standardizes the mean differences, thus placing all effect sizes on a common metric, allowing for comparison across studies. For research that did not include means and standard deviations, effect sizes were calculated from reported tests of significance.

$$g = d \times J \tag{3.1}$$

where d is

$$d = \left(\frac{t}{\sqrt{n}} \right) \times \sqrt{2 \times (1 - r)} \tag{3.2}$$

and the correction factor for J is

$$J = 1 - \left(\frac{3}{(4 \times df - 1)} \right) \tag{3.3}$$

where $df = n_{(total)} - 1$

Means and standard errors for effect sizes for students with disabilities and typically developing students are reported and examined separately.

Study as the unit of analysis: Research hypothesis 1 results.

Overall results comparing students with disabilities and their typically developing peers, with study as the unit of analysis, are reported in Table 15. Table 15 shows that the Q-test for the distribution of observed effect sizes for students with disabilities and typically developing students, $Q_{(33)} = 650.08$ and $Q_{(30)} = 403.16$, respectively, was statistically significant ($p < 0.001$). This suggests that there is heterogeneity in conditions for each group; i.e., non-accommodated versus accommodated conditions, differences that are not readily accounted for by sampling variation. That is to say, the true effect size does vary from study to study due to heterogeneity in effect size and within study error. While the Q-test value for the remaining analyses is reported, discussion will be limited as, with a single exception (see p. 201), Q-test values were statistically significant thus, selection of the random-effects model for further analysis was deemed appropriate.

While use of the random-effects model was decided upon *a priori*, as confidence intervals generated with random-effects models do not overstate the degree of precision for the meta-analytic findings, and this model does not produce any substantial Type I bias for mean effects significance tests and moderators, or interactions (Hunter & Schmidt, 2000): results from the Q-test support this decision.

Table 15: *Comparison Between Students With and Without Disabilities -* $\overline{ES}$ *Estimates, Confidence Intervals, & Q-statistics*[a]

		Mean effect size & 95% CI (Hedges' g)					**Heterogeneity**		
Comparison group	**k**	$\overline{ES}$ [b]	**Std Err**[b]	**LL**[b]	**UL**[b]	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
Fixed effects									
students w/ disabilities	34	0.42	0.01	0.40	0.44	< 0.001	650.08	33	< 0.001
students w/o disabilities[c]	31	0.12	0.01	0.11	0.14	< 0.001	403.16	30	< 0.001
Random effects									
students w/ disabilities	34	0.36	0.06	0.25	0.48	< 0.001			
students w/o disabilities[c]	31	0.19	0.04	0.12	0.26	< 0.001			

[a] Study was used as the unit of analysis, all substudy information was combined

[b] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

[c] students w/o disabilities refers to typically developing students

Students with disabilities.

Under the random-effects model, the mean effect size for students with disabilities is 0.36 while it is 0.19 for their typically developing peers (Table 15). This indicates that there is a small positive mean effect for test accommodations for students with disabilities while there is an even smaller, albeit statistically significant, mean effect for their typically developing peers. These results indicate that both students with disabilities and typically developing students benefit from test accommodations. This was not surprising given current special education classification requirements, whereby some typically developing students might qualify for and receive special education services if classification practices were slightly more lenient.

For students with disabilities, the 34 different effect sizes calculated ranged from -0.24 to 1.43 (see Appendix L for effect sizes and standard errors calculated for students with disabilities).There were 28 effects (82%) positive and 6 effects (18%) negative, thus test accommodations appear to have a positive effect for students with disabilities.

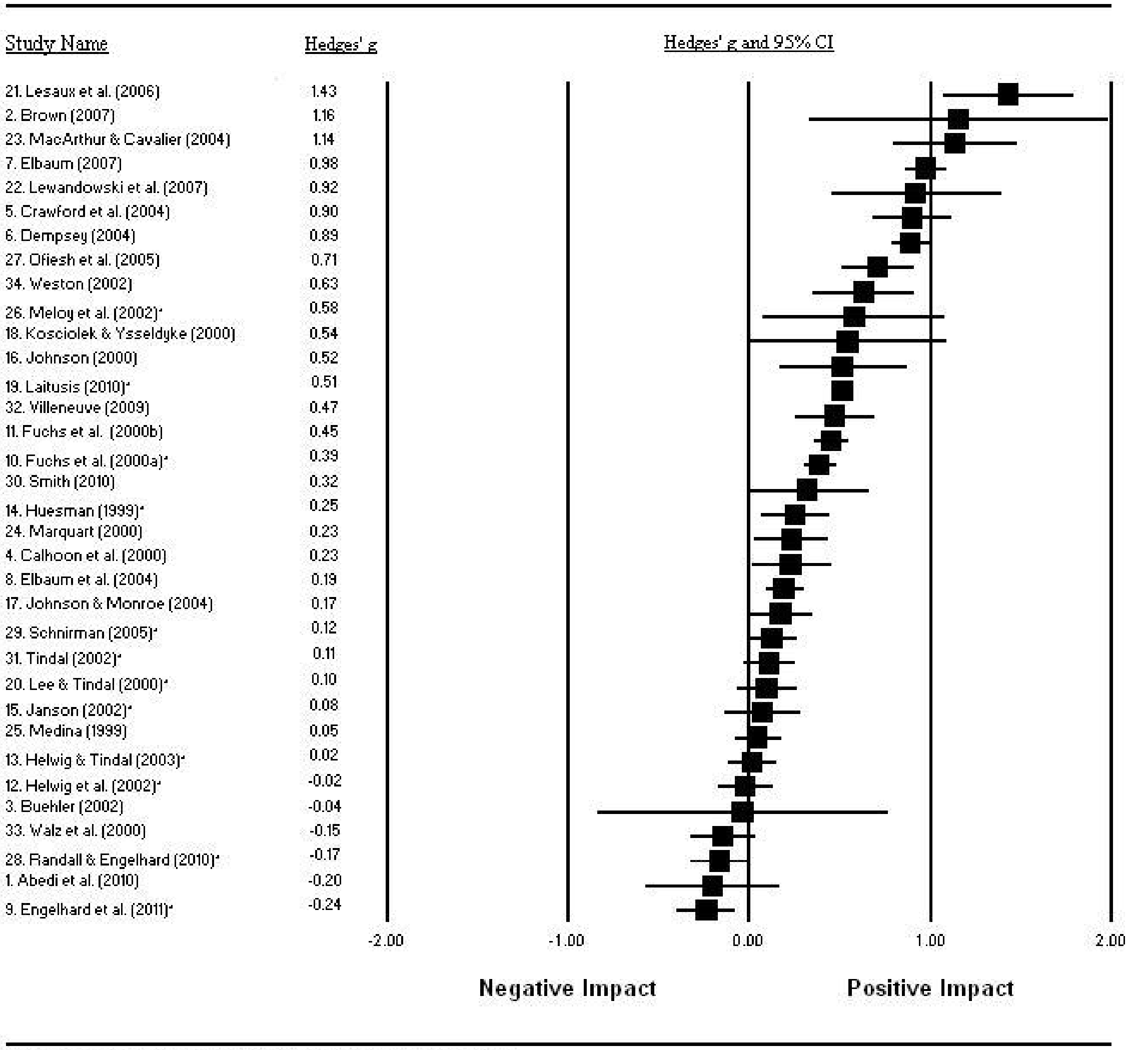
There were 21 statistically significant effects in total, with 19 positive effects and only 2 negative effects. The majority of these positive effect sizes, 19 (68%) were statistically significant while 13 (32%), were not. Most negative effect sizes were not statistically significant, 67% of non-significant effects or 12% of the total effects, for this group.

Effect sizes were categorized using Cohen's (1992) labels for "mean" effect size with 0.8 deemed large, 0.5 deemed medium, and 0.2 deemed small, as lower-bound estimates. Effect sizes in the positive range were large, k = 7 or 21%, medium, k = 6 or 18%, or small, k = 7 or 21% with 2 (6%) negative small effects. The 12 effects (35%) ranging between -0.17 and 0.193 were considered very small.

The preponderance of statistically significant effects, both positive and negative, were small, medium or large with a few exceptions. One study with a medium effect size, Kosciolek and Ysseldyke, 2000 (18), and one study with a small effect size, Smith, 2010 (30), did not reach statistical significance as values adjusted for error spanned the midpoint interval of zero. Only 1 of the 2 negative small effect sizes was statistically significant, Engelhard et al., 2011 (9). Thus, we see the majority of effects were positive, with one-fifth of these being large, statistically significant effects.

The standard error, a measure of precision, is on the same scale as effect size and ranged from 0.02 to 0.44 across all studies included in the analysis. Studies with the largest standard error, Brown, 2007 (2) and Buehler, 2002 (3), are considered less precise than Laitusis, 2010 (19) and Fuchs et al. (2000a) (10). One-half of the standard errors were smaller than 0.10.

The forest plot in Figure 3 displays effect sizes for the 34 primary studies examining accommodation effects for students with disabilities, bounded by their respective confidence intervals.



* effect size computed used combined substudies (students with disabilities)

Figure 3: Forest Plot of Effect Size Estimates for Students with Disabilities – Study as the Unit of Analysis

Ten of the 12 very small effects spanned the midpoint interval of zero as summarized in the forest plot of effect sizes for students with disabilities (Figure 3). Since these effects are considered both, very small and span zero, we can infer a null effect of test accommodations for students with disabilities for 35% of primary studies examined. All studies contributed almost equal weighting to the overall results, with no one study being particularly dominant in the analysis, as would be expected under the random-effects model.

Figure 3 reveals that most individual study effect size estimates and the overall mean effect size estimate were relatively precise, with two exceptions, Brown, 2007 (2) and Buehler, 2002 (3). This figure also shows that the majority of confidence intervals around effects sizes did not include zero, were statistically significant, and were positive, thus providing evidence for the positive impact of test accommodations for students with disabilities. Only 15 of the study effect sizes, 44%, fall inside the confidence interval for the overall mean effect.

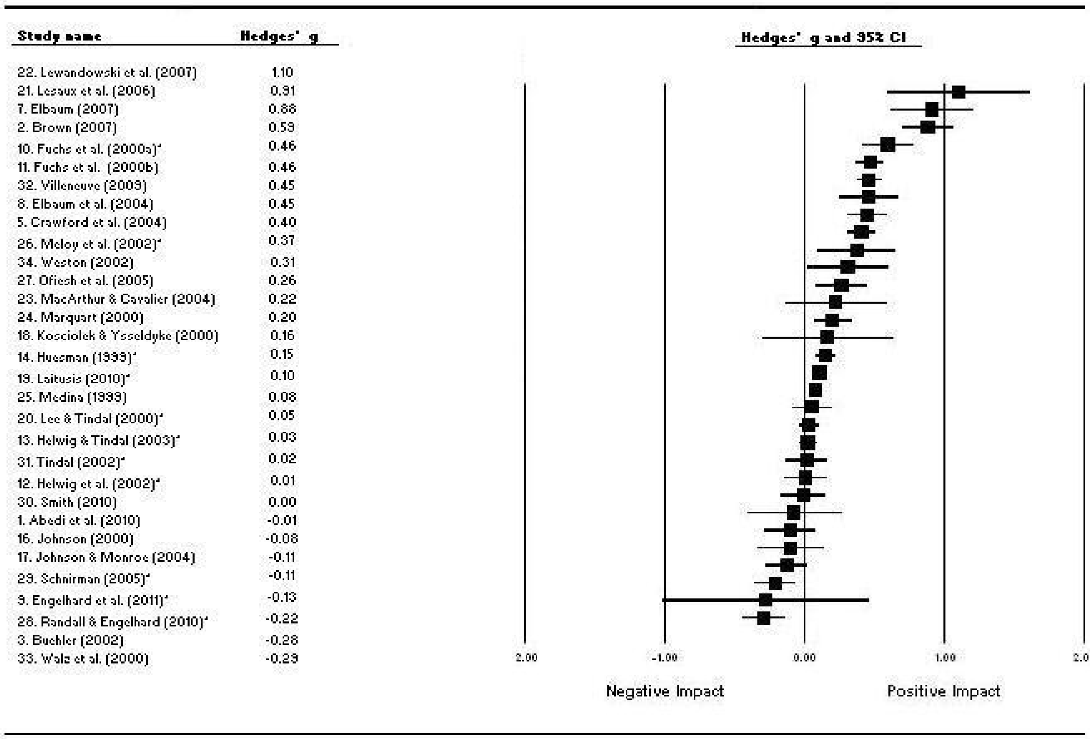
Typically developing students.

The 31 effect sizes calculated for typically developing students ranged from -0.29 to 1.10 (see Appendix M for effect sizes and standard errors calculated for typically developing students). There were 23 positive effects (74%) and 8 negative effects (26%). Of the total number of effect sizes, 18 (58%) were statistically significant while 13 (42%) were not. While 16 of the statistically significant effects (70% of positive effects, 52% of total effects) were positive, almost one-third of these effects (7; 30% of positive effects, 23% of total effects) were not statistically significant. Two of the 8 negative effects (25% of negative effects, 6% of total effects) were statistically significant.

Categorizing effect size, we see 3 large (9.7%), 1 medium (3.2%), and 9 small, (29.0%) positive effects and 3 small negative effects (9.7%). Almost one-half (15:31 or 48.4%) of effects, ranging between -0.13 and 0.20, were very small. All but 1 of the non-trivial, positive effect sizes, MacArthur & Cavalier, 2004 (23), were statistically significant. One of the 2 non-trivial, negative effect sizes was statistically significant, while the other, Buehler, 2002 (2), was not. For typically developing students the majority of effects were positive, with over one-fifth of these being small, statistically significant effects.

Standard errors for typically developing students ranged from 0.01 to 0.38 across all studies included in the analysis. The effect sizes for Buehler, 2002 (3) and Lewandowski et al., 2007 (22) were less precise than Medina, 1999 (25) and Laitusis, 2010 (19).One-half of the standard errors were smaller than 0.08.

Figure 4 provides a visual display of the 31 effect sizes for typically developing students.



* effect size computed used combined substudies (typically developing students)

Figure 4: Forest Plot of Effect Size Estimates for Typically Developing Students – Study as the Unit of Analysis

For these students, 11 of the 15 very small effects had confidence intervals which included zero (Figure 4). As for the students with disabilities, we can infer a null effect of test accommodations for typically developing students for 35% of primary studies examined. Similarly, almost all studies contributed equal weighting to the overall results, with no one study being particularly dominant in the analysis.

Most individual study effect size estimates and the overall mean effect size estimate were relatively precise, with two exceptions, Lewandowski et al., 2007 (22) and Buehler, 2002 (3). Eight of the study effect sizes (26%) fell inside the confidence interval for the overall mean effect. Additionally, as seen in Figure 4, close to one-half (41%) of effects sizes spanned the zero midpoint interval and can be considered very small, as well as non-significant, providing evidence for a lack of effect for test accommodations for typically developing students.

Figure 5 provides an expanded, graphical representation of the effects of test accommodations on students with disabilities and their typically developing peers.

Students with disabilities			Interval midpoint	Typically developing students			
			1.45				
		21, LDR, RD, ET, 1.43	1.40				
			1.35				
			1.25				
			1.20				
		2, LDR, SC, RA, 1.16	1.15	22, OH, MA, ET, 1.10			
		23, LD , WR, SD, 1.14	1.10				
			1.05				
			1.00				
		8, LD , MA, RA, 0.98	0.95				
		22, OH, MA, ET, 0.92	0.90	21, LDR, RD, ET, 0.91			
	6, LD , O, ET, 0.89	5, SE, WR, ET, 0.90	0.85	8, LD , MA, RA, 0.88			
			0.80				
			0.75				
		27, LDR, RD, ET, 0.71	0.70				
			0.65				
		34, LD , MA, RA, 0.63	0.60				
		26, LDR, SC, RA, 0.58	0.55	2, LDR, SC, RA, 0.59			
19, LDR, RD, RA, 0.51	16, LDR, MA, RA, 0.52	18, SE, RD, RA, 0.54	0.50	10, LD , MA, ET, 0.46			
	11, LD , RD, ET, 0.45	32, LDR, RD, ET, 0.47	0.45	11, LD , RD, ET, 0.46	32, LDR, RD, ET, 0.45		
			0.40	7, LD , RD, RA, 0.45	5, SE, WR, ET, 0.40		
		10, LD , MA, ET, 0.39	0.35	26, LDR, SC, RA, 0.37			
		30, LD , RD, SA, 0.32	0.30	34, LD , MA, RA, 0.31			
		14, LD , RD, ET, 0.25	0.25	27, LDR, RD, ET, 0.26			
	4, LDRM, MA, RA, 0.23	24, SE, MA, ET, 0.23	0.20	23, LD , WR, SD, 0.22			
	17, SE, MA, SL, 0.17	7, LD , RD, RA, 0.19	0.15	24, SE, MA, ET, 0.20	18, SE, RD, RA, 0.16		
	31, SE, MA, RA, 0.11	29, LD, MA, RA, 0.13	0.10	14, LD , RD, ET, 0.15	19, LDR, RD, RA, 0.10		
	15, SE, MA, RA, 0.08	20, LDR, MA, RA, 0.10	0.05	25, LD , O, ET, 0.08	20, LDR, MA, RA, 0.05		
	13, SE, MA, RA, 0.02	25, LD , O, ET, 0.05	0.00	13, SE, MA, RA, 0.03	31, SE, MA, RA, 0.02	12, LDR, MA, RA, 0.01	30, LD , RD, SA, 0.00
	3, LDR, RD, ET, -0.04	12, LDR, MA, RA, -0.02	-0.05	1, SE, RD, ST, -0.01			
		33, SE, RD, ET, -0.15	-0.10	16, LDR, MA, RA, -0.08			
		28, SE, RD, RA, -0.17	-0.15	17, SE, MA, SL, -0.11	29, LD, MA, RA, -0.11	9, SE, MA, CA, -0.13	
		1, SE, RD, ST, -0.20	-0.20	28, SE, RD, RA, -0.22			
		9, SE, MA, CA, -0.24	-0.25	3, LDR, RD, ET, -0.28	33, SE, RD, ET, -0.29		
			-0.30				

RA = Read Aloud	E = Elementary
ST = Segmented Text	EM = Elementary & Middle
SL = Simplified Language	M = Middle
SA = Special Acoustics	MS = Middle & Secondary
ET = Extended Time	S = Seconday
CA = Calculator Use	C = College
SD = Scribe for dictation	A = Adult
LD = Learning Disability	MA = mathematics
LDR = Learning Disability (Reading)	RD = reading/language arts
	SC = science
LDRM = Learning Disability (Reading & Math)	WR = writing
	O = Other
OH = Other Health Impaired	
SE = Special Education	

Note: Each cell contains the (1) study number, (2) assessment accommodation, (3) disability classification, (4) assessment subject, and (5) effect size estimate (Hedges' g)

Figure 5: Graph of Hedges' g Effect Size Estimates for Students with Disabilities Compared to Typically Developing Students - Study as Unit of Analysis

From Figure 5 we see that two main groups of students with disabilities, those receiving special education assistance and those classified with learning disabilities, comprise the majority of students with disabilities. We also see that most of the assessments relied on math or reading content and the bulk of the accommodations were either for extended time or for reading aloud. Examination of the distribution of effect sizes about the interval midpoints shows that students with disabilities were more likely to be positively impacted by test accommodations than their typically developing peers.

Overall, percentages of statistically significant effect size estimates for students with typical development mirror the results found for students with disabilities. Effect size estimation for Buehler, 2002 (3) and Laitusis, 2010 (19) was less precise for both students with disabilities and those with typical development. Effect sizes appear to be measured more precisely for typically developing students as compared to students with disabilities. The overall mean effect size for students with disabilities, 0.36, albeit small-to-medium, reached statistical significance, $p < 0.001$. Test accommodations have a very small, statistically significant mean effect (0.19, $p < 0.001$) for typically developing students. The overall effect for students with disabilities, while small and statistically significant (0.36, $p < 0.001$), is almost double the mean effect size for their typically developing peers. These results may be interpreted, cautiously, to lend support to the differential boost hypotheses, whereby students with disabilities are positively impacted by test accommodations while their typically developing peers are also affected, albeit minimally.

Substudy as the unit of analysis: Research hypothesis 1 results.

Overall results, comparing students with disabilities and typically developing students, with substudy as the unit of analysis, are presented in Table 16. The Q-test for the distribution of observed effect sizes for both students with disabilities and those with typical development, $Q_{(61)} = 782.27$ and $Q_{(56)} = 512.14$, was statistically significant ($p < 0.001$). The overall mean effect size, under the random-effects model, was 0.30 for students with disabilities and 0.17 for their typically developing peers. Mirroring the results using study as the unit of analysis, there was a small positive mean effect for test accommodations for students with disabilities and a very small, statistically significant mean effect for their typically developing peers.

Table 16: *Comparison Between Students With and Without Disabilities - $\overline{ES}$ Estimates, Confidence Intervals, & Q-statistics*[a]

		Mean effect size & 95% CI (Hedges' g)					**Heterogeneity**		
Comparison group	**k**	$\overline{ES}$ [b]	**Std Err**[b]	**LL**[b]	**UL**[b]	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
			Fixed effects						
students w/ disabilities	62	0.42	0.01	0.40	0.44	< 0.001	782.27	61	< 0.001
students w/o disabilities[c]	57	0.13	0.01	0.11	0.14	< 0.001	512.14	56	< 0.001
			Random effects						
students w/ disabilities	62	0.30	0.04	0.21	0.38	< 0.001			
students w/o disabilities[c]	57	0.17	0.03	0.11	0.22	< 0.001			

[a] Substudy was used as the unit of analysis

[b] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

[c] students w/o disabilities refers to typically developing students

Students with disabilities.

Substudies and studies included in the following section will be referred to as studies. There were 62 different effect sizes ranging from -0.57 to 1.43 for students with disabilities, with just over 75% of these values being positive (see Appendix N for effect sizes and standard errors calculated for students with disabilities).

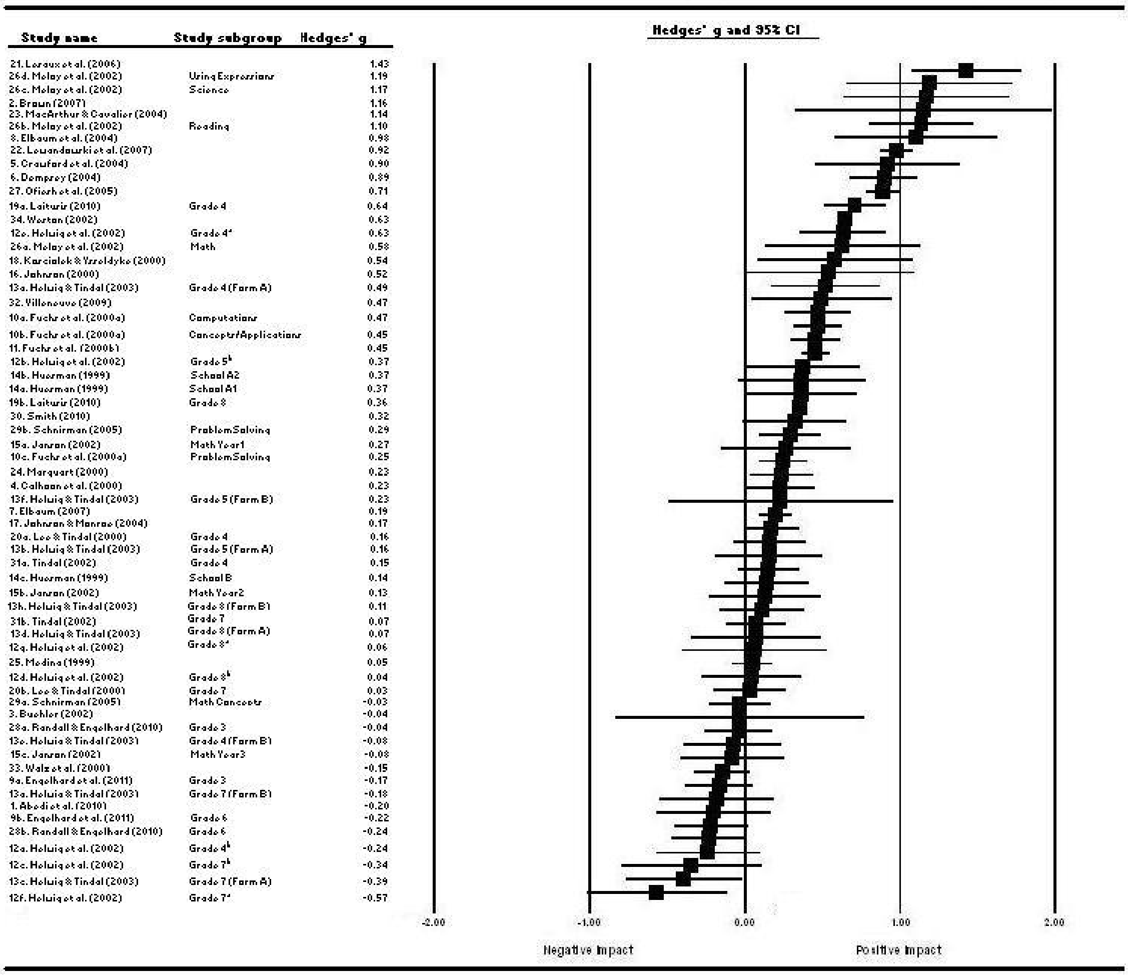
Of this total, 30 of the effects were statistically significant and 32 were not. Twenty-eight of the statistically significant effects were positive, with only two negative statistically significant effects. That is, most of the negative effect sizes, 87% of negative effects or 21% of the total effects, for this group were not significant.

There were 47 positive effects and 15 negative effects. The positive effects were large, k = 10 or 16%, medium, k = 7 or 11%, or small, k = 16 or 26% with six small negative effects, 10%, and one medium negative effect, 2%. Twenty-two effects (35%) ranging between -0.18 and 0.19 were considered very small.

Most of the statistically significant positive and negative effects were small, medium or large effects with few exceptions. One study with a medium effect size, Kosciolek and Ysseldyke, 2000 (18), and five studies with small effect sizes, Helwig et al. 2002 (12b), Huesman, 1999 (14b), Smith, 2010 (30), Janson, 2002 (15a), and Helwig and Tindal, 2003 (13f), did not reach statistical significance. Only two of the seven small and medium negative effect sizes were statistically significant, Helwig and Tindal, 2003 (13c) and Helwig et al., 2002 (12f). Thus, we see the majority of effects were positive, with almost one-fifth of these being large, statistically significant effects.

The standard error for the effect sizes ranged from 0.02 to 0.42 across all studies included in the analysis. Studies with the largest standard error, Brown, 2007 (2) and Buehler, 2002 (3), are considered less precise than Laitusis, 2010 (19a) and Laitusis, 2010 (19b).

The forest plot in Figure 6 displays the effect sizes for the 62 primary studies examining accommodations for students with disabilities, bounded by their respective confidence intervals.



Study name	Study subgroup	Hedges' g
21. Leroux et al. (2006)		1.43
26d. Meloy et al. (2002)	Using Expressions	1.19
26c. Meloy et al. (2002)	Science	1.17
2. Brown (2007)		1.16
23. MacArthur & Cavalier (2004)		1.14
26b. Meloy et al. (2002)	Reading	1.10
8. Elbaum et al. (2004)		0.98
22. Lewandowski et al. (2007)		0.92
5. Crawford et al. (2004)		0.90
6. Dempsey (2004)		0.89
27. Ofiesh et al. (2005)		0.71
19a. Laitusis (2010)	Grade 4	0.64
34. Weston (2002)		0.63
12e. Helwig et al. (2002)	Grade 4[a]	0.63
26a. Meloy et al. (2002)	Math	0.58
18. Kosciolek & Ysseldyke (2000)		0.54
16. Johnson (2000)		0.52
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	0.49
32. Villeneuve (2009)		0.47
10a. Fuchs et al. (2000a)	Computations	0.47
10b. Fuchs et al. (2000a)	Concepts/Applications	0.45
11. Fuchs et al. (2000b)		0.45
12b. Helwig et al. (2002)	Grade 5[b]	0.37
14b. Huesman (1999)	School A2	0.37
14a. Huesman (1999)	School A1	0.37
19b. Laitusis (2010)	Grade 8	0.36
30. Smith (2010)		0.32
29b. Schnirman (2005)	Problem Solving	0.29
15a. Janson (2002)	Math Year1	0.27
10c. Fuchs et al. (2000a)	Problem Solving	0.25
24. Marquart (2000)		0.23
4. Calhoon et al. (2000)		0.23
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.23
7. Elbaum (2007)		0.19
17. Johnson & Monroe (2004)		0.17
20a. Lee & Tindal (2000)	Grade 4	0.16
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	0.16
31a. Tindal (2002)	Grade 4	0.15
14c. Huesman (1999)	School B	0.14
15b. Janson (2002)	Math Year2	0.13
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	0.11
31b. Tindal (2002)	Grade 7	0.07
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.07
12g. Helwig et al. (2002)	Grade 8[a]	0.06
25. Medina (1999)		0.05
12d. Helwig et al. (2002)	Grade 8[b]	0.04
20b. Lee & Tindal (2000)	Grade 7	0.03
29a. Schnirman (2005)	Math Concepts	-0.03
3. Buehler (2002)		-0.04
28a. Randall & Engelhard (2010)	Grade 3	-0.04
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	-0.08
15c. Janson (2002)	Math Year3	-0.08
33. Walz et al. (2000)		-0.15
9a. Engelhard et al. (2011)	Grade 3	-0.17
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.18
1. Abedi et al. (2010)		-0.20
9b. Engelhard et al. (2011)	Grade 6	-0.22
28b. Randall & Engelhard (2010)	Grade 6	-0.24
12a. Helwig et al. (2002)	Grade 4[b]	-0.24
12c. Helwig et al. (2002)	Grade 7[b]	-0.34
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	-0.39
12f. Helwig et al. (2002)	Grade 7[a]	-0.57

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

Figure 6: Forest Plot of Effect Size Estimates for Students with Disabilities – Substudy as the Unit of Analysis

Twenty-one of the twenty-two very small effects spanned the interval midpoint of zero as displayed in the forest plot (Figure 6). With these effects being very small and spanning zero, they were considered to have a trivially small, or null, effect. All studies contributed almost equal weighting to the overall results, with no one study being dominant in the analysis. Figure 6 demonstrates that, while most individual study effect size estimates and the overall mean effect size estimate were precise, there were four notable exceptions, Brown, 2007 (2), Helwig & Tindal, 2003 (13f), Buehler, 2002 (3), and Helwig et al., 2002 (12f). This figure also shows that a large proportion of effects sizes, 27:62 or 44%, did not include zero in the interval, were statistically significant, and were positive, again, providing evidence for the positive impact of test accommodations for students with disabilities. One-quarter, k = 16, of the study effect sizes fall inside the confidence interval for the overall mean effect.

Typically developing students.

Typically developing students were represented by 57 different effect sizes ranging from -0.39 to 1.87, with just 70% of these effect sizes being positive (see Appendix O for effect sizes and standard errors calculated for these students). Twenty-five (44%) of these effects were statistically significant, while 32 (56%) were not. Twenty-two of the statistically significant effects were positive, with only three significant negative effects. That is, most of the effect sizes for negative effects were not significant, 87% of non-significant effects or 24% of the total effects for this group.

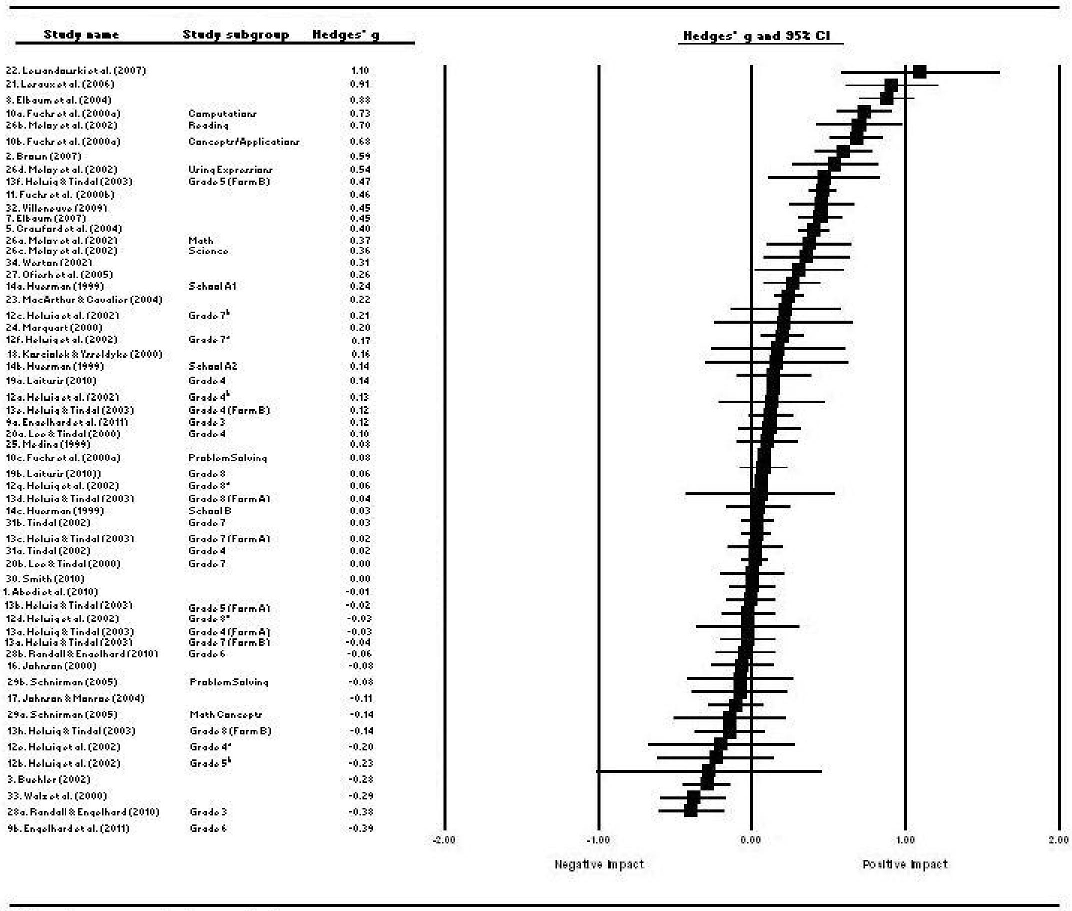
Forty effect sizes were positive while 17 were negative. Large, k = 3 or 5%, medium, k = 5 or 9%, and small, k = 21 or 21%, positive effects were seen. There were five (9%) small negative effects. The remaining 32 effects (56%) ranging from -0.20 to 0.20, were considered very small.

Most statistically significant positive and negative effects were small, medium or large effects with few exceptions. Four studies with small effect sizes, two positive; MacArthur & Cavalier, 2004 (23) and Helwig et al., 2002 (12c); and two negative; Helwig et al. 2002 (12b) and Buehler, 2002 (3), did not reach statistical significance.

Consequently, we see the majority of effects were trivially small and non-significant, 29:57 or 51%.

The standard error for the effect sizes ranged from 0.02 to 0.38 across all studies included in the analysis. Studies with the largest standard error, Lewandowski et al., 2007 (22) and Helwig et al., 2002 (12g), are considered less precise than Laitusis, 2010 (19a) and Medina, 1999 (25).

Figure 7 displays effect sizes for the 57 primary studies for typically developing students, bounded by their respective confidence intervals.



[a] Condition order: not accommodated - accommodated
[b] Condition order: accommodated - not accommodated

Figure 7: Forest Plot of Effect Size Estimates for Typically Developing Students – Substudy as the Unit of Analysis

For typically developing students, 29 of the 32 trivially small effects spanned the midpoint interval of zero (Figure 7), leading us to infer a trivially small, or null, effect for test accommodations in 51% of the studies examined for this population of students. There was very little variability in the weighted contribution of each study to the overall results, again, with no individual study being dominant in the analysis.

For the most part individual study effect size estimates and the overall mean effect size estimate were precise, with three exceptions, Lewandowski et al., 2007 (22), Helwig et al., 2002 (12g) and Buehler, 2002 (3). Sixteen effect sizes (28%) fall inside the confidence interval for the overall mean effect, with 1, Laitusis, 2010 (19b), fully enclosed within the confidence interval. In addition, over one-half (53%) of effects sizes spanned the zero midpoint interval, can be considered trivially small and were non-significant, again, providing evidence for negligible impact of test accommodations for typically developing students.

Figure 8 provides an expanded, graphical representation of the effects of test accommodations on students with disabilities and their typically developing peers.

Students with disabilities					Interval midpoint	Typically developing students						
					1.45							
				21, LDR, RD, ET, 1.43	1.40							
					1.35							
					1.25							
					1.20							
		2, LDR, SC, RA, 1.16	26c, LDR, SC, RA, 1.17	26d, LDR, SC, RA, 1.20	1.15	22, OH, MA, ET, 1.10						
			26b, LDR, SC, RA, 1.10	23, LD , WR, SD, 1.14	1.10							
					1.05							
					1.00							
				8, LD , MA, RA, 0.98	0.95							
				22, OH, MA, ET, 0.92	0.90	21, LDR, RD, ET, 0.91						
			6, LD , O, ET, 0.89	5, SE, WR, ET, 0.90	0.85	8, LD , MA, RA, 0.88						
					0.80							
					0.75							
				27, LDR, RD, ET, 0.71	0.70	10a, LD , MA, ET, 0.73						
					0.65	26b, LDR, SC, RA, 0.70	10b, LD , MA, ET, 0.68					
		12e, LDR, MA, RA, 0.63	34, LD , MA, RA, 0.63	19a, LDR, RD, RA, 0.64	0.60							
				26a, LDR, SC, RA, 0.58	0.55	2, LDR, SC, RA, 0.59						
			16, LDR, MA, RA, 0.52	18, SE, RD, RA, 0.54	0.50	26d, LDR, SC, RA, 0.54						
11, LD , RD, ET, 0.45	10b, LD , MA, ET, 0.45	10a, LD , MA, ET, 0.47	32, LDR, RD, ET, 0.47	13a, SE, MA, RA, 0.49	0.45	13f, SE, MA, RA, 0.47	11, LD , RD, ET, 0.46	32, LDR, RD, ET, 0.45				
					0.40	7, LD , RD, RA, 0.45	5, SE, WR, ET, 0.40					
	19b, LDR, RD, RA, 0.36	14a, LD , RD, ET, 0.37	14b, LD , RD, ET, 0.37	12b, LDR, MA, RA, 0.37	0.35	26a, LDR, SC, RA, 0.37	26c, LDR, SC, RA, 0.36					
				30, LD , RD, SA, 0.32	0.30	34, LD , MA, RA, 0.31						
			15a, SE, MA, RA, 0.27	29b, LD, MA, RA, 0.29	0.25	27, LDR, RD, ET, 0.26						
	13f, SE, MA, RA, 0.23	4, LDRM, MA, RA, 0.23	24, SE, MA, ET, 0.23	10c, LD , MA, ET, 0.25	0.20	14a, LD , RD, ET, 0.24	23, LD , WR, SD, 0.22	12c, LDR, MA, RA, 0.21				
31a, SE, MA, RA, 0.15	13b, SE, MA, RA, 0.16	20a, LDR, MA, RA, 0.16	17, SE, MA, SL, 0.17	7, LD , RD, RA, 0.19	0.15	24, SE, MA, ET, 0.20	12f, LDR, MA, RA, 0.17	18, SE, RD, RA, 0.16				
		13h, SE, MA, RA, 0.11	15b, SE, MA, RA, 0.13	14c, LD , RD, ET, 0.14	0.10	14b, LD , RD, ET, 0.14	19a, LDR, RD, RA, 0.14	12a, LDR, MA, RA, 0.13	13e, SE, MA, RA, 0.12	9a, SE, MA, CA, 0.12	20a, LDR, MA, RA, 0.10	
		12g, LDR, MA, RA, 0.07	13d, SE, MA, RA, 0.07	31b, SE, MA, RA, 0.08	0.05	25, LD , O, ET, 0.08	10c, LD , MA, ET, 0.08	19b, LDR, RD, RA, 0.06	12g, LDR, MA, RA, 0.06			
		20b, LDR, MA, RA, 0.03	12d, LDR, MA, RA, 0.04	25, LD , O, ET, 0.05	0.00	13d, SE, MA, RA, 0.04	14c, LD , RD, ET, 0.03	31b, SE, MA, RA, 0.03	13c, SE, MA, RA, 0.02	31a, SE, MA, RA, 0.02	20b, LDR, MA, RA, 0.00	30, LD , RD, SA, 0.00
		28a, SE, RD, RA, -0.04	3, LDR, RD, ET, -0.04	29a, LD, MA, RA, -0.03	-0.05	1, SE, RD, ST, -0.01						
			15c, SE, MA, RA, -0.08	13e, SE, MA, RA, -0.08	-0.10	13b, SE, MA, RA, -0.02	12d, LDR, MA, RA, -0.01	13a, SE, MA, RA, -0.03	13g, SE, MA, RA, -0.04	28b, SE, RD, RA, -0.06	16, LDR, MA, RA, -0.08	29b, LD, MA, RA, -0.08
				33, SE, RD, ET, -0.15	-0.15	17, SE, MA, SL, -0.11	29a, LD, MA, RA, -0.14	13h, SE, MA, RA, -0.14				
			13g, SE, MA, RA, -0.18	9a, SE, MA, CA, -0.17	-0.20	12e, LDR, MA, RA, -0.20						
	12a, LDR, MA, RA, -0.24	28b, SE, RD, RA, -0.24	9b, SE, MA, CA, -0.22	1, SE, RD, ST, -0.20	-0.25	12b, LDR, MA, RA, -0.24	3, LDR, RD, ET, -0.28	33, SE, RD, ET, -0.29				
				12c, LDR, MA, RA, -0.34	-0.30							
				13c, SE, MA, RA, -0.39	-0.35	28a, SE, RD, RA, -0.38	9b, SE, MA, CA, -0.39					
					-0.40							
					-0.45							
					-0.50							
				12f, LDR, MA, RA, -0.57	-0.55							
					-0.60							

RA = Read Aloud	E = Elementary
ST = Segmented Text	EM = Elementary & Middle
SL = Simplified Language	M = Middle
SA = Special Acoustics	MS = Middle & Secondary
ET = Extended Time	S = Seconday
CA = Calculator Use	C = College
SD = Scribe for dictation	A = Adult
LD = Learning Disability	MA = mathematics
LDR = Learning Disability (Reading)	RD = reading/language arts
	SC = science
LDRM = Learning Disability (Reading & Math)	WR = writing
	O = Other
OH = Other Health Impaired	
SE = Special Education	

Note: Each cell contains the (1) study number, (2) assessment accommodation, (3) disability classification, (4) assessment subject, and (5) effect size estimate (Hedges' g)

Figure 8: Graph of Hedges' g Effect Size Estimates for Students with Disabilities Compared to Typically Developing Students - Substudy as Unit of Analysis

Comparison of results between students with disabilities and typically developing students.

Two main groups of students with disabilities, those receiving special education assistance and those classified as learning disabled, comprise the majority of students with disabilities seen in Figure 8. As well, most of the assessments relied on math or reading content and the bulk of the accommodations were for either extended time or reading aloud. Distribution of effect sizes about the interval midpoints provides visual confirmation that students with disabilities were more likely to be positively impacted by assessment accommodations, more values being above 0.20, than typically developing peers, more values hovering around 0.20 and below.

Percentages of statistically significant effect sizes for typically developing students are very similar to those found for students with disabilities. Precision of estimates appears similar for the two groups; however, less precise study estimates for the students with disabilities were not from the same primary study as those for typically developing peers. Further, effect size appears to more precisely measured for typically developing students when compared to students with disabilities. A small, statistically significant mean effect (0.30) was found for impact of test accommodations for students with disabilities. The statistically significant overall mean effect for typically developing students (0.17) was considered very small. Although results for typically developing students indicate test accommodations have a positive effect for these students, the effect is considered trivially small. On the other hand, the impact of test accommodations for students with disabilities, while small, is nontrivial. As was the case using study as the unit of analysis, the results using substudy as the unit of analysis support the differential boost hypotheses.

Study and substudy as the unit of analysis: A comparison.

Several coded studies contained data that could be used to calculate more than one effect size per group (students with disabilities and typically developing students). Two parallel analyses were executed, one using study as the unit of analysis and one using substudy as the unit of analysis. A comparison of the results from both analyses follows. The following convention will be used throughout the remainder of this document: study as the unit of analysis will be referred to as study or study results, substudy as the unit of analysis will be referred to as substudy or substudy results.

Sixty-five effect sizes, 34 for students with disabilities and 31 for typically developing students, were calculated for study results while 119 effect sizes, 62 for students with disabilities and 57 for typically developing students, were calculated for substudy results. Differences between these numbers is product of multiple data points for 12 studies for students with disabilities and 11 studies with multiple data points for typically developing students. Reasons for multiple data points vary from non-aggregation of data for participants from multiple grades to the same group of participants taking different assessments.

As would be expected, the study results were extremely similar for substudy results since both drew from the same samples of students for each subgroup, students with disabilities and their typically developing peers. For example, both study and substudy results, percentages of statistically significant effect sizes for students with disabilities are very similar to those for typically developing students, around 60% for study results and about 45% for substudy results. Precision of effect size estimation was also similar for study and substudy results, with effect sizes being more precisely measured for typically developing students as compared to their peers, students with disabilities.

Table 17 provides a comparison of effect sizes for students with disabilities across study and substudy results. Most combined and disaggregated data provide similar effect sizes estimates for students with disabilities, with few exceptions. The greatest difference between effect sizes were for Helwig et al., 2002 (12), Helwig and Tindal, 2003 (13), and Schnirman, 2005 (29). These differences can be accounted for by the amount of variability between each of the effect size estimates. For example, substudy effect size estimates for Helwig et al., 2002 (12) ran from small, negative effects (-0.24) to small, positive effects (0.21) while the study effect size was trivially small and positive (0.01). Meloy et al., 2002 (26) also showed variability between study and substudy results. However, this is based on a decision not to combine effect size estimates for study results as the assessments used were across content areas. The most commonly assessed content area, math, was selected from among the possible choices of content area. This effect size estimate was, therefore, present in both study and substudy results. The same decision was not made for Fuchs et al., 2000a (10) and Schnirman, 2005 (29) as all measures used in each of these primary investigations assessed the same content area.

Table 17: *Comparison of Effect Size Estimates with Study and Substudy as Unit of Analysis - Students with Disabilities*

Unit of analysis: Study		**Unit of analysis: Substudy**	
Study name	$\overline{ES}$ [a]	**Study name**	**ES**[b]
Multiple grades			
9. Engelhard et al. (2011)	-0.24	9a. Engelhard et al.	-0.17
		9b. Engelhard et al.	-0.22
19. Laitusis (2010)	0.51	19a. Laitusis	0.64
		19b. Laitusis	0.36
20. Lee & Tindal (2000)	0.10	20a. Lee & Tindal	0.16
		20b. Lee & Tindal	0.03
28. Randall & Engelhard (2010)	-0.17	28a. Randall & Engelhard	-0.04
		28b. Randall & Engelhard	-0.24
31. Tindal (2002)	0.11	31a. Tindal	0.15
		31b. Tindal	0.07
Form &/or order effects			
12. Helwig et al. (2002)	-0.02	12a. Helwig et al.	-0.24
		12b. Helwig et al.	0.37
		12c. Helwig et al.	-0.34
		12d. Helwig et al.	0.04
		12e. Helwig et al.	0.63
		12f. Helwig et al.	-0.57
		12g. Helwig et al.	0.07
13. Helwig & Tindal (2003)[a]	0.02	13a. Helwig & Tindal	0.49
		13b. Helwig & Tindal	0.16
		13c. Helwig & Tindal	-0.39
		13d. Helwig & Tindal	0. 067
		13e. Helwig & Tindal	-0.08
		13f. Helwig & Tindal	0.23
		13g. Helwig & Tindal	-0.18
		13h. Helwig & Tindal	0.1 09
Group effects			
14. Huesman (1999)	0.25	14a. Huesman	0.37
		14b. Huesman	0.37
		14c. Huesman	0.14
Multiple years of data			
15. Janson (2002)	0.08	15a. Janson	0.27
		15b. Janson	0.13
		15c. Janson	-0.08
Multiple tests with same research participants			
10. Fuchs et al. (2000a)	0.39	10a. Fuchs et al. (a)	0.47
		10b. Fuchs et al. (a)	0.45
		10c. Fuchs et al. (a)	0.25
26. Meloy et al. (2002)	0.58	26a. Meloy et al.	0.58
		26b. Meloy et al.	1.10
		26c. Meloy et al.	1.17
		26d. Meloy et al.	1.20
29. Schnirman (2005)	0.12	29a. Schnirman	-0.03
		29b. Schnirman	0.29

[a] ES is Hedges' *g* effect size estimate for individual studies

[b] $\overline{ES}$ is Hedges' *g* mean effect size estimate

A comparison of effect sizes for typically developing students across study and substudy results is presented in Table 18. As was the case for students with disabilities, most combined and disaggregated data provide similar estimates for effect sizes, with few exceptions. The greatest difference between effect sizes were for Engelhard et al., 2011(9), Helwig et al., 2002 (12), and Randall and Engelhard, 2010 (28). These differences can be accounted for by the amount of variability between each of the effect size estimates. For example, substudy effect size estimates for Randall and Engelhard, 2010 (28) were trivially small and negative (-0.06) and small and negative (-0.38) while the study effect size was small and negative (-0.22). As was seen previously, and for the same reasons, Meloy et al., 2002 (26) also showed variability between study and substudy results.

Table 18: *Comparison of Effect Size Estimates with Study and Substudy as Unit of Analysis - Typically Developing Students*

Unit of analysis: Study		**Unit of analysis: Substudy**	
Study name	$\overline{ES}$ [a]	**Study name**	**ES**[b]
Multiple grades			
9. Engelhard et al. (2011)	-0.13	9. Engelhard et al.[a]	0.12
		9b. Engelhard et al.	-0.39
19. Laitusis (2010)	0.10	19a. Laitusis	0.14
		19b. Laitusis	0.06
20. Lee & Tindal (2000)	0.05	20a. Lee & Tindal	0.10
		20b. Lee & Tindal	0.00
28. Randall & Engelhard (2010)	-0.22	28a. Randall & Engelhard	-0.38
		28b. Randall & Engelhard	-0.06
31. Tindal (2002)	0.02	31a. Tindal	0.02
		31b. Tindal	0.03
Form &/or order effects			
12. Helwig et al. (2002)	0.01	12a. Helwig et al.	0.13
		12b. Helwig et al.	-0.24
		12c. Helwig et al.	0.21
		12d. Helwig et al.	-0.03
		12e. Helwig et al.	-0.20
		12f. Helwig et al.	0.17
		12g. Helwig et al.	0.06
13. Helwig & Tindal (2003)	0.03	13a. Helwig & Tindal	-0.03
		13b. Helwig & Tindal	-0.02
		13c. Helwig & Tindal	0.02
		13d. Helwig & Tindal	0.04
		13e. Helwig & Tindal	0.12
		13f. Helwig & Tindal	0.47
		13g. Helwig & Tindal	-0.04
		13h. Helwig & Tindal	-0.14
Group effects			
14. Huesman (1999)	0.15	14a. Huesman	0.24
		14b. Huesman	0.14
		14c. Huesman	0.03
Multiple tests with same research participants			
10. Fuchs et al. (2000a)	0.46	10a. Fuchs et al. (a)	0.73
		10b. Fuchs et al. (a)	0.68
		10c. Fuchs et al. (a)	0.08

Unit of analysis: Study		Unit of analysis: Substudy	
Study name	$\overline{ES}$ [a]	**Study name**	**ES**[b]
26. Meloy et al. (2002)	0.37	26a. Meloy et al.	0.37
		26b. Meloy et al.	0.70
		26c. Meloy et al.	0.36
		26d. Meloy et al.	0.54
29. Schnirman (2005)	-0.11	29a. Schnirman	-0.14
		29b. Schnirman	-0.08

[a] ES is Hedges' *g* effect size estimate for individual studies

[b] $\overline{ES}$ is Hedges' *g* mean effect size estimate

Patterns of effect sizes for study and substudy results are also similar (see Figures 5 and 8, graphs of Hedges' g for students with and without disabilities). Additionally, conclusions based on overall results for both studies were the same. Overall mean effect size for students with disabilities (study mean effect size = 0.36, substudy mean effect size = 0.30 for the random-effects model), albeit small, reached statistical significance, $p < 0.001$. As well, in both cases the statistically significant overall mean effect for typically developing students (study effect size = 0.19, substudy effect size = 0.17) was considered very small.

Empirically we do not appear to lose much information by combining substudy effect size estimates to produce study estimates. However, with respect to the substantive nature of the research purposes espoused in each of the primary research studies, with multiple data points and the presentation of disaggregated results in these primary research studies, the remaining research hypotheses were addressed using substudy (subgroup) within study as the unit of analysis.

Overall we may conclude that, while the evidence for providing assessment accommodations for students with disabilities is not as compelling as was hoped, students with disabilities did benefit from assessment accommodations. Thus, this leads us to examine which accommodation, or accommodations, is more effective for these students.

Research hypothesis 2 results.

Research hypothesis 2 asked if each of the four test accommodation categories, presentation, response, setting, and timing/scheduling, constituted an effective accommodation for students with disabilities. Hedges' g was used to calculate both the effect size estimates for each substudy and the mean effect size estimates for each test accommodation category.

Accommodation category: Research hypothesis 2 results.

Table 19 provides overall results examining test accommodation categories, with substudy as the unit of analysis. For the most part, only two categories of accommodations were empirically examined with any frequency: presentation and timing/scheduling. As there were so few studies exploring response and setting test accommodations, these accommodations were not subject to scrutiny at the aggregate level.

Table 19: *Comparison Between Accommodations for Students with Disabilities -* $\overline{ES}$ *, Confidence Intervals, & Q-statistics*

		Mean effect size & 95% CI (Hedges' g)					**Heterogeneity**		
Accommodation	**k**	$\overline{ES}$ ^{a}	**Std Err**^{a}	**LL**^{a}	**UL**^{a}	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
				Fixed effects					
Presentation	41	0.42	0.01	0.40	0.45	< 0.001	491.44	40	< 0.001
Response	3	0.06	0.07	-0.09	0.20	0.449	47.99	2	< 0.001
Timing/Scheduling	17	0.45	0.02	0.41	0.50	< 0.001	216.35	16	< 0.001
				Random effects					
Presentation	41	0.22	0.06	0.12	0.33	< 0.001			
Response	3	0.24	0.38	-0.50	0.98	0.525			
Setting	1	0.32	0.17	-0.02	0.66	0.061			
Timing/Scheduling	17	0.47	0.09	0.30	0.64	< 0.001			

[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

The Q-tests for the distribution of observed effect sizes for presentation and timing/scheduling accommodations, $Q_{(40)} = 491.45$ and $Q_{(16)} = 216.35$, were statistically significant ($p < 0.001$). Overall mean effect sizes for presentation and timing/scheduling accommodations under the random-effects model were 0.22 and 0.47, respectively. This indicates that there was a small positive effect for both presentation and timing/scheduling test accommodations for students with disabilities.

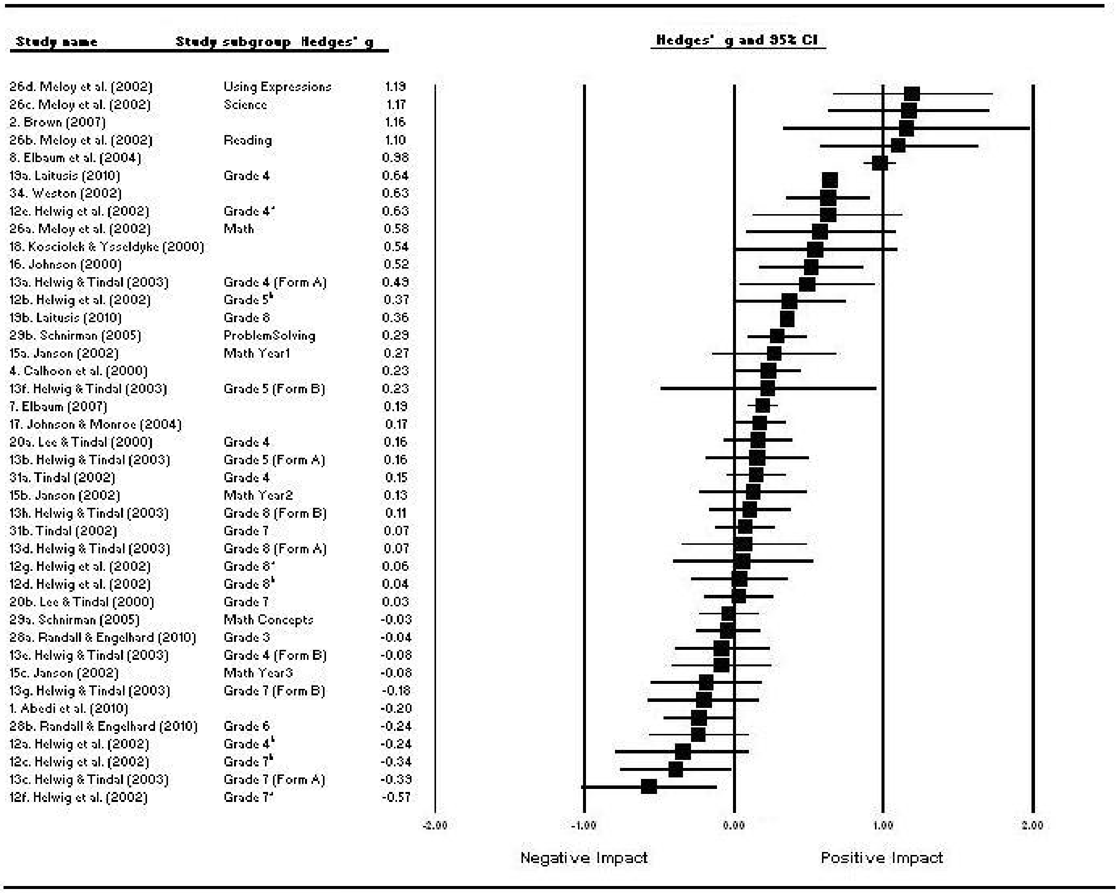
Presentation accommodations.

Forty-one effect size estimates, based on 18 studies, were calculated for the presentation accommodation (see Appendix P for effect sizes and standard errors calculated for students with disabilities by accommodation category). Effect sizes for this test accommodation ranged from -0.57 to 1.19. While there were many very small, positive effect sizes (k = 12, 29%), a sizeable portion (44%) of the positive effect size estimates were equally distributed between small (k = 7), medium (k = 6), and large (k = 5) effects. Only 15% of the effect size estimates were negative, thus pointing to the positive impact of presentation accommodations for students with disabilities. Of the total number of effect size estimates, 17 of the effects were statistically significant and 24 were not. There were 15 statistically significant positive effects and only 2 statistically significant negative effects. That is, most of the negative effects, 82% of negative effects or 22% of the total effects, for this group were not significant.

Most of the statistically significant effects, both positive and negative, ranged from small to large. Four studies spanned zero, Kosciolek and Ysseldyke, 2000 (18) with a medium effect, and Helwig et al. 2002 (12b), Janson, 2002 (15a), and Helwig and Tindal, 2003 (13f) with small effects, and did not reach statistical significance. Only 2 of the 11 small and medium negative effect sizes were statistically significant, Helwig & Tindal, 2003 (13c) and Helwig et al., 2002 (12f). Consequently, the majority of effects were both positive and reached statistical significance.

Standard errors for presentation accommodations ranged from 0.02 to 0.42 across all studies included in the analysis. The effect size estimates for Brown, 2007 (2) and Helwig and Tindal, 2003 (13f) were less precise than Laitusis, 2010 (19a) and Laitusis, 2010 (19b).

The forest plot in Figure 9 displays the effect sizes for the 41 primary studies examining presentation accommodations for students with disabilities, bounded by their respective confidence intervals.



[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

Figure 9: Forest Plot of Effect Size Estimates for Presentation Accommodations

The 95% confidence interval of effects for the weighted average effects for the 41 presentation accommodation studies, displayed in Figure 9, showed differences across presentation accommodations that were highly variable, and differed markedly by study. Figure 9 also reveals that most study effect estimates and the overall mean effect size estimate were relatively precise, with two exceptions, Brown, 2007 (2) and Helwig and Tindal, 2003 (13f). Only 13 of the study effect sizes (32%) fell inside the confidence interval for the overall mean effect, and none was fully enclosed within the confidence interval. Additionally, eight of the effect sizes, which span a portion of the overall mean effect size also spanned zero. As well, almost all very small effects, 16 of 17 or 39% of all effects, spanned zero. Since these effects are both considered very small and span zero, we can conclude that in 39% of presentation accommodation studies little or no effect was seen. Conversely, 32% (13 of 41) of all included studies had statistically significant small to large, positive effect size estimates. Consequently, we may cautiously infer that presentation accommodations have a positive but small impact ($\overline{ES} = 0.22$, $p <$ 0.001) for students with disabilities.

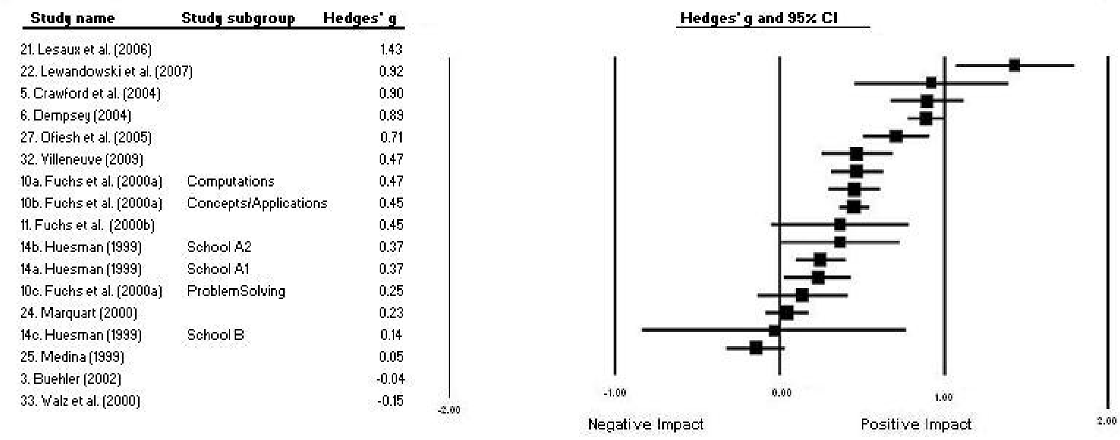
Timing/scheduling accommodations.

There were 15 positive and 2 negative effects, for a total of 17 effect size estimates for timing/scheduling test accommodations. Twelve of the effects (71%) were statistically significant, with nine of these (53%) being both significant and positive. The positive effects were large, k = 4 or 24%, medium, k = 1 or 6%, or small, k = 8 or 47% with two very small negative effects (12%). Four effects (24%) ranged from -0.15 to 0.14 and were regarded as very small.

Standard errors for timing/scheduling accommodations extended from 0.05 to 0.41 with the effect for Buehler, 2002 (3) being less precisely estimated than those for

Dempsey, 2004 (6) and Fuchs et al., 2000b (11). Forty-one percent of the standard errors were 0.09 or smaller.

Figure 10 displays a forest plot with effect sizes, bounded by a 95% confidence interval, for all 17 of the primary studies exploring timing/scheduling accommodations for students with disabilities.



Study name	Study subgroup	Hedges' g
21. Lesaux et al. (2006)		1.43
22. Lewandowski et al. (2007)		0.92
5. Crawford et al. (2004)		0.90
6. Dempsey (2004)		0.89
27. Ofiesh et al. (2005)		0.71
32. Villeneuve (2009)		0.47
10a. Fuchs et al. (2000a)	Computations	0.47
10b. Fuchs et al. (2000a)	Concepts/Applications	0.45
11. Fuchs et al. (2000b)		0.45
14b. Huesman (1999)	School A2	0.37
14a. Huesman (1999)	School A1	0.37
10c. Fuchs et al. (2000a)	ProblemSolving	0.25
24. Marquart (2000)		0.23
14c. Huesman (1999)	School B	0.14
25. Medina (1999)		0.05
3. Buehler (2002)		-0.04
33. Walz et al. (2000)		-0.15

[a] Condition order: not accommodated - accommodated
[b] Condition order: accommodated - not accommodated

Figure 10: Forest Plot of Effect Size Estimates for Timing/Scheduling Accommodations

There were few studies with negative effect sizes, one, Buehler, 2002 (3) being imprecisely estimated; i.e., having a fairly large standard error. Most effect sizes, 11 or 65%, were both positive and did not span zero. As well, there was very little variability in the weighted contribution of each study to the overall results, with no individual study being dominant in the analysis. Several of the studies, eight or 47%, fell inside the confidence interval for the overall effect, with three, Fuchs et al., 2000a (10a), Fuchs et al., 2000a (10b), and Fuchs et al., 2000b (11) fully enclosed within the confidence interval. Examination of the distribution of effect sizes about the interval midpoint shows that students with disabilities were likely to be positively impacted by timing/scheduling test accommodations.

As there were only three effect size estimates for response test accommodations it is not possible to discuss overall mean effect size. MacArthur and Cavalier, 2004 (23) with a statistically significant, large effect size of 1.13, was fairly precisely estimated (standard error = 0.17). Effect size estimates for Engelhard et al., 2011 (9a) (ES = -0.17) and Engelhard et al., 2011 (9b) (ES = -0.22) did not reach statistical significance.

There was only one empirical study of setting test accommodations, Smith, 2010 (30). The estimated effect size (0.32) while not statistically significant as the 95% confidence interval spanned zero, was fairly precisely estimated (s.e. = 0.17).

Figure 11 provides an expanded, graphical representation of the effects of different categories of test accommodations on students with disabilities.

Presentation Accommodations				Interval Midpoint	Timing-Scheduling Accommodations			
				1.45				
				1.40	21, ET, LDR, RD, 1.43			
				1.35				
				1.25	22, ET, OH, MA, 0.92			
				1.20				
	2, RA, LDR, SC, 1.16	26c, RA, LDR, SC, 1.17	26d, RA, LDR, SC, 1.20	1.15				
			26b, RA, LDR, SC, 1.10	1.10	23, SD, LD , WR, 1.14			
				1.05				
				1.00				
			8, RA, LD , MA, 0.98	0.95				
				0.90	23, ET, LDR, RD, 0.92			
				0.85	5, ET, SE, WR, 0.90	6, ET, LD , O, 0.89		
				0.80				
				0.75				
				0.70	27, ET, LDR, RD, 0.71			
				0.65				
	12e, RA, LDR, MA, 0.63	34, RA, LD , MA, 0.63	19a, RA, LDR, RD, 0.64	0.60				
			26a, RA, LDR, SC, 0.58	0.55				
		16, RA, LDR, MA, 0.52	18, RA, SE, RD, 0.54	0.50				
			13a, RA, SE, MA, 0.49	0.45	32, ET, LDR, RD, 0.47	10a, ET, LD , MA, 0.47	10b, ET, LD , MA, 0.45	11, ET, LD , RD, 0.45
				0.40				
		19b, RA, LDR, RD, 0.36	12b, RA, LDR, MA, 0.37	0.35	14b, ET, LD , RD, 0.37	14a, ET, LD , RD, 0.37	30, SA, LD , RD, 0.32	
				0.30				
		15a, RA, SE, MA, 0.27	29b, RA, LD, MA, 0.29	0.25				
	7, RA, LD , RD, 0.19	13f, RA, SE, MA, 0.23	4, RA, LDRM, MA, 0.23	0.20	10c, ET, LD , MA, 0.25	24, ET, SE, MA, 0.23		
31a, RA, SE, MA, 0.15	13b, RA, SE, MA, 0.16	20a, RA, LDR, MA, 0.16	17, SL, SE, MA, 0.17	0.15				
		13h, RA, SE, MA, 0.11	15b, RA, SE, MA, 0.13	0.10	14c, ET, LD , RD, 0.14			
	12g, RA, LDR, MA, 0.07	13d, RA, SE, MA, 0.07	31b, RA, SE, MA, 0.07	0.05	25, ET, LD , O, 0.05			
		20b, RA, LDR, MA, 0.03	12d, RA, LDR, MA, 0.04	0.00	3, ET, LDR, RD, -0.04			
		28a, RA, SE, RD, -0.039	29a, RA, LD, MA, -0.034	-0.05				
		15c, RA, SE, MA, -0.08	13e, RA, SE, MA, -0.08	-0.10	33, ET, SE, RD, -0.15			
			13g, RA, SE, MA, -0.18	-0.15	9a, CA, SE, MA, -0.17			
	12a, RA, LDR, MA, -0.24	28b, RA, SE, RD, -0.24	1, ST, SE, RD, -0.20	-0.20	9b, CA, SE, MA, -0.22			
				-0.25				
			12c, RA, LDR, MA, -0.34	-0.30				
			13c, RA, SE, MA, -0.39	-0.35				
				-0.40				
				-0.45				
				-0.50				
			12f, RA, LDR, MA, -0.57	-0.55				
				-0.60				

RA = Read Aloud	Response Accommodations
ST = Segmented Text	Setting Accommodations
SL = Simplified Language	
SA = Special Acoustics	E = Elementary
ET = Extended Time	EM = Elementary & Middle
CA = Calculator Use	M = Middle
SD = Scribe for dictation	MS = Middle & Secondary
	S = Seconday
LD = Learning Disability	C = College
LDR = Learning Disability (Reading)	A = Adult
LDRM = Learning Disabilit (Reading & Math)	MA = mathematics
	RD = reading/language arts
OH = Other Health Impaire	SC = science
SE = Special Education	WR = writing
	O = Other

Note: Each cell contains the (1) study number, (2) assessment accommodation, (3) disability classification, (4) assessment subject, and (5) effect size estimate (Hedges':g)

Figure 11: Graph of Hedges' g Effect Size Estimates for Presentation Accommodations Compared to Timing/Scheduling Accommodations

Comparison of accommodation categories.

The only two test accommodation categories with empirical data to warrant examination of overall mean effect size estimates, presentation and timing/scheduling, have dissimilar distributions and appear to be differentially effective, as seen in Figure 11.

Empirical research; as represented by individual ‘boxes’ in the figure; for the presentation accommodation was spread across math, science, and reading content. Research appeared to be conducted equally across the two major groups of students with disabilities, students with learning disabilities and students receiving special education services, although in just under one-quarter of all studies, medium to large effect size estimates were for research participants with learning disabilities. The distribution of effect sizes about the interval midpoints visually confirms that students with disabilities are likely to be positively impacted by presentation accommodations, with just under one-half of the effect sizes being small (0.20 to 0.50) to large (0.80 and greater). While there were many extremely small effects (39%) only 15% of all effects were negative. This suggests that presentation accommodations, for the most part, positively affected the test scores of students with disabilities, with a statistically significant overall mean effect size estimate of 0.22.

The distribution of effect size estimates for timing/scheduling accommodations in Figure 11 was more compelling than that for presentation accommodations. Almost all timing/scheduling accommodations were used with students with learning disabilities. Use of these accommodations was spread across almost all content areas: math, reading, and writing. With 77% of effect sizes producing at least a small effect (0.20 to 0.50), evidence points to the positive impact of timing/scheduling accommodations for students with disabilities.

Overall, the mean effect sizes for the presentation ($\overline{ES}$ = 0.22) and timing/scheduling ($\overline{ES}$ = 0.47) test accommodations were small, albeit statistically significant. To provide greater clarification of these results, each accommodation category was further broken down. Results for specific accommodations; e.g., read-aloud test accommodations, follow.

Specific accommodation category: Research hypothesis 2 results.

Table 20 presents overall results for the specific test accommodations investigated. As was previously noted, only presentation and timing/scheduling test accommodation categories were tested with any frequency. Specific test accommodations falling under these two accommodation categories, with acceptable numbers of primary studies to warrant further investigation, were read aloud and extended time. The medium used for read-aloud accommodations varied, ranging from computer presentation to unfamiliar proctor reading the test questions and responses, or the entire test, aloud. The extended time test accommodation was also varied and ranged from 20 minutes to three days. As other specific test accommodations only had one or two representative studies, results from these studies were not subjected to further examination.

Table 20: *Comparison between Specific Accommodations -* $\overline{ES}$ *, Confidence Intervals, & Q-statistics*

		Mean effect size & 95% CI (Hedges' g)					**Heterogeneity**		
Accommodation	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
				Fixed effects					
Presentation									
ReadAloud	39	0.43	0.01	0.41	0.46	< 0.001	472.47	38	< 0.001
Response									
Calculator	2	-0.19	0.08	-0.35	-0.03	0.02	0.11	1	0.744
Timing/Scheduling									
ExtendedTime	17	0.45	0.02	0.41	0.50	< 0.001	216.35	16	< 0.001
				Random effects					
Presentation									
ReadAloud	39	0.24	0.06	0.13	0.35	< 0.001			
SegmentedText	1	-0.20	0.19	-0.58	0.17	0.285			
SimplifiedLanguage	1	0.17	0.09	0.00	0.35	0.057			
Response									
Calculator	2	-0.19	0.08	-0.35	-0.03	0.021			
Dictation(scribe)	1	1.14	0.17	0.80	1.48	< 0.001			
Setting									
SpecialAcoustics	1	0.32	0.17	-0.02	0.66	0.061			
Timing/Scheduling									
ExtendedTime	17	0.47	0.09	0.30	0.64	< 0.001			

[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

The Q-tests for the distributions of observed effect sizes for the read-aloud and extended-time test accommodations, $Q_{(38)} = 472.47$ and $Q_{(16)} = 216.35$, were statistically significant ($p < 0.001$). Overall mean effect sizes for the read-aloud and extended-time accommodations under the random-effects model were 0.24 and 0.47, respectively, indicating a small and small-to-medium positive effect for students with disabilities using these test accommodations.

Read-aloud accommodation.

While effect size information for all specific accommodations was included in the meta-analysis, due to the scant number of studies included for segmented text, simplified language, and calculator use accommodations, only read-aloud and extended-time accommodations were aggregated. Effect size estimates for the read-aloud test accommodation (k = 39) ranged from -0.57 to 1.19 (see Appendix P for effect sizes and standard errors calculated for students with disabilities by specific accommodation). There were 29 positive effects (74%) and 10 negative effects (26%). Seventeen (44%) of these effect size estimates were statistically significant while 22 (56%) were not. There were 15 statistically significant positive effects (52% of positive effects, 38% of total effects), with only 2 statistically significant negative effects (20% of negative effects, 5% of total effects).

Examining the effect size categories, we find 5 large (13%), 6 medium (15%), and 7 small (18%) positive effects with 1 medium (3%) and 4 small (10%) negative effects. Less than one-quarter of effect size estimates (k = 4, 24%), which ranged between -0.18 and 0.19, were considered extremely small.

The standard errors for the read-aloud accommodation ranged from 0.02 to 0.42. As was seen with the analysis of test accommodation categories, the standard errors for the Brown, 2007 (2) and Helwig and Tindal, 2003 (13f) studies were less precise than those for Laitusis, 2010 (19a) and Laitusis, 2010 (19b). Approximately 50% of these standard errors were less than 0.17.

Extended-time accommodation.

Results for the extended-time accommodation were summarized previously (see pp. 188 - 190). As was seen with overall effect size estimates for presentation and timing/scheduling test accommodations categories, overall mean effects for read-aloud (0.24) and extended-time (0.47) test accommodations were small, albeit statistically significant, providing evidence of the positive impact of these specific test accommodations for students with disabilities.

Ancillary analysis: Students with learning disabilities versus students requiring special education services.

An ancillary analysis based on category of disability was performed. Two main categories of students with disabilities were represented in the majority of primary studies conducted; students with learning disabilities and students receiving special education services. Thus, examination of test accommodation effects for these two groups of students was conducted. It must be noted that students receiving special education service does include students with learning disabilities so there is some overlap in the two groups under investigation.

Overall results, mean effect size estimates, comparing the effects of test accommodations for students with learning disabilities with those for students receiving special education services are presented in Table 21. The Q-tests for the distributions of observed effect sizes for students with learning disabilities and students receiving special education services, $Q_{(22)} = 100.90$ and $Q_{(37)} = 492.69$, were statistically significant ($p < 0.001$). Overall, a statistically significant, positive mean effect was found for students with learning disabilities ($\overline{ES} = 0.42$, $p < 0.001$), while the mean effect size for students receiving special education services was very small, 0.07, and not statistically significant ($p = 0.305$).

Table 21: *Comparison between Students with Learning Disabilities & Receiving Special Education Services -* $\overline{ES}$*, Confidence Intervals, & Q-statistics*

		Mean effect size & 95% CI (Hedges' g)					Heterogeneity		
Comparison group	**k**	$\overline{ES}$ ^{a}	**Std Err**^{a}	**LL**^{a}	**UL**^{a}	***p* (ES)**	**Q-value**	**df (Q)**	***p* (Q)**
				Fixed effects					
Special Education	23	0.07	0.03	0.01	0.12	0.017	100.90	22	< 0.001
Learning Disabilities	38	0.48	0.01	0.46	0.51	< 0.001	492.69	37	< 0.001

		Mean effect size & 95% CI (Hedges' g)					Heterogeneity		
Comparison group	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p* (ES)**	**Q-value**	**df (Q)**	***p* (Q)**
				Random effects					
Special Education	23	0.07	0.06	-0.06	0.19	0.305			
Learning Disabilities	38	0.42	0.05	0.32	0.52	< 0.001			

[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Students with learning disabilities.

Table 22 presents mean effect size estimates comparing specific test accommodation categories for students with learning disabilities. The two categories of test accommodations examined most frequently for students with learning disabilities were read aloud and extended time. The Q-values for these test accommodations, $Q_{(22)}$ = 329.60 and $Q_{(12)}$ = 147.92, respectively, were statistically significant ($p < 0.001$). The overall effect size estimates for the read-aloud and extended-time accommodations were 0.36 and 0.48, respectively, indicating that there were small positive effects for students with learning disabilities using these two test accommodations. Results for dictation, although reaching statistical significance, and special acoustics were not subject to further examination as each had only one representative study.

Table 22: *Comparison between Accommodations (Students with Learning Disabilities) -* $\overline{ES}$ *Estimates, Confidence Intervals, & Q-statistics*

		Mean effect size & 95% CI (Hedges' g)					Heterogeneity		
Accommodation	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
				Fixed effects					
Read aloud	23	0.48	0.01	0.45	0.51	0.000	329.60	22	< 0.001
Extended time	13	0.48	0.02	0.43	0.53	0.000	147.92	12	< 0.001
				Random effects					
Read aloud	23	0.36	0.07	0.22	0.50	0.000			
Extended time	13	0.48	0.09	0.30	0.65	0.000			
Dictation (scribe)	1	1.14	0.17	0.80	1.48	0.000			
Special acoustics	1	0.32	0.17	-0.02	0.66	0.061			

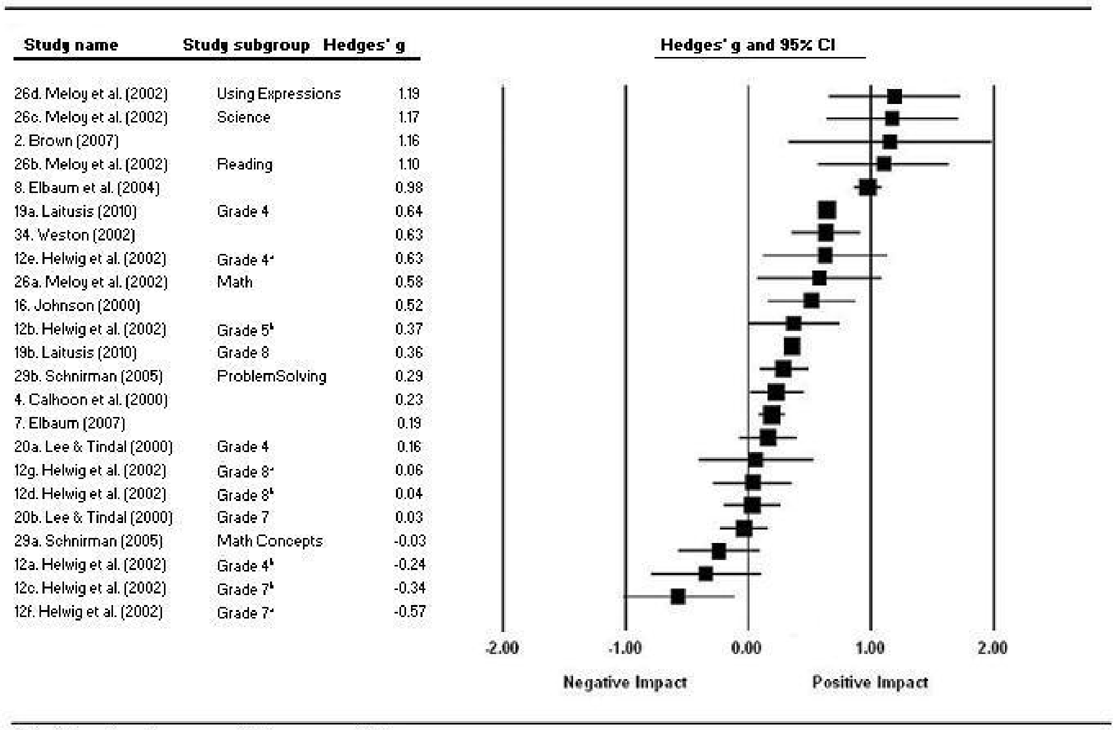
[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

There were 23 effect sizes estimated for the read-aloud test accommodation that ranged from -0.57 to 1.19 (see Appendix P for effect sizes and standard errors calculated for students with learning disabilities by specific accommodation). The majority of effect size values (61%) were above 0.20 and were equally distributed between small (k = 4, 17%), medium (k = 5, 22%), and large (k = 5, 22%) effects. Less than one-quarter (17%) of the effect sizes were negative. Two-thirds of effect size estimates (k = 15, 66%) were statistically significant. Of these, almost all (k = 14, 93%) of statistically significant effects (61% of total effects) were positive effects, indicating that the read-aloud test accommodation positively affected scores for students with learning disabilities.

There were few effects that did not reach statistical significance (k = 8, 35%) and of these, one was a small, positive effect (Helwig et al. 2002 (12b)) and two were small, negative effects (Helwig et al. 2002 (12a), Helwig et al. 2002 (12c)). Five effect size estimates were extremely small, with only one of these being statistically significant and positive (Elbaum, 2007 (7)).

Standard errors for the read-aloud accommodation ranged from 0.02 to 0.42. As was previously noted, standard errors for the Brown, 2007 (2) study was less precise than those for Laitusis, 2010 (19a) and Laitusis, 2010 (19b). Approximately 50% of these standard errors were less than 0.17.

The 23 effect sizes for the primary studies examining read-aloud test accommodations for students with learning disabilities, bounded by their respective confidence intervals, are presented in the forest plot in Figure 12.



Study name	Study subgroup	Hedges' g
26d. Meloy et al. (2002)	Using Expressions	1.19
26c. Meloy et al. (2002)	Science	1.17
2. Brown (2007)		1.16
26b. Meloy et al. (2002)	Reading	1.10
8. Elbaum et al. (2004)		0.98
19a. Laitusis (2010)	Grade 4	0.64
34. Weston (2002)		0.63
12e. Helwig et al. (2002)	Grade 4[a]	0.63
26a. Meloy et al. (2002)	Math	0.58
16. Johnson (2000)		0.52
12b. Helwig et al. (2002)	Grade 5[b]	0.37
19b. Laitusis (2010)	Grade 8	0.36
29b. Schnirman (2005)	ProblemSolving	0.29
4. Calhoon et al. (2000)		0.23
7. Elbaum (2007)		0.19
20a. Lee & Tindal (2000)	Grade 4	0.16
12g. Helwig et al. (2002)	Grade 8[a]	0.06
12d. Helwig et al. (2002)	Grade 8[b]	0.04
20b. Lee & Tindal (2000)	Grade 7	0.03
29a. Schnirman (2005)	Math Concepts	-0.03
12a. Helwig et al. (2002)	Grade 4[b]	-0.24
12c. Helwig et al. (2002)	Grade 7[b]	-0.34
12f. Helwig et al. (2002)	Grade 7[a]	-0.57

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

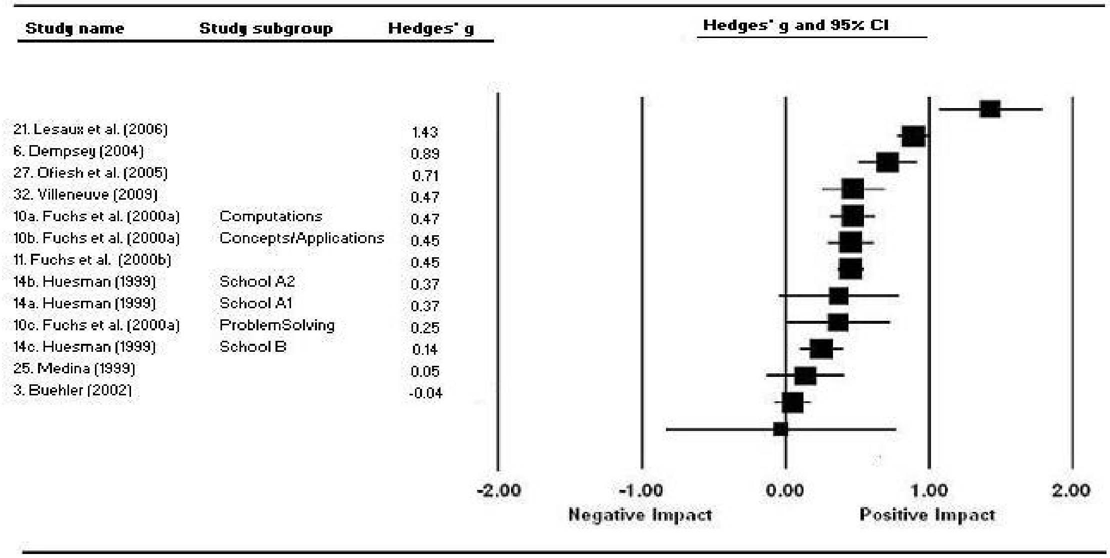
Figure 12: Forest Plot of Effect Size Estimates for Read-Aloud Accommodations for Students with Learning Disabilities

Figure 12 illustrates the variability between the effect size estimates for studies examining read-aloud test accommodations for learning disabled students. Most effect sizes estimates were relatively precise, as demonstrated by the 95% confidence interval bounding the effect size estimates, with Brown, 2007 (2) being the least precise of the studies examined. Nine of the studies (39%) fell within the confidence interval for the overall mean effect with one, Laitusis, 2010 (19b), fully enclosed within the interval. Of the effect size estimates spanning zero, all but one (Helwig et al. 2002 (12b)), were very small effects (5 of 23, 22%) or small and negative (2 of 23 or 9%). Thus, in 31% of read-aloud accommodation studies we see a small negative, or no, effect. However, 61% (14 of 23) of all included studies exhibited statistically significant small to large, positive effect size estimates. As a result, we may infer that read-aloud accommodations have a positive impact ($\overline{ES}$ = 0.36, $p < 0.001$) on students with learning disabilities.

Extended time yielded 12 positive and 1 negative effect size estimates. Just over two-thirds (k = 9, 69%) of these effects were statistically significant, with all nine being both significant and positive. The positive effects were large, k = 2 or 15%, medium, k = 1 or 8%, or small, k = 7 or 54%. Three effects (24%) ranged from -0.15 and 0.14 were regarded as extremely small.

Standard errors for extended time ranged from 0.05 to 0.41, with the effect for Buehler, 2002 (3) being less precisely estimated than those for Dempsey, 2004 (6) and Fuchs et al., 2000b (11). Fifty-four percent of standard errors were 0.10 or smaller.

Figure 13 provides a forest plot with effect sizes, bounded by a 95% confidence interval, for the 13 primary studies investigating extended-time accommodations for students with learning disabilities.



[a] Condition order: not accommodated - accommodated
[b] Condition order: accommodated - not accommodated

Figure 13: Forest Plot of Effect Size Estimates for Extended-Time Accommodations for Students with Learning Disabilities

For the extended-time accommodation there were few studies with negative effect sizes. One study, Buehler, 2002 (3), was imprecisely estimated; i.e., having a fairly large standard error, as demonstrated in Figure 13**.** The majority of effect size estimates, 9 or 69%, are both positive and do not span zero. For the most part, there was very little variability in the weighted contribution of each study to the overall results. However, the Buehler, 2002 (3) study was not weighted as heavily as other studies included in the analysis. Several of the studies, six or 46%, fall inside the confidence interval for the overall effect, with three, Fuchs et al., 2000a (10a), Fuchs et al., 2000a (10b), and Fuchs et al., 2000b (11) fully enclosed within the confidence interval. Inspection of the effect size distribution about the interval midpoint shows that students with learning disabilities were likely to be positively affected by extended time.

Students receiving special education services.

Although the overall mean effect size estimates for students receiving special education services were very small and statistically non-significant, further analysis of test accommodations was conducted to provide a better understanding of the effects of test accommodations on this group of students, as well as a comparison to students with learning disabilities. Table 23 presents overall results for specific test accommodations for students receiving special educations services. Further analysis were not conducted for extended-time, segmented text, simplified language, and calculator-use test accommodations as three or fewer primary studies were used to calculate overall mean effect size estimates for these accommodations. Subsequently only information for the read-aloud test accommodation is provided.

The Q-test for the distribution of observed effect sizes for the read-aloud test accommodation, $Q_{(15)} = 23.68$, was not statistically significant ($p = 0.071$). While there was no statistical cause to suspect heterogeneity for reading test accommodations, results of previous analyses and substantive thinking led us to pursue the random-effects model to estimate overall mean effect size for this specific test accommodation. The overall mean effect for the read-aloud accommodation under the random-effects model for students with receiving special education services was both very small and not statistically significant ($\overline{ES} = 0.04$, k = 16, $p = 0.48$).

Table 23: *Comparison Between Accommodations (Students Receiving Special Education Services) -* $\overline{ES}$ *, Confidence Intervals, & Q-statistics*

		Mean effect size & 95% CI (Hedges' g)					**Heterogeneity**		
Accommodation	**k**	$\overline{ES}$ [a]	**Std Err**[a]	**LL**[a]	**UL**[a]	***p*(ES)**	**Q-value**	**df (Q)**	***p*(Q)**
				Fixed effects					
Read aloud	16	0.03	0.04	-0.05	0.11	0.43	23.68	15	0.071
Extended time	3	0.26	0.06	0.15	0.37	0.00	52.27	2	< 0.001
Calculator use	2	-0.19	0.08	-0.35	-0.03	0.02	0.11	1	0.744
				Random effects					
Read aloud	16	0.04	0.05	-0.06	0.14	0.48			
Extended time	3	0.33	0.30	-0.26	0.92	0.28			
Segmented text	1	-0.20	0.19	-0.58	0.17	0.28			
Simplified language	1	0.17	0.09	0.00	0.35	0.06			
Calculator Use	2	-0.19	0.08	-0.35	-0.03	0.02			

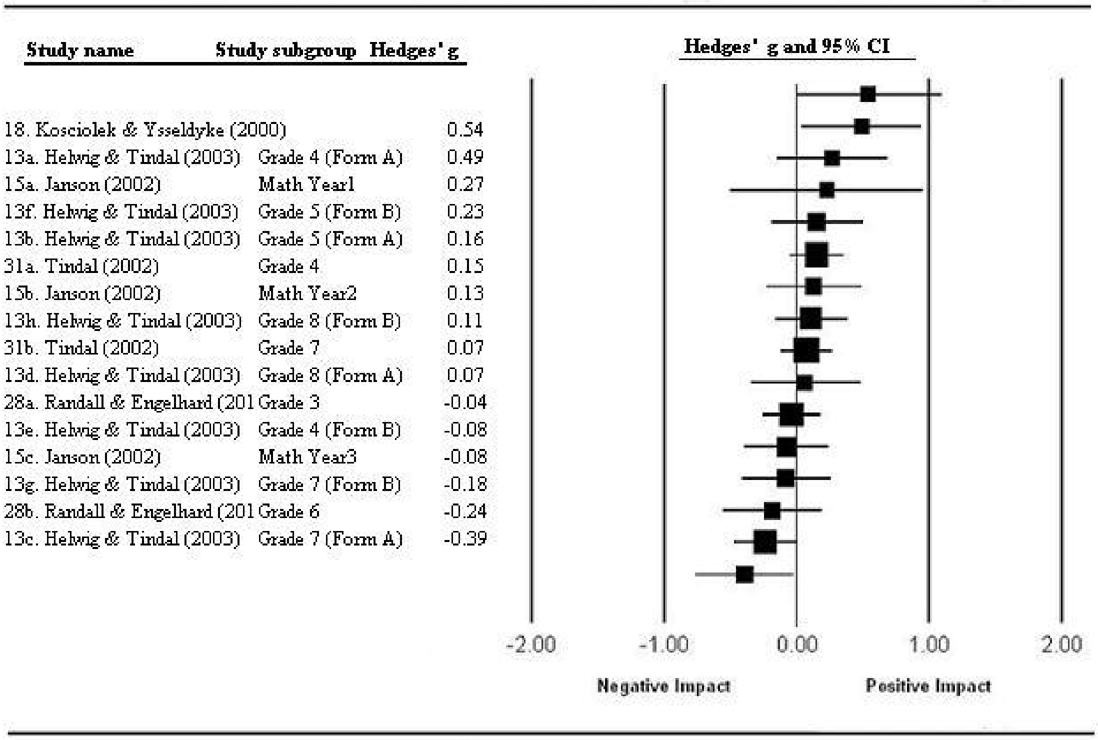
[a] $\overline{ES}$ is Hedges' *g* mean effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Effect size estimates for the read-aloud test accommodation (k = 16) ranged from -0.39 to 0.54 (see Appendix P for effect sizes and standard errors calculated for students requiring special education services by specific accommodation). Almost two-thirds of the effect size estimates (k = 10, 63%) were positive, and just over one-third (k = 6, 37%) were negative. Only 13% (2) of these effect size estimates were statistically significant. Of the statistically significant effects, 1 was positive (1% of positive effects, 7% of total effects) and 1 was negative (17% of negative effects, 7% of total effects).

There were 1 medium (6%) and 3 small (19%) positive effects, with 2 small (13%) negative effects. Almost two-thirds of all effect size estimates (k=10) which ranged between -0.18 and 0.16, were considered extremely small.

The standard errors for the read-aloud accommodation ranged from 0.10 to 0.37. The standard errors for the Helwig and Tindal, 2003 (13f) study was less precise than those Tindal, 2002 (31a) and Tindal, 2002 (31b). Approximately 50% of these standard errors were less than 0.18.

Effect size estimates for the 16 primary studies examining read-aloud test accommodations for students receiving special education services, bounded by their respective confidence intervals, are presented in the forest plot in Figure 14.



Study name	Study subgroup	Hedges' g
18. Kosciolek & Ysseldyke (2000)		0.54
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	0.49
15a. Janson (2002)	Math Year1	0.27
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.23
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	0.16
31a. Tindal (2002)	Grade 4	0.15
15b. Janson (2002)	Math Year2	0.13
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	0.11
31b. Tindal (2002)	Grade 7	0.07
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.07
28a. Randall & Engelhard (201	Grade 3	-0.04
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	-0.08
15c. Janson (2002)	Math Year3	-0.08
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.18
28b. Randall & Engelhard (201	Grade 6	-0.24
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	-0.39

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

Figure 14: Forest Plot of Effect Size Estimates for Read-Aloud Accommodations for Students Receiving Special Education Services

The majority of effect size estimates for the read-aloud accommodation (63%) for this group of students was very small. Overall, only one estimate, Helwig and Tindal, 2003 (13f), was imprecisely estimated; i.e., having a fairly large standard error, as demonstrated in Figure 16. As well, there was little variability in the weighted contribution of each study to the overall results. Just over one-third of the studies, six or 38%, fell inside the confidence interval for the overall mean effect size estimate, although none was fully enclosed within the confidence interval. Inspection of the effect size distribution shows that students receiving special education services were not likely to be affected, either positively or negatively, by the extended-time test accommodation.

Primary research included a larger variety of test accommodations for students requiring special education services as compared with their learning-disabled counterparts. As students receiving special education services, generally speaking, are considered to be a more heterogeneous group of students than students with learning disabilities are, this would be expected. On the other hand, primary research conducted using students with learning disabilities, being considered a more homogenous group relatively speaking, used a more targeted approach examining test accommodations focusing more frequently on read-aloud and extended-time test accommodations. Overall we saw a small and statistically significant impact for read-aloud ($\overline{ES} = 0.24$, $p < 0.001$) and extended-time ($\overline{ES} = 0.47$, $p < 0.01$) test accommodations for students with disabilities. However, when we disaggregate this group we see that the effects were intensified for students with learning disabilities ($\overline{ES} = 0.36$, $p < 0.001$ for read aloud; 0.48, $p < 0.001$ for extended time) while negligible, and not statistically significant, for students receiving special education services ($\overline{ES}$ = 0.04, p = 0.48 for read aloud; N/A for extended time).

As previously discussed, students receiving special education services also include students with learning disabilities so there is some overlap. However, it does appear that the more specific we can be regarding type of disability, the better able we are to target appropriate accommodations that have a positive and statistically significant impact.

Results for the Meta-regression Analyses

With statistically significant overall mean effect size estimates for students with disabilities ($\overline{ES}$ = 0.30, k = 62) and typically developing students ($\overline{ES}$ = 0.17, k = 57), and statistically significant Q-values of $Q_{(61)} = 782.27$ and $Q_{(56)} = 512.14$, respectively, we see that there is heterogeneity beyond that expected for sampling variation. The results of studies of test accommodations examined were not universally and uniformly effective. To more fully examine the unexplained variation, potential moderators for primary and specific test accommodations were identified, and meta-regression analyses were performed. Additionally, meta-regression, rather than a comparison of the mean effect size for the each of the types of test accommodations, more effectively helps answer the question of which type of test accommodation more effectively removes construct-irrelevant variance from the test scores for students with disabilities.

Use of the random-effects meta-regression model was decided *a priori*. It was expected that the impact of the categorization variables would capture some, not all, of the true variation among the estimated effect sizes, providing impetus for using the random-effects model (Borenstein et al., 2009). Additionally, the Q-values for all meta-analyses were statistically significant and use of the random-effects model is consistent with assumptions regarding the distribution of the effect sizes across the studies collected (Wilson, 2006, ppt). For the random-effects model the effect size “is the mean of the true effect sizes for all studies with a given value of the [moderators]” (Borenstein et al., 2009, p. 195). It should be noted that for the random-effects meta-regression model, as opposed to the fixed-effects model, weights assigned to each study are more moderate, confidence intervals are wider, and there is less likelihood of moderator p-values reaching statistical significance. Using Wilson’s meta-regression macro (2007: metareg.sps), the random-effects model was estimated via iterative maximum likelihood.

For the purposes of the meta-regression, both test accommodation and disability classification were re-categorized. Students requiring special education services, special education, and students with disabilities, learning disability, comprised the disability classification. Learning disability in reading and learning disability in reading and math constituted the learning disability category. Assessment accommodation was re-categorized with the aggregate categories of presentation and timing (timing/scheduling): segmented text and read aloud constituted the presentation category and extended time formed the timing category. When there were more than two levels for a categorical variable, indicator variables were created. As well, to provide a more detailed examination of the potential effect of the selected moderators, the data were differentially split and three datasets were created: all test accommodation data for students with disabilities, timing and presentation accommodation data for students requiring special education services, and timing and presentation data for students with learning disabilities. As there were so little data for setting (k = 1) and response (k = 3) test accommodation categories, representative studies for setting and response categories were dropped from the second and third datasets.

Separate meta-regressions analyses were run for the following conceptual groupings:

- Researcher-manipulated variable directed towards reducing construct-irrelevant variance for students with disabilities; i.e., test accommodation
- Population description; i.e., descriptions for students with disabilities
- Assessment description; i.e., assessment content and assessment format
- Dissemination; i.e., type of publication, publication year

Separate meta-regressions were performed to evaluate differential predictions of the effectiveness of assessment accommodations, represented by the effect sizes for each included study. To better understand whether each variable set differentially contributed to the overall model, each variable set was run independently. The information was compared to the model that included all the moderators in an effort to understand whether the overall model, or one or more individual variable sets, better explained the heterogeneity in the model. For example, it may be that a single variable set such as the researcher-manipulated variable, test accommodation, really does explain the largest portion of the variability in the model, while other variable sets contribute negligible information. This effect becomes less obvious for the overall model if the researcher -manipulated variable interacts with one of the assessment description variables, such as test format. Hedges’ g effect size estimate (ES) was the dependent variable. Type of test accommodation was expected to have an effect on efficacy of test accommodations. A meta-regression using predictors from all the variable groupings, and an examination of the overall model, was conducted following the initial set of analyses.

Meta-regression research hypothesis.

The current study addressed the following hypothesis for the meta-regression portion of the current research:

> Research Hypothesis 3: Which type of accommodation(s)–Presentation, Response, Setting, or Timing/Scheduling–more effectively remove construct-irrelevant variance from target students’ test scores?

Research hypothesis 3 results.

To effectively address the research hypothesis posed, mean effect size, R^2, variable weights, and the significance for each variable set were calculated. In an effort to understand the impact of the variable sets, and to examine whether variable sets explained more of the heterogeneity in the model as a set or as an independent variable, the same information was calculated for the overall model where each variable could be examined individually. Additionally, to investigate the amount of variance explained by the researcher-manipulated variable, test accommodation, the model was run excluding test accommodation and results were compared to the full model. Further, only statistically significant variables from the overall model were entered into the equation to evaluate the amount of variance that could be explained.

Effect of test accommodation on test scores for students with disabilities.

Table 24 provides results of four meta-regressions of effect size on conceptual variable sets. Effect size, proportion of variance explained (R^2), residual variance test significance (Q), and individual variable unstandardized (b) and standardized (β) beta weights, together with p-values, for each variable set are listed in Table 24.

Table 24: *Random-effects Model for Students with Disabilities - All Data*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Each Variable Set

Variable	k	Mean ES	R^2	p ($Q_{residual}$)	b	β	p(b)
Researcher manipulated variable		0.30	0.07	0.261			
Test Accommodation (Timing)	17				0.22	0.24	0.297
Test Accommodation (Presentation)	41				-0.02	-0.02	0.940
Population Description		0.30	0.22	0.216			
Disability Classification	38[a]				0.27	0.31	0.006
Grade Level/s (Elementary)	27				-0.28	-0.33	0.085
Grade Level/s (Middle school)	29				-0.40	-0.48	0.011
Dissemination		0.30	0.01	0.252			
Publication Year	62				0.00	-0.04	0.756
Publication Type (Journal)	40				0.15	0.17	0.371
Publication Type (Dissertation)	15				0.11	0.11	0.561
Test Characteristics		0.32	0.35	0.560			
Test Content (Math)	37				-0.67	-0.80	0.000
Test Content (Reading/LA)	18				-0.37	-0.40	0.026
Test Format	47[b]				-0.23	-0.23	0.043

[a] total for students with learning disabilities
[b] total for multiple-choice format

Population description and test characteristic variable sets explained the greatest amounts of variability for change in test score, R^2=0.22 and R^2 =0.35, respectively. The remaining variable sets explained relatively little variance: R^2 =0.07 and R^2 =0.01, for researcher-manipulated variables and dissemination. Both population description and test characteristic variable sets had statistically significant beta weights. For population description the disability classification, students with learning disabilities as compared to students receiving special education services, and middle school grade level, students in middle school as compared to students in elementary school and students in secondary school or college, were statistically significant. All test characteristic variables, math content (math content as compared to reading and other test contents), reading content (reading content as compared to math and other test contents), and test format (multiple choice format as compared to other test formats) were statistically significant. Thus, population description and test characteristic moderator variables were related to the effect size. With $Q_{(residual)}$ values that were not statistically significant we can conclude that the moderators were related to effect size and aid in explaining the heterogeneity seen in effect sizes.

Table 25 provides the overall model when all moderator variables were entered into the meta-regression. Entering all the variables into a regression model yielded statistically significant results for two of the three test characteristic moderator variables, math and reading content. The overall model explained more of the unique variability for change in test score, R^2=0.42, than did any of the separate moderator variable sets. When only statistically significant moderator variables were entered into the regression equation, math and reading content, the R^2 fell to 0.28 (see Appendix Q). When all but the researcher-controlled, test accommodation, moderator variables were entered into the meta-regression equation the R^2 was also 0.42 (see Appendix Q), thus it appears that test accommodations, timing versus all other accommodations and presentation versus all other accommodations, does not provide additional information to explain the heterogeneity in effect sizes. None of the test accommodation categories appears to uniquely effectively remove construct-irrelevant variance to explain the change in test score for students with disabilities, nor do the test accommodations appear to provide an explanation for the improvement seen.

Table 25: *Random-effects Model for Students with Disabilities - All Data*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Overall Model

Variable	**k**	**Mean ES**	**R^2**	p ($Q_{residual}$)	b	β	p(b)
Overall Model		0.32	0.42	0.459			
Test Accommodation (Timing)	17				0.12	0.13	0.687
Test Accommodation (Presentation)	41				0.03	0.04	0.925
Disability Classification	38[a]				0.16	0.19	0.138
Grade Level/s (Elementary)	27				-0.02	-0.03	0.907
Grade Level/s (Middle school)	29				-0.09	-0.10	0.646
Publication Year	62				0.00	0.02	0.875
Publication Type (Journal)	40				-0.01	-0.01	0.960
Publication Type (Dissertation)	15				-0.19	-0.20	0.345
Test Content (Math)	37				-0.63	-0.74	0.001
Test Content (Reading/LA)	18				-0.38	-0.42	0.043
Test Format	47[b]				-0.14	-0.14	0.367

[a] total for students with learning disabilities
[b] total for multiple-choice format

Effect of timing and presentation accommodations on test scores for students with disabilities.

Change in student test score, proportion of variance explained (R^2), residual variance test significance (Q), and individual variable unstandardized (b) and standardized (β) beta weights, together with p-values, for each variable set are presented in Table 26. Only studies containing data for timing and presentation accommodations are included in the analysis as there were few setting (k = 1) and response (k = 3) studies.

Table 26: *Random-effects Model for Students with Disabilities - Timing & Presentation Accommodation Data Only*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Each Variable Set

Variable	**k**	**Mean ES**	**R^2**	p ($Q_{residual}$)	b	β	p(b)
Researcher manipulated variable		0.30	0.07	0.274			
Test Accommodation	17[a]				0.24	0.27	0.027
Population Description		0.30	0.17	0.207			
Disability Classification	36[b]				0.24	0.29	0.017
Grade Level/s (Elementary)	25				-0.19	-0.23	0.262
Grade Level/s (Middle school)	28				-0.33	-0.40	0.053
Dissemination		0.30	0.02	0.236			
Publication Year	58				0.01	0.04	0.739
Publication Type (Journal)	37				0.13	0.15	0.431
Publication Type (Dissertation)	14				0.09	0.09	0.630
Test Characteristics		0.30	0.32				
Test Content (Math)	35				-0.65	-0.78	0.000
Test Content (Reading/LA)	17				-0.32	-0.35	0.063
Test Format	47[c]				-0.26	-0.24	0.034

[a] total for timing accommodation
[b] total for students with learning disabilities
[c] total for multiple-choice format

As in the previous meta-regression analysis, population description, R^2=0.17, and test characteristic, R^2 =0.32, variable sets explained the greatest amounts of variability for mean increase in test score. Researcher-manipulated variables, R^2 =0.07, and dissemination, R^2 =0.02 explained little of the heterogeneity. Statistically significant beta weights were found for population description and test characteristic variable sets. Disability classification (b = 0.24, β= 0.29) was statistically significant (p = 0.017) and positively related to the change in test score. Math content (b = -0.65, β= -0.78) was statistically significant (p < 0.001) and negatively related to effect size, as was test format (b = -0.26, β= -0.24, p = 0.034). That is to say, students with disabilities did not perform as well in an accommodated condition if they were being tested in math content. While the researcher-manipulated variable, test accommodation (timing versus presentation), was statistically significant (p = 0.027), the R^2 was very small (0.07) and did not contribute much to explaining the heterogeneity in the model. Thus, population description and test characteristic moderator variables are related to the effect size. As previously seen, the $Q_{(\text{residual})}$ values were not statistically significant, thus the moderators were considered to be related to effect size and provided information to aid in explaining the heterogeneity seen.

An overall model for the meta-regression is provided in Table 27. When all of the moderator variables were entered into the meta-regression model only one of the moderator variables, math content, was statistically significant. The overall model explained more of the heterogeneity for mean increase in test score, R^2=0.38, than any of the moderator variable sets. When the single statistically significant moderator variable, math content, was entered into the meta-regression equation the R^2 fell to 0.16 (see Appendix R). When the test accommodation moderator variable was not entered into the meta-regression model the R^2 remained the same, 0.38 (see Appendix R). As was seen when all test accommodations were examined previously, test accommodations do not help explain variance in the model. Thus, neither timing nor presentation test accommodations effectively remove construct-irrelevant variance for students with disabilities, nor do they aid in explaining any improvement seen.

Table 27: *Random-effects Model for Students with Disabilities - Timing & Presentation Accommodation Data Only*
Mean Effect Size, R^2**,** Significance of the Residual Q, Variable Weight and Significance for the Overall Model

Variable	**k**	**Mean ES**	**R^2**	p **(Q$_{residual}$)**	b	β	**p(*b*)**
Overall Model		0.30	0.38	0.339			
Test Accommodation	17[a]				0.09	0.10	0.553
Disability Classification	36[b]				0.17	0.20	0.116
Grade Level/s (Elementary)	25				0.00	0.00	0.982
Grade Level/s (Middle school)	28				-0.07	-0.08	0.720
Publication Year	58				0.00	0.04	0.796
Publication Type (Journal)	37				-0.02	-0.02	0.926
Publication Type (Dissertation)	14				-0.18	-0.19	0.367
Test Content (Math)	35				-0.60	-0.72	0.001
Test Content (Reading/LA)	17				-0.35	-0.39	0.061
Test Format	47[c]				-0.15	-0.14	0.347

[a] total for timing accommodation
[b] total for students with learning disabilities
[c] total for multiple-choice format

Effect of timing and presentation accommodations on test scores for students with learning disabilities.

Effect size, proportion of variance explained (R^2), residual variance test significance (Q), individual variable beta weights and their associated p-values, for each moderator variable set are presented in Table 28**.** Only studies containing data for timing or presentation test accommodations are included in the analysis. Only the test characteristic variable set, R^2 =0.30, explained a sizeable portion of the variability for mean increase in test score. Researcher-manipulated variables, R^2 =0.02, population description, R^2 =0.06, and dissemination, R^2 =0.08, variable sets provide little explanation for the heterogeneity seen. A single statistically significant beta weight (b = -0.65, β = -0.77, p < 0.001) for math content was found. Math content was negatively related to effect size, change in test score, for students with learning disabilities.

Table 28: *Random-effects Model for Students with Learning Disabilities - Timing & Presentation Accommodation Data Only*[a]

Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Each Variable Set

Variable	k	Mean ES	R^2	p ($Q_{residual}$)	b	β	p(b)
Researcher manipulated variable		0.41	0.02	0.247			
Test Accommodation	13[b]				0.11	0.12	0.430
Population Description		0.41	0.06	0.211			
Grade Level/s (Elementary)	14				-0.22	-0.26	0.254
Grade Level/s (Middle school)	17				-0.31	-0.36	0.108
Dissemination		0.41	0.08	0.185			
Publication Year	36				0.03	0.24	0.124
Publication Type (Journal)	23				0.11	0.13	0.635
Publication Type (Dissertation)	10				0.01	0.01	0.981
Test Characteristics		0.40	0.35	0.207			
Test Content (Math)	19				-0.65	-0.77	0.000
Test Content (Reading/LA)	12				-0.19	-0.21	0.309
Test Format	29[c]				-0.20	-0.19	0.168

[a] for this subset of data test accommodation category data and specific test accommodation data are the same

[b] total for timing accommodation

[c] total for multiple-choice format

The overall meta-regression model is presented in Table 29. Entering all moderator variables into the meta-regression model, R^2=0.48, explained more of the variance for effect size, mean increase in test score, than any of the variable sets. Additionally, only one of the moderator variables, math content, was statistically significant. Entering the single statistically significant moderator variable, math content, into the meta-regression equation reduced R^2 to 0.28 (see Appendix S). When the researcher-manipulated moderator variable, test accommodation, was not entered into the meta-regression model the R^2 remained the same, 0.48, as the overall model (see Appendix S). Neither timing nor presentation test accommodations aid in explaining improvement in test scores, effect size, for students with learning disabilities nor do they appear to remove construct-irrelevant variance.

Table 29: *Random-effects Model for Students with Learning Disabilities - Timing & Presentation Accommodation Data Only*[a]

Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for the Overall Model

Variable	**k**	**Mean ES**	$\mathbf{R^2}$	p ($\mathbf{Q_{residual}}$)	b	β	**p(*b*)**
Overall Model		0.40	0.48	0.182			
Test Accommodation	13[b]				-0.05	-0.06	0.780
Grade Level/s (Elementary)	14				-0.04	-0.04	0.838
Grade Level/s (Middle school)	17				-0.05	-0.06	0.813
Publication Year	36				0.02	0.12	0.457
Publication Type (Journal)	23				0.03	0.04	0.887
Publication Type (Dissertation)	10				-0.28	-0.30	0.318
Test Content (Math)	19				-0.82	-0.98	0.000
Test Content (Reading/LA)	12				-0.30	-0.35	0.153
Test Format	29[c]				-0.23	-0.22	0.204

[a] for this subset of data test accommodation category data and specific test accommodation data are the same

[b] total for timing accommodation

[c] total for multiple-choice format

Test accommodations, construct irrelevance, and effect size.

What was consistently demonstrated across all meta-regression analyses was that a substantial proportion of the heterogeneity could be explained by test characteristics and, in some instances, descriptive characteristics of the population under investigation. However, very little of the heterogeneity in the meta-regression model was explained by test accommodations. Neither timing nor the presentation test accommodations appeared to be uniquely effective in removing construct-irrelevant variance for the students with disabilities. Construct-irrelevant variance was better explained by content of the assessment, specifically math content being negatively related to effect size, and specific disability group, specifically students with learning disabilities when compared to the more general group of students with disabilities. These findings also held when a subset of this population, students with learning disabilities, was examined.

Chapter Four

Discussion

With a growing number of students identified as requiring special education services, and the increased use of high-stakes and large-scale assessments to monitor academic progress at the student, school, district, and state levels, issues regarding the utility of the these types of assessments abound. One of the most frequently recommended methods to minimize construct-irrelevant variance and difficulty on these assessments is use of test accommodations (Kieffer, Lesaux, Rivera, & Francis, 2009). Appropriate accommodations provide direct, or indirect, support to minimize factors irrelevant to the content, or construct, being assessed and allow students with disabilities the opportunity to demonstrate their knowledge and skills with minimal impedance.

The present study was designed to provide a quantitative synthesis of experimental and quasi-experimental research on the efficacy and validity of test accommodations for students with disabilities participating in high-stakes assessment programs. Previous analyses in this area tended to be narrative syntheses of the research literature and, as such, are considered more subjective than the quantitative synthesis used. Employing meta-analysis, the study was designed to build on the work of Chui and Pearson (1999), as well as narrative syntheses of the research (Bolt & Thurlow, 2004; Calahan Laitusis, 2004; Cormier et al., 2010; Elliott, McKevitt, & Kettler, 2002; Johnstone et al., 2006; Sireci et al., 2003; Thompson et al. , 2002; Thurlow & Bolt, 2001;

Tindal & Fuchs, 2000; Zenisky & Sireci, 2007; Zuriff, 2000). Additionally, meta-regression, previously not attempted in this area of research, was employed to further our understanding of the heterogeneity in effect sizes seen when evaluating the effect of assessment accommodations for students with disabilities.

Thirty-four studies, from mid-1999 through mid-2011, investigating testing accommodations for students with disabilities comprised the dataset used in the present analysis. Separate effect sizes were calculated for students with disabilities and their typically developing peers. The 34 separate studies (34 for students with disabilities, 31 for typically developing students) were analyzed using each study as the unit of analysis, aggregating results across separate subunits. With 12 studies providing more than one unit of analysis, or study, 119 separate effect sizes (62 for students with disabilities, 57 for typically developing students) were coded and analyzed using substudy as the unit of analysis.

Summary of findings.

The current study investigated three separate, linked research hypotheses. The first two hypotheses were investigated using quantitative meta-analytic techniques, while the final research hypothesis was analyzed using meta-regression.

Meta-analysis.

The first meta-analytic research hypothesis focused on differences between students with disabilities, typically developing students, via effect sizes for each group. Effect size statistics, based on Hedges' g, were used to investigate differences between these two groups. The second research hypothesis focused on which types of test accommodations were efficacious. Results for presentation and timing/scheduling assessment accommodations were presented separately. Results for setting and response assessment accommodations were not included as the number of effect sizes for each of these accommodation categories was very small, thus making inferences would be tenuous at best. Again, effect size statistics, based on Hedges' g, were used to explore this hypothesis.

Differential boost.

The first research hypothesis focused on whether or not there was empirical support for delivering test accommodations to students with disabilities as opposed to typically developing peers, and was explored as a question of differential boost (Fuchs & Fuchs, 1999). As it was felt that some typically developing students might benefit from assessment accommodations, though not to the same extent as their disabled peers, differential boost was selected to frame answers to the first research hypothesis. While differential boost and the interaction hypothesis (Sireci et al., 2005) propose similar assumptions with respect to students with disabilities, that is, that students with disabilities will exhibit test score gains in accommodated versus non-accommodated conditions, they diverge on their assumptions with respect to typically developing students. The interaction hypothesis, in its strictest interpretation, posits that typically developing students will not benefit from assessment accommodations whereas differential boost postulates differences between the two groups but does not dismiss the possibility that these students will make gains in an accommodated condition, albeit substantially less than students with disabilities.

For the study level analysis the Q-test for the distribution of observed effect sizes was statistically significant both for students with disabilities and their typically developing peers. Similar results were obtained when substudy was used as the unit of analysis. Results of Q-test indicated that there were differences in effects sizes for students with disabilities and peers with typical development that were not readily accounted for by sampling variation. Thus, random-effects models were used.

It should be noted that overall mean effect size across groups, students with disabilities and typically developing students, was not used. Focus was on dispersion of effect sizes within each group and not overall effect size. Hence, results reported are for each group.

With study as the unit of analysis, the mean effect size for students with disabilities was 0.36 (k = 34, $p < 0.001$) and 0.19 (k = 31, $p < 0.001$) for typically developing students. When using substudy as the unit of analysis similar results were found with the mean effect size for students with disabilities being 0.30 (k = 62, $p < 0.001$) and 0.17 (k = 57, $p < 0.001$) for typically developing students. For the study level analysis, these effect sizes represented 5,740 students with disabilities with 8,877 typically developing peers totaling 14,617 participants, while representing 5,338 students with disabilities and 8,491 typically developing peers for a total of 13,829 participants for the substudy level analysis. The differences between the numbers of participants for the two analyses reflect differences in how participants were counted when studies were aggregated at the study level versus disaggregate at the substudy level. Specifically, demographics presented for study level incorporated all research participants included in effect size calculations while demographics for substudy only incorporated research participants once, even in instances where these participants would have taken more than one version of a test as the number of participants at the substudy level was broken out by specific information for each substudy. In both analyses the mean effect size for students with disabilities, albeit small (Cohen, 1992) and statistically significant, was one-third to almost one-half larger than that for their typically developing peers. The mean effect size for typically developing peers, although statistically significant, was considered very, or trivially, small.

Results from these analyses lend support to the differential boost hypotheses, whereby students with disabilities are positively impacted by test accommodations while their typically developing peers gain little from test accommodations.

Presentation test accommodations.

The second research hypothesis focused on the efficacy of specific assessment accommodations for students with disabilities. Analyses for presentation accommodations were conducted across the entire group of students with disabilities, as well as being broken down by type of disability, learning disability and students requiring special education services.

The overall mean effect size, using a random-effects model, for the presentation assessment accommodation ($\overline{ES}$ = 0.22, k = 41, $p < 0.001$) was small and statistically significant. Presentation accommodations were categorized as read-aloud, segmented text, and simplified language specific accommodation categories. The overall mean effect size for the specific category of read-aloud accommodation was 0.24 (k = 39, $p < 0.001$)

and although small was statistically significant. As there was only one effect size, each, for segmented text and simplified language assessment accommodations, these results were reported but not examined.

While overall mean effects for students with disabilities provided some insight into the efficacy of presentation accommodations, one further analysis was conducted to see if specific category of disability would provide additional insight into the efficacy of this specific test accommodation. As most studies provided information on type of disability under investigation, these data were available and were used to create two categories, students with learning disabilities and students requiring special education services. Once disaggregated we saw that the effect for students with learning disabilities intensified ($\overline{ES}$ = 0.36, k = 23, $p < 0.001$) while it was negligible ($\overline{ES}$ = 0.04, k = 16, $p = 0.48$) for students requiring special education services.

The findings indicate that the use of presentation assessment accommodations had a statistically significant, albeit small, impact on the performances of students with disabilities. This effect intensified for students with learning disabilities when students with learning disabilities and students requiring special education services were studied separately. Again, it must be noted that students receiving special education services also include students with learning disabilities so there is some overlap. Although limited by this overlap, we do see that the more specific we are about type of disability, the better able we appear to be in targeting appropriate accommodations to positively (e.g., statistically significant) impact students with disabilities.

Timing/scheduling test accommodations.

With the focus on the efficacy of assessment accommodations by specific accommodation category for the second research hypothesis, results for timing/scheduling test accommodations for students with disabilities were presented separately. As with the results for presentation assessment accommodations, analyses for timing/scheduling accommodations were conducted across the entire group of students with disabilities, as well as being broken out by type of disability, learning disability or students requiring special education services.

The overall effect size for the timing/scheduling assessment accommodation ($\overline{ES}$ = 0.47, k = 17, $p < 0.001$) was small, bordering on medium, and statistically significant. Again, one further analysis was conducted to see if specific disability category might provide added insight into the efficacy of assessment accommodations. While the overall mean effect for timing/scheduling accommodations ($\overline{ES}$ = 0.48, k = 13, $p < 0.001$) for students with learning disabilities was statistically significant, bordering on being considered a medium effect, disaggregation did not intensify the results for this group. As there were only three effect size estimates for students requiring special education services these results were reported but not examined.

The findings indicate that the use of timing/scheduling assessment accommodations had a small to medium, statistically significant impact on the performances of students with disabilities. This effect remained consistent for students with learning disabilities.

Meta-regression.

The third research hypothesis focused on removal of construct-irrelevant variance from the test scores of students with disabilities. In addition, effects on test score for students with disabilities due to use of assessment accommodations was explored. Potential moderating variables were entered into a meta-regression analysis in an effort to ascertain which, if any, variables aided in removing construct-irrelevant variance as well as helping explain test score improvement for students with disabilities. With statistically a significant overall effect size estimate ($\overline{ES} = 0.30$, k = 62, $p < 0.001$) and a statistically significant Q-value ($Q_{(61)} = 782.27$, $p < 0.001$) for students with disabilities, heterogeneity beyond sampling variation was present. Meta-regression analyses, using the random-effects model were performed for (i) students with disabilities across all data collected, (ii) students with disabilities across presentation and timing/scheduling data only, and (iii) students with learning disabilities across presentation and timing/scheduling data only. The dependent variable was represented by Hedges' g effect size estimates in the meta-regression analyses.

Effect of test accommodation on test scores for students with disabilities.

When sets of moderator variables were analyzed separately, population description and test characteristic variable sets were found to explain the greatest amounts of variability for mean increase in test score; R^2=0.22 and R^2 =0.35, respectively while researcher-manipulated variables and dissemination explained relatively little variance; R^2 =0.07 and R^2 =0.01 respectively. Additionally, beta weights for population description and test characteristic variables sets were statistically significant. With

$Q_{(residual)}$ values that were not statistically significant, we can conclude that the moderators were related to effect size and aid in explaining the heterogeneity seen and that population description and test characteristic moderator variables were related to the effect size.

More of the unique variability for mean increase in test score, R^2=0.42, was explained by the overall model than by any single moderator variable set. Entering only statistically significant moderator variables into the regression equation, math and reading content, decreased variability accounted for, R^2 = 0.28. When all but test accommodations, the researcher-controlled moderator variables, were entered into the meta-regression equation, the R^2 was also 0.42. Consequently, test accommodations, timing versus all other accommodations and presentation versus all other accommodations, did not appear to provide additional information to explain the heterogeneity in the model.

Effect of timing and presentation accommodations on test scores for students with disabilities.

As was seen with in the meta-regression analysis including all the research data, the greatest amount of variability in mean increase in test score for separate variable sets was explained by population description, R^2=0.17, and test characteristic, R^2 =0.32. The researcher-manipulated variables, R^2 =0.07, and dissemination, R^2 =0.02 variable sets explained little of the heterogeneity seen. Both population description and test characteristic variable sets had statistically significant beta weights, albeit only disability classification (b = 0.24, β= 0.29, p = 0.017), math content (b = -0.65, β= -0.78, p = <

0.001), and test format (b = -0.26, β= -0.24, p = 0.034) beta weights for these variable sets were statistically significant. Disability classification was positively related to change in test score while math content and test format were negatively related to change in test score.

Again, the overall model explained more of the heterogeneity for mean increase in test score, R^2=0.38, than any of the moderator variable sets. When the only statistically significant moderator variable, math content, was entered into the meta-regression equation the R^2 fell to 0.19. However, the R^2 increased to 0.42 when the test accommodation moderator variable was not entered into the meta-regression model.

As previously, we see that neither timing nor presentation test accommodations effectively removed construct-irrelevant variance for students with disabilities, nor did they aid in explaining any improvement from non-accommodated to accommodated condition seen.

Effect of timing and presentation accommodations on test scores for students with learning disabilities.

The greatest portion of variability in effect size was explained by a single variable set, test characteristic (R^2 =0.30). Variable sets for researcher-manipulated variables, R^2 =0.02, population description, R^2 =0.06, and dissemination, R^2 =0.08, provide little explanation for the heterogeneity seen. As seen previously, once the data were reduced to data for presentation and timing/scheduling test accommodations, the only single statistically significant beta weight (b = -0.65, β= -0.77, p < 0.001) was for math content.

When entering all moderator variables into the meta-regression model, the overall model, R^2=0.48, explained more of the variance in effect size than any variable set. Additionally, once type of disability was accounted for, there was an increase in explained variance, R^2=0.42 for all data across all students with disabilities and R^2=0.38 for presentation and timing/scheduling test accommodation data across all students with disabilities.

When only math content, the single statistically significant moderator variable, was entered into the meta-regression model, the meta-regression equation reduced R^2 to 0.28. It did not make any noticeable difference whether the assessment accommodation variable was entered into the meta-regression equation (R^2=0.48) or not (R^2=0.48).

To reiterate, neither presentation nor timing/scheduling assessment accommodations aid in explaining effect size for students with learning disabilities, nor do they appear to remove construct-irrelevant variance.

Across the three separate sets of meta-regression analyses, there were statistically significant results for population description and test characteristic variable sets, specifically math content. Test format, specifically multiple-choice, was also found to be statistically significant for the first two sets of regression analyses. As well, disability classification was statistically significant in the two first sets of regression analyses. Disability classification was not included in the third analyses as it was used to structure the model, whereby the variance for the moderator variables was examined for students with learning disabilities. Thus, the findings from the meta-regression analyses demonstrate little evidence of moderating effects of any of the reported characteristics of studies, research participants, or assessments. Rather, the findings from the meta-regressions suggest that systematic differences across studies may have been, in part, due to differences in test content. However, when considering test content it becomes apparent that the test content is intimately intertwined with the type of assessment accommodation used. For example, for students with disabilities most studies of presentation; i.e., read-aloud assessment accommodations were for math assessments, 30:39 (77%) and most studies of timing/scheduling; i.e., extended-time, were for reading and language arts, 9:17 (53%). Similar outcomes were seen for students with learning disabilities; i.e., read-aloud assessment accommodations were for math assessments, 16:23 (70%) and most studies of timing/scheduling; i.e., extended-time, were for reading and language arts, 8:13 (62%).

Relation of results of this study to research in the field.

Educators and policy-makers require robust evidence regarding the effectiveness of testing accommodations for students with disabilities to make valid accommodation choices for these students. While there has been much primary research and qualitative research syntheses in this area, there has not been a quantitative synthesis, across the entire set of assessment accommodations for all students with disabilities since Chiu and Pearson (1999). In the ensuing years, there has been enough primary research in the field of assessment accommodations for students with disabilities to make meta-analysis useful. Zenisky and Sireci (2007) called for “[the] completion of more well-constructed meta-analyses of specific accommodations [as] one strategy that researchers should consider” (p. 17). Although meta-analysis does not overcome all of the pitfalls seen in the existing primary research in this area, using meta-analysis to aggregate and quantitatively analyze existing research has the potential to provide a more rigorous examination of the data collected to date. Given the increasing importance of large-scale assessments and the increasingly high stakes attached to assessment results for states, districts, schools and students, the current synthesis of research work has valuable implications for researchers, policy makers, and educators. Additionally, there has been a call to extend information regarding the efficacy of assessment accommodations through examination of potential moderating effects (Kieffer et al., 2009). With the addition of meta-regression to provide a statistical means to delve deeper into possible explanations for excess variance and extend effect size findings provided through a meta-analysis of existing research studies, the current research helps answer that call.

Most qualitative syntheses of the primary research into assessment accommodations for students with disabilities point to mixed results for use of test accommodations (Bolt & Thurlow, 2004; Cormier et al., 2010; Johnstone et al., 2006; Thompson et al., 2002; Thurlow & Bolt, 2001; Zenisky & Sireci, 2007; Zuriff, 2000). Conversely, and in keeping with the results of the current research work, Sireci et al. (2003), Sireci et al., (2005), and Thurlow (2007) found extended time improved the performance of students with disabilities more than for typically developing peers. As well, and again in keeping with the results of the current study, Sireci et al. (2005) found that read-aloud, oral, assessment accommodations on mathematics tests generally showed improved performance for some students with disabilities. Additionally, while this author agrees with Fuchs, Fuchs, and Capizzi (2005), certain accommodations have been shown to benefit some students with learning disabilities, and no single accommodation has been shown to benefit all students with learning disabilities: quantitative analyses of the primary research indicated that students with learning disabilities benefit from presentation and timing/scheduling test accommodations more often than not.

The current research supports the notion of differential boost, as did the original findings of Chiu and Pearson (1999). Students with disabilities ($\overline{ES}$ = 0.30 for the current study, $\overline{ES}$ = 0.16 for Chiu & Pearson, 1999) perform differentially better than their typically developing peers ($\overline{ES}$ = 0.17 for the current study, $\overline{ES}$ = 0.06 for Chiu & Pearson, 1999). It is expected that Chiu and Pearson's inclusion of English language learners, together with the addition of more studies examining simplified language, explain the lower mean effect size found for students with disabilities in their study. This explanation is consistent with findings in the field. For example, Pennock-Roman and Rivera (2011) posted overall effect sizes of 0.053 (plain English, restricted time) and 0.108 (plain English, no time constraints) for English language learners with overall effect sizes of -0.008 (plain English, restricted time) and 0.064 (plain English, no time constraints) for their English-speaking peers.

It must be noted that although those involved in examining assessment accommodations for students with disabilities have identified four areas of accommodation (presentation, response, setting, and timing/scheduling) data were only available for two types of assessment accommodations (presentation and timing/scheduling) for the present meta-analysis. While the current research examining the aggregated work on presentation and timing/scheduling accommodations, specifically read-aloud and extended-time accommodations, does point to the ability of these accommodations to ‘level the playing field’ for students with disabilities, the present work is unable to address response and setting accommodations without location of additional primary analyses in these areas.

Issues in Meta-analysis

As the current study is a quantitative synthesis of the research literature, limitations are not bounded in the same manner as they are with primary research. Limitations are manifold and include issues with the variables under investigation; e.g., test accommodations, those limitations found with the primary studies, issues with coding the primary study information, and issues with the statistical techniques employed; i.e., meta-analysis and meta-regression.

Coding primary study information added layers of complexity to the present meta-analysis and meta-regression analyses that were not originally expected. The unit of analysis was not clear-cut unless there was only one set of data for students with disabilities and/or their typically developing peers. For example, while deciding on unit of analysis for a primary study that contained data for two different grade levels was simple, when a primary study contained subtests or different content areas demarcation of the unit of analysis became less obvious. As well, the research designs used in the primary studies have grown much more complex since Chiu and Pearson’s (1999) meta-analysis in this area. The addition of the “maximum potential thesis” (Zuriff, 2000), “differential boost” (Fuchs & Fuchs, 1999), and interaction hypothesis (Sireci et al., 2003; Sireci et al., 2005) provided new ways of thinking about research into test accommodations and how best to examine test accommodations. Designs moved from examining test score boost for students with disabilities to differences in test score boost between students with disabilities and typically developing peers; i.e., differential boost. Further, many of the studies located, approximately 10%, did not include adequate information to be included in the present meta-analysis. In some instances, correlation coefficients for primary studies using repeated measures designs could not be located or estimated. As well, studies were missing information on the number of students assessed, results for both the non-accommodated and accommodated conditions, standard deviations for the conditions, and/or t-test and p-value information in some instances. Additionally, the various categories of students with disabilities were not always well defined. Frequently, classification relied on participant's use of an individualized education plan (IEP) but did not include information on the classification contained in the IEP; e.g., primary study listed participants received special education services with no further breakdown. Although it had been hoped that more specific disability information would have been provided so that assessment accommodation could be correlated with specific types of disabilities, such as cognitive impairment or seriously emotionally disturbance, only one such study was located.

With the increase in the complexity of research design, issues with determining appropriate effect size calculation arose which ultimately led to the need to make decisions regarding the aggregation of effect sizes based on different calculations. While data based on different research designs, hence different effect size calculations, may be aggregated or disaggregated it was felt that aggregating the data based on substantive lines (Borenstein et al., 2009) was the most appropriate method.

It should be noted that correction for Type I error, whereby finding a statistically significant effect size when none was present, was not applied to the calculated effect sizes.

Meta-analytic research has a number of limitations. Lipsey and Wilson (2001) present a common weakness as "the amount of effort and expertise it takes" (p. 7), lack of sensitivity to "important issues" (p. 7) due to the structured, mechanical processes used, the mix of studies which can be included in a meta-analysis, and "mixing of study findings of different methodological quality in the same meta-analysis" (p. 9). Further limitations for meta-analysis, noted by Borenstein et al. (2009), include the file drawer problem, whereby the sample of studies selected was biased and important studies were ignored.

While the first limitation says much about those attempting to conduct research using meta-analysis and cannot be remedied without experience, this researcher reviewed each step of the meta-analytic process with a methodology expert and spent much time reviewing pertinent literature on meta-analysis, substantive (e.g., what constitutes a good meta-analysis), methodological (e.g., how to calculate an appropriate effect size), and empirical (e.g., Gregg & Nelson's (2010) meta-analysis on test accommodations for transitioning adolescents with learning disabilities). In an attempt to curtail the effect of the second limitation, efforts were made to add a descriptive component to the summary of the findings by providing as much context as possible. Additionally, information for all effect size estimates was included in the text of the present research. In a study that relies on primary research where attempts at replication between studies can be tenuous, some mixing of apples and oranges is expected. However, the third limitation, coding information such as type of analysis (i.e., boost, differential boost), statistic used (i.e., t-test, ANOVA, ANCOVA, etc.), and use of the coded information in the analysis was hoped to help curb this issue. As well, the research hypotheses were designed to aggregate studies with more similar components through multiple meta-analyses. This was evidenced by starting with a comparison between students with and without disabilities, moving to an examination of accommodation categories for students with disabilities, then to an examination of specific accommodations for students with disabilities, culminating with an analysis of specific accommodations for students with learning disabilities as compared to students receiving special education services. The mixing of studies of different methodological quality is difficult to address and appears to be the most contentious issue among meta-analysts. However, research by Ahn and Becker (2011), through a Monte Carlo study, recommend against the use of quality weights in meta-analysis as their addition does not significantly change results found. It should be noted that most studies used in the meta-analyses, 29:34 or 85%, went through some type of peer-review process as they were published in peer-reviewed journals or were dissertations that would have been reviewed by a dissertation committee. Limitations due to selecting a biased sample, based on the notion that only studies with high treatment effects are published, did not appear to be problem for the present research. Studies with and without treatment effects; i.e., positive effects for accommodated conditions, were found in the primary research literature, so much so that syntheses of the research literature in the area pointed to the mixed results from research on testing accommodations. To address the final limitation the present study cast as wide a net as possible to find research in the area conducted between mid-1999 through mid-2011. Further, all studies that could possibly be coded were coded. Studies were not dropped from the analysis unless necessary statistical data, data regarding test accommodation and/or data regarding the participants, could not be found in the study, by locating additional work on the primary research (e.g., a report and a journal article reporting on the same primary research), or by contacting the primary researchers involved. Additionally, to allow for further examination of included studies by other researchers, information for studies that were dropped from the analysis or could not be located have been included in the appendices (see Appendix H for a list of excluded studies and Appendix I for a list of irretrievable studies, respectively).

A further limitation for meta-analysis, specifically as a statistical technique, was noted based on the type of primary research collected. At present there do not appear to be any methods to compute an effect size for mean difference for multiple group comparisons. For example, primary research that contains comparisons of students with learning disabilities, students with behavioral issues, and students with speech/language disabilities is problematic, as an effect size based on the aggregate comparison does not appear to be possible. While an effect size can be computed for any two of these three groups, trying to analyze data collected for all three groups in a single analysis is not yet possible. Comparing multiple assessment accommodations in a single analysis for a primary study produces the same problem. The present study did not incorporate results from the single study found that had multi-group data for disability type. However, data from primary research simultaneously assessing the impact of multiple test accommodations were used. When faced with data for more than one test accommodation in a single study data from the most commonly studied test accommodation and data for the nonaccommodated condition were selected for use in the meta-analyses. This did limit the breadth of the types of accommodations that were analyzed, with lesser-studied accommodations being discarded for some studies; 3:34 or 9%.

Issues in Meta-regression

Shortcomings of meta-regression methods cited by Higgins and Thompson (2004) are "substantially inflated false-positive rates when heterogeneity is present, when there are few studies, and when there are many covariates[; i.e., moderators]" (p. 1663), and "… fixed effect meta-regression [being] likely to produce seriously misleading results in the presence of heterogeneity" (p. 1663). To counter these shortcomings only nine to 11 moderators were examined for 36 to 62 substudies and the random-effects model was employed. As the number of included 'studies;' i.e., substudies, was considered sizeable it was felt that the risk of identifying spurious associations was decreased (Higgins & Thompson, 2004).

While much data were collected, the data used for the meta-regression were considered *lumpy*, particularly for indicator variables; i.e., the data, being ordinal or categorical, were difficult to structure. Much of the critical information, necessary for inclusion in a meta-regression, was missing from located studies.

An additional shortcoming of meta-regression relates to the associations derived from meta-regressions: these associations must be thought of as observational. Causal relationships drawn from randomized comparisons provide relatively strong interpretations of data while those drawn from meta-regression cannot be viewed in the same light (Thompson & Higgins, 2002). Specifically, averages of student characteristics for each study were used as moderators in the meta-regression and cannot be thought of in the same way as they were in the primary analysis. Although moderators were pre-specified to avoid data dredging, the associations for the current research must still be thought of as observational. Moreover, while primary studies included in the meta-analyses benefited from using randomization in the original experimental or quasi-experimental design, meta-regression analyses performed no longer benefit from this randomization. As well, variables, which differ between studies in a meta-analysis, may be highly correlated and produce bias by confounding.

Additionally, while meta-regression can be used to explain heterogeneity of treatment effects between studies through use of carefully selected moderators, in this case differences between unaccommodated and accommodated test scores, the presence of 'residual' heterogeneity must be recognized, as it is not realistic to presume all of the heterogeneity has been explained (Thompson & Higgins, 2002). In an effort to account for as much of the heterogeneity as possible the random-effects model was used for the current research as the random-effects analysis provides wider confidence intervals than the fixed-effects model.

At this point in time meta-regression is not a very flexible technique. Researchers are limited by the types of regression analyses available for this technique and cannot easily conduct standard or hierarchical regression analyses, nor are they able to easily and effectively leverage results from structural modeling.

Limitations

One potential inadequacy with assessment accommodations can be tracked to the inability of some students to effectively use the accommodation due to constraints of their specific disability. For example, Burch (2002) points to the limited effectiveness of extended time on reading tests for poor readers. If research participants are unable to decode the words on the test administered, no amount of time will help demonstrate their ability to answer comprehension questions on the test. As well, some assessment accommodations may be of limited potency. For instance, extending the time limits on a one-hour assessment by 10% or less; i.e., five minutes, may not have the potency to induce a treatment effect for the accommodated condition whereas an increase of five minutes on a test of one-half an hour; i.e., 17%, may be much more potent.

The sheer numbers of different types of assessment accommodations make research on testing accommodations difficult. Cahalan-Laitusis (2004) found difficulties pursuing research into assessment accommodations for tests of writing due to several factors, including the multiple types of accommodations being employed by test users; e.g., states and school districts.

The most pressing concern for primary study researchers addressing the efficacy of test accommodations for students with disabilities was the limited number of students that were available to take part in the research. With 13% of the population requiring special education services, ‘small n’ studies become a common issue. Another pressing issue was related to the heterogeneity of the group of students under study. Primary researchers noted that students with disabilities are a heterogeneous group and often, when providing assessment accommodations in research situations, this variation has not been taken into account. While some studies allowed for this heterogeneity by providing research participants with teacher-recommended test accommodations in addition to the accommodation under study, this made disentangling the effect of the assessment accommodation and the provision of other, dissimilar teacher-recommended accommodations tenuous. It also precluded inclusion of studies with this type of design from the current meta-analytic research. It must also be noted that the same could be said for students with learning disabilities, as they are also a heterogeneous mix of individuals. In addition, many of the parametric research techniques; e.g., ANOVA, that provide useful data for meta-analyses cannot be utilized if the number of participants in the primary research is particularly small. Further, wide varieties of limitations were listed for the primary studies included in the current analyses. These limitations ranged from non-representativeness of the sample; e.g., too many white participants, and lack of homogeneity of the group of students with disabilities under investigation to self-pacing for read-aloud accommodations and ceiling effects for extended-time accommodations where students did not require additional time to complete the assessment used in the investigation.

Conclusion

Current trends in use of assessment accommodation relate to the enactment of NCLB (2001) and the need to test *all* students, including those with disabilities. With both an increase in testing and an increase in the numbers of students requiring special education services, the proliferation and use of testing accommodations has burgeoned. While

> some general accommodation decision-making and implementation guidelines can be obtained from a synthesis of currently available research, more empirical study is warranted to further investigate the effects of testing accommodations for students with disabilities (Bolt & Thurlow, 2005, p. 151).

To this end, many primary studies on assessment accommodations have been conducted. However, there have been only two limited, quantitative syntheses conducted since 1999 (Chui & Pearson). At the same time, it should be noted that there is no 'one-size-fits-all' test accommodation (Abedi, Hofestetter, & Lord, 2004) for students with disabilities. Rather there is a range of test accommodations that may aid in allowing students with disabilities demonstrate what they know and can do. Additionally, as noted by Gregg and Nelson (2010), "[a]ccommodations are not the source of differential performance… they simply mediate learning" (p. 233) and "do not supply the knowledge necessary to pass tests" (p. 231). To both prevent test accommodations from being a source of differential performance between students with and typically developing students and allow students with disabilities access to tests, test accommodations must not change the construct, or content, being tested. The present quantitative synthesis of the research tried to both address the gap in the body of research and examine the efficacy of the variety of test accommodations commonly used.

Meta-analyses examining the differences between scores for students with disabilities as compared to their typically developing peers provided evidence of differential boost. Results for students with disabilities showed, at best, small to moderate overall effects with these students not benefiting, as compared to their typically developing peers, as much as would be expected. As with Chiu and Pearson's meta-analysis (1999), we must proceed cautiously with the interpretation of these average effects as "…a wide variety of accommodations exist, the statuses of student are specific, and the implementations of accommodations vary in nature and quality" (p. 3). Additionally, it must be noted that the presence of this small to moderate effect for students with disabilities does not mean that all such students benefited from the accommodation, nor does it mean that those who benefited from the positive effect benefited equally. Further, some portion of the students with disabilities included may have a compromised neurological system such that no accommodation would allow them to demonstrate ability in the area being assessed. There was variability within the group of students with disabilities such that some benefited more than others did and the overall level of benefit was in the small to moderate range. Their typically developing counterparts did not receive the same level of benefit, although the benefit received was statistically significant. Additionally, other factors may also have helped account for this result. As some of the typically developing students that might have received special services did not, given the current special education classification requirements, this is to be expected. The potency of the various accommodations, some accommodations to more effectively 'level the playing field' than others, was considered to add to the variability in effect size estimates. With respect to typically developing students, these students, in some instances, also benefited from assessment accommodations. Some members in this group most likely have undiagnosed disabilities, thus also benefit from accommodations provided. As well, a recent study by Lewandowski et al. (2007), examining extended time for students with attention deficit disorder, found that removal of ceiling effects and allowing extended time benefited students developing typically, as they were able to accomplish more work than students with disabilities with the addition of extra time.

Demographics from the present quantitative synthesis of assessment accommodation research produced a number of findings that were expected. With national focus on large-scale assessment in core content areas; i.e., math and reading, it was not surprising to find that these two content areas were the most frequently studied. As well, students with learning disabilities were the most likely disability group to be included as research participants in primary studies of test accommodations.

Meta-analyses examining the efficacy of different categories of assessment accommodations did provide evidence for timing/scheduling and presentation assessment accommodations. Similar analyses examining specific test accommodations also provided evidence for the efficacy of extended-time and read-aloud accommodations. As was the case for differential boost, the evidence also showed small to moderate overall effects for these accommodations. Unfortunately, there were so few studies of setting and response accommodations even tentative conclusions regarding their efficacy were not possible. As was the case with differential boost, the heterogeneity within the different categories of accommodations may partially explain the lack of strong overall effects. For example, read-aloud accommodations have at least three distinct qualities, (i) administrator of the accommodation, (ii) content being read aloud, and (iii) time students were expected to wait between responding to questions. The administrator of the accommodations varied from classroom teacher to videotaped presentation, while the content being read aloud varied from test questions to the entire test, and with elapsed time varying from study to study it is likely that some participants became bored and did not attend to the task at hand. While grouping each of these specific test accommodations on a more granular level had been attempted, there were too few studies in each grouping to provide useful information in the meta-analyses. Narrowing the amount of heterogeneity by limiting the studies to those for students with learning disabilities did provide stronger evidence for use of extended-time and read-aloud accommodations. Further, matching accommodation to learning profile for students with disabilities; e.g., use of IEP- or teacher-recommended accommodations, might potentially increase the effect seen. However, few studies using this type of approach provide data; i.e., means and standard deviations for a group of students requiring the same, single accommodation matched to their learning profile, which can be used in a meta-analysis.

Meta-regression analyses conducted did point to two groups of moderator variables that may help explain some of the heterogeneity in the analyses, specifically population description and test characteristics. Population characteristics that were statistically significant, disability characteristics; i.e., students with learning disabilities as compared with all other students with disabilities, and grade level; i.e., middle school students with disabilities as compared to all other grade levels, accounted for some of the variability. However, when removing learning disability as a moderator variable; i.e., only examining students with learning disabilities, the population descriptor, grade level, was no longer significant. Note that this is not to say that the preponderance of studies in the analysis for students with learning disabilities was conducted with middle school students. Rather, this moderator did not provide sufficient information to account for the variance in the analyses once type of disability was controlled for. Test characteristics, specifically math content as compared to all other test content, also helped explain some of the heterogeneity found across the studies included in the analysis. However, the researcher-manipulated variable of test accommodation provided almost no explanation for the variability found. It is postulated that there is entanglement between test accommodation and test content. Test content may well be a proxy for the assessment accommodation used, particularly when the test accommodation was part of the test design, as was the case with the read-aloud accommodation for the studies used in the analyses.

While the results of the meta-analysis yielded small to moderate overall effects, and the meta-regression was only able to account for a portion of the variability of these effects, it is believed that this information provided some meaningful insight as small effects "may be highly meaningful for an intervention that requires few resources and imposes little on the participants" which "may be more meaningful for serious and fairly intractable problems " (Wilson, 2006, slide 3). Findings of differential boost, that is, allowing for the possibility that typically developing students may receive some (albeit less) benefit from test accommodations, is not tantamount to saying that there is no reduction in construct-irrelevant variance. It is postulated that reduction in construct-irrelevant variance proceeds along a continuum and, while evidence of differential boost does not account for a large reduction in construct-irrelevant variance, some portion of this variance is reduced.

Much has been said about the potential for leveling the playing field or closing the gap between students with and without disabilities by providing assessment accommodations to students with disabilities. This, taken at face value, only provides part of the picture when it comes to decreasing construct-irrelevant variance and, possibly, increasing fairness in assessment. It must be noted that 'leveling the playing field' may not be the same as 'closing the gap'. While we might be able to make the playing field more level for those with disabilities, this does not ensure that the gap between these students and their typically developing peers is, indeed, closed. Sireci et al. (2005) suggested that the goal of the accommodation must be considered when conducting research or discussing assessment accommodations. For example, if the goal of the accommodation (i.e., access to a graphing calculator) is to obtain a more precise measure of students' abilities, as opposed to 'leveling the playing field' or 'closing the gap', then the accommodation should be offered, even when it benefits both students with disabilities and typically developing students. Sireci et al. (2005), in reviewing the test accommodation literature for students with disabilities, argued that if the performance of both groups of students improves with an accommodation the assessment was, most likely, too restrictive in the first place. Whilst this is considered true for assessment, it takes focus away from an even more salient issue. We must continue to work on trying to close the gap between students with disabilities and their typically developing peers when it comes to curricular goals. Students with disabilities should not be limited to a ‘watered-down’ version of their typically developing peers’ curriculum. Inasmuch as possible, students with disabilities must be provided with an education that allows them to reach their full potential. Poor performance on large-scale assessment for students receiving special education services does not imply that they are incapable of mastering content contained on these assessments. Rather it may be providing a wake-up call to educators, inviting us to provide these students with targeted, explicit, and rigorous instruction to better prepare them to successfully negotiate the academic world (Kieffer et al., 2009).

Suggestions for Future Research

There have been over 20 years of research into accommodations for students with disabilities that include several syntheses of the literature (Bolt & Thurlow, 2004; Cormier et al., 2010; Johnstone et al., 2006; Sireci et al., 2003; Sireci et al., 2005; Thompson et al., 2002; Thurlow, 2007; Thurlow & Bolt, 2001; Zenisky & Sireci, 2007; Zuriff, 2000), a meta-analysis conducted in 1999 (Chui & Pearson, 1999), and two meta-analyses examining subsets of issues with respect to accommodations and students with disabilities (Elbaum, 2006; Gregg & Nelson, 2012). Nonetheless, the body of studies to sample from for a meta-analysis has remained insufficient and, at this point in time, does not provide information for lesser used assessment accommodations which were considered necessary to fully address the research hypotheses posed by this author. There are, literally, over 100 potential accommodations states, school districts, and schools can draw from to help level the playing field for students with disabilities, with the effectiveness of only a handful of different types of accommodations being addressed through empirical research. The two most frequently addressed accommodations, extended-time (k = 13) and read-aloud (k = 16), represented 85% of all the usable studies for the present quantitative syntheses. With this in mind, several avenues available for future research in the area of test accommodations; primary research, extensions to current research, and future directions; are presented.

The findings for the current work focused primarily on two categories of accommodations, timing/scheduling and presentation, represented by two specific types of accommodations, extended-time and read-aloud, for which there was robust evidence. However, there was little or no focus on other accommodations, specifically setting and response, as there were so few studies conducted for these assessment accommodations. Given the wide variety of assessment accommodations in use, future research should investigate other innovative and/or widely used methods for accommodating students with disabilities that have not yet been studied. Additionally, for a truly comprehensive meta-analysis in this area, indeed for a more well-rounded literature on assessment accommodations, more data on these accommodations are needed. For example, examining accommodations matched to learning profile for students with disabilities; e.g., use of IEP- or teacher-recommended accommodations, might potentially increase the effect seen. Students who might understand the concept but are not able to demonstrate their knowledge through use of an accommodation under study might better be able to display their understanding if the accommodation matched their learning profile. Crafting several studies using this type of approach has the potential to provide needed data; i.e., statistical data such as means and standard deviations for a group of students requiring the same, single accommodation matched to their learning profile, which could be used to better understand how a ‘well-matched’ and articulated test accommodation increases the efficacy of the test accommodation to remove construct-irrelevant variance in student test score. Additionally, possible future meta-analyses would better be able to use this data in additional summary quantitative research.

Pennock-Roman and Rivera (2011) noted that some

> [a]ccommodation conditions having nontrivial effect sizes that do not reach statistical significance may be worth examining in future research to verify whether the effect sizes are replicable and statistically significant with a larger sample size (p. 17).

Many such effect sizes were found for the present meta-analyses and data for these studies is provided in the text of this dissertation.

Primary researchers, even when restricted by the method of dissemination of their research, should endeavor to include certain statistics as these statistics aid in both understanding the magnitude of their findings and in quantitative meta-analytic research. Statistics that are often missing in published research include p-values and, for studies using repeated measures designs, correlations between test scores for non-accommodated and accommodated conditions for each group participating in the research study. It is also recommended that statistical values of non-significant results be included as imputation of means and standard deviations for studies missing these values is not possible at this juncture.

Clear operationalization of the assessment accommodation under investigation is of paramount importance. It is recommended that researchers use a common framework, such as the NCEO categories, to provide clear and precise information on the specific category of accommodation that was implemented. This information provides both readers and those wishing to perform secondary analyses on the assessment accommodation with a clear picture of exactly which accommodation was used and how it was implemented.

Ability to account for the potency of the accommodation needs to be taken into account. While disaggregating the types of accommodations and running separate analyses may aid in understanding differences between groups, it is believed that a more direct approach is preferable. Possibly, primary researchers and those involved in test accommodation research could develop a scale that would allow researchers to rank the ‘strength’ of the accommodation condition. While this not considered a trivial task for most types of test accommodations, it could prove infinitely useful when trying to understand trends in the research in this area. Researchers might start with a relatively simpler scale; for example, a scale for extended time, whereby scaled rankings could be given based on the percentage of extra time given. This proposal may not be appropriate for all types of assessment accommodations, as many do not have ‘degrees’ of implementation. Some test accommodations may only be an all or none proposition, such as dictation to a scribe.

“[I]t is also important to ensure that students are comfortable with the accommodation prior to receiving it on a test, and that they are receiving the accommodation during instruction” (Bolt & Thurlow, 2004, p. 149). When students with disabilities are not familiar with the accommodations used in the research being conducted, they are more likely to perform poorly. While the same might be said for students with typical development, these students appear to be less likely to be distracted by test accommodations they are unfamiliar with.

Providing complete information on key factors such as the assessment used, including reliability and validity information, and descriptive data collected such as test format would allow for the inclusion of more studies in meta-regression studies, as well as provide for a more granular look at possible trends in assessment accommodation research.

Several extensions to the current research work are recommended. While it was not in the purview of research hypotheses posed, examining the differential boost hypothesis using a further breakdown of students with disabilities into two groups, students with learning disabilities and students requiring special education services may provide a more potent effect for test accommodations for these students. When examining results for students with learning disabilities it was noted that these students did receive more benefit from extended-time and read-aloud accommodations than other groupings of students with disabilities.

During the data collection phase this researcher noted that several studies on computer adaptive testing (CAT) have been conducted over the past several years. The present research could be extended to include assessment accommodation research on CAT if a literature review on the comparability of CAT to paper and pencil assessments shows such assessments are, indeed, equivalent for this population. CAT could be coded as a separate accommodation, or divided into subgroupings based on the NCEO categories for assessment accommodations. Additions based on CAT studies of assessment accommodations might provide data on setting and response accommodations that are sorely lacking.

Becker (1988), cited in Morris and DeShon (2002), provides effect size calculations that allow for an independent group by repeated measures design. This could be used for the Engelhard et al. (2011) study included in the current analysis. The current meta-analysis selected the 'pretest' means collected for independent groups for the first year of the study as this was considered to more appropriately match the current research hypothesis. However, calculating effect sizes based on two years worth of data, which included pretest-posttest repeated measures information across independents groups; i.e., students with disabilities and typically developing students, might provide a more appropriate examination of the Engelhard et al. (2011) data. The effect sizes calculated could then be entered into the *Comprehensive Meta-Analysis V.2.2.050* program and compared with the results of the present study.

In their recent work, Pennock-Roman and Rivera (2012) proposed 'differential boost index,' which they used when conducting a meta-analysis on assessment accommodations for English language learners. Using this index with data on students with disabilities may provide a more comprehensive interpretation of the results as it aids in establishing whether "the improvement for the focal group [; i.e., students with disabilities] is relatively larger than for the reference group [; i.e., typically developing students]" (p. 3).

Future directions for quantitative meta-analytic methods recommended include the extension of programs such as *Comprehensive Meta-Analysis V.2.2.050* to include effect size computations that can be used with more complex research, such as that proposed by Becker (1988). To be able to effectively aggregate research with complex or multivariate designs meta-analytic methodology would need to be extended such that effect sizes for multivariate research designs, or designs that contain independent groups and repeated measures research designs within the same study, are available. Until this can be accomplished work by primary researchers included in meta-analyses need to employ simpler research designs or meta-analysts must select portions of the data from primary research that can be analyzed, leading to potential apples and oranges or garbage-in, garbage-out issues.

Finally, to reiterate Gregg and Nelson (2012)

> the impact of test accommodations on the validity of test scores should be investigated more thoroughly by future researchers. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) define validity as the 'degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test' (p. 184).

Policy implications.

Fairness is a primary consideration in all aspects of testing. Careful standardization of tests and administration conditions helps to ensure that all test takers are given a comparable opportunity to demonstrate what they know and how they can perform in the area being tested. Fairness implies that every test taker has the opportunity to prepare for the test and is informed about the general nature and content of the test, as appropriate to the purpose of the test. Fairness also extends to the accurate reporting of individual and group test results (Joint Committee on Testing Practices, 2004, p. 2).

NCLB assessment policies emphasis on assessment of all students coupled with the disaggregated group reporting for students with disabilities and other groups necessitate the use of valid measures of student performance. For students with disabilities it is often necessary to provide accommodations to ensure the measures used are accessible. At the same time, the content and/or underlying constructs must not be altered. Policy makers, researchers, and educators require access to the information on test accommodation research for students with disabilities and this information needs to be understandable and, where possible, implementable. To this end, it is necessary to review and summarize the research conducted on test accommodations for students with disabilities. This provides policy makers, researchers and educators with the opportunity "…to question whether changes in assessment and accommodations policies need to be made" (Cormier et al., 2010, p. 18).

Policy makers and educators need to learn more about test accommodations that are appropriate to use with students with disabilities in high-stakes testing situations. Connecting instructional accommodations and testing accommodations to allow students with disabilities the chance to become facile in using the accommodation prior to using it in a testing situation is vital. Those involved with testing students with disabilities have an obligation to ensure that students with disabilities have the ability to demonstrate what they can and cannot do. To this end, documentation of accommodation use, implementation, and efficacy must be tracked and disseminated to the educational community.

Research literature on test accommodations for students with disabilities has not yet reached ‘critical mass,’ whereby definitive statements regarding differential performance of students with disabilities and their typically developing peers or the efficacy of specific test accommodations can be made. Much of the research on test accommodations points to their limited effectiveness in improving the performance of students with disabilities, specifically students with learning disabilities (Kieffer et al., 2009). While results from the present quantitative research synthesis point to the benefits of extended-time and read-aloud accommodations with this population of students, being more effective for students with learning disabilities, there is still more work to be done.

It should be noted that, although this research points to efficacy for read-aloud and extended-time test accommodations, implementation of these accommodations on high-stakes assessments for students with disabilities or all students should not be seen as a panacea which allows these students to demonstrate their knowledge. Such implementation has the possibility of invalidating the high-stakes assessment as it could invalidly boost the test scores of the students receiving the accommodation. As well, the accommodation might not meet the needs of certain students receiving the accommodation and thus not allow these students to display their knowledge. Simply providing a test accommodation to a subgroup of students; i.e., students with disabilities or all students does not mean that we have effectively provided students with the tools to demonstrate what they have learned. Rather, it might hinder effective assessment and evaluation of student learning. Careful consideration and examination of test accommodations is necessary if we are to ensure that all students are provided with a means to display their abilities on high-stakes assessment. Simply providing a test accommodation is not enough.

Lesser studied accommodations; i.e., test accommodations that may be seen as more esoteric, are generally only used infrequently with students with disabilities, and usually only with students with complex, combined disabilities. Therefore, it is recommended that, for purposes of test validity, these results not be included in overall results for a school, school district, or state/province until such time that the test accommodation has been included in primary research analyses, or universal test design is in place, and is being used to alleviate issues with test scores arising from these students taking these types of accommodations. Depending upon the purpose for use of the results it would unfair to include test scores from students taking the assessment with lesser used accommodations, aggregated in the final results particularly at the class or school level where such results have the potential to exert a much stronger influence on aggregate information. This is particularly true for high-stakes assessment, whereby funding or other important decisions are made based on the results of the assessment. While this will not address the use of accommodations for this specific group of students with disabilities, it is considered the more prudent course of action. However, the results from these accommodations may still be validly used to inform teachers and administration about these students' abilities, that is, what they *can* and *cannot do* with respect to the content being examined.

While the efficacy of assessment accommodations for students with disabilities continues to be a major topic for educational researchers, we must not lose sight of larger issues. Firstly, when students are excluded from assessment situations we are less likely to target services to students with disabilities as their progress is inconsequential and *does not count* (Bolt & Thurlow, 2006). Secondly, there is still a gap between the performance of students with disabilities and their typically developing peers. This gap cannot be diminished by leveling the playing field for assessments using test accommodations, rather it must be addressed through improving instruction for these learners (Kieffer et al., 2009).

It is hoped that the present research will be used to extend research into the field of test accommodations for students with disabilities by providing guidance into selection of areas for further research. Additionally, it is hoped that, in some small part, the research conducted helps to inform those working in the area of universal test design with information regarding the potency of the test accommodations and their ability to impact assessment validity. Further, this research attempted to provide useful information for local, state, and federal testing agencies, as well as independent testing agencies, regarding the efficacy of various testing accommodations.

References

References prefaced with a numeral; e.g., (01), are included in the meta-analyses and meta-regression analyses.

Abedi, J., Hofestetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74,* 1 - 28.

(01) Abedi, J., Kao, J. C., Leon, S., Mastergeorge, A. M., Sullivan, L., Herman, J., & Pope, R. (2010). Accessibility of segmented reading comprehension passages for students with disabilities. *Applied Measurement in Education, 23*(2), 168 - 186.

Ahn, S., & Becker, B. J. (2011). Incorporating quality scores in meta-analysis. *Journal of Educational Behavioral Statistics, 36*(5), 555 - 585.

Algozzine, B. (1993). Including students with disabilities in systemic efforts to measure outcomes: Why ask why? In National Center on Educational Outcomes (Ed.). *Views on inclusion and testing accommodations for student with disabilities.* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/3a/03.pdf

Allington, R. L., & McGill-Frazen, A. (1992). Unintended effects of educational reform in New York. *Educational Policy, 6*(4), 397 - 414.

Amendments to the Individuals with Disabilities in Education Act, Public Law No. 108-446 (2004).

Americans with Disabilities Act of 1990, Pub. L. No. 101-336, (1991). Retrieved from http://www.ada.gov/archive/adastat91.htm

Authority: 20 U.S.C. 1412(a)(8), 1417(c) (March, 1999). Rules and Regulations. *Federal Register 64* (48), 12460.

Baker, H. (2008). *Effects of plain language revision on item difficulty, discrimination and DIF*. (Doctoral dissertation, University of Denver, 2008). Retrieved from Dissertations & Theses @ University of Denver. (AAT 3337045)

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research, 63*(1), 69 - 93.

Barkley, Russell A. (2006). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment (Third Edition).* New York, NY: Guilford Publications.

Bolt, S., Krentz, J., & Thurlow, M (2002). *Are we there yet? Accountability for the performance of students with disabilities (Technical report 33).* National Center on Educational Outcomes, University of Minnesota, MN. Retrieved from http://cehd.umn.edu/nceo/OnlinePubs.

Bolt, S., & Thurlow, M. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education, 25*(3), 141 - 152.

Bolt, S., & Thurlow, M. (2006). *Item-level effects of the read-aloud accommodation for students with reading disabilities (NCEO Synthesis report 65).* National Center on Educational Outcomes, University of Minnesota, MN. Retrieved from http://cehd.umn.edu/nceo/OnlinePubs.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Bouck, E. C., & Yadav, A. (2008). Assessing calculators as assessment accommodations for students with disabilities. Assistive Technology Outcomes and Benefits, 5(1), 19 - 28.

(2) Brown, D. W. (2007). *The role of reading in science: Validating graphics in large-scale science assessment.* Unpublished doctoral dissertation, University of Oregon, Oregon.

(3) Buehler, K. L. (2002). *Standardized group achievement tests and the accommodation of additional time.* (Doctoral dissertation, Indiana State University, 2001). Retrieved from ProQuest Dissertations & Theses. (AAT 3050241)

(4) Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23*(4), 271 - 282.

Christensen, L., Lazarus, S., Crone, M., & Thurlow, M. (2008). *2007 state policies on assessment participation and accommodations for students with disabilities (NCEO Synthesis Report No. 69*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes Retrieved from http://cehd.umn.edu/nceo/OnlinePubs/Synthesis69/Synthesis69.pdf

Chiu, C., & Pearson, P. (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficiency students.* Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT (ERIC Document Reproduction Service No. ED 433 362).

Clapper, A., Morse, A., Lazarus, S., Thompson, S., & Thurlow, M. (2003). *2003 state policies on assessment participation and accommodation for students with disabilities (NCEO Synthesis Report 56).* Retrieved from: http://www.cehd.umn.edu/nceo/OnlinePubs/Synthesis56.html

Code of Fair Testing Practices in Education (2004). Washington, DC: Joint Committee on Testing Practices. Retrieved from http://www.apa.org/science/FinalCode.pdf

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155 - 159.

Cooper, H., & Hedges, L. (Eds.). (1994). *The handbook of research synthesis.* New York, NY: Russell Sage Foundation.

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008 (NCEO Technical Report 56*). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://cehd.umn.edu/nceo/OnlinePubs

Council for Exceptional Children (2002). *No child left behind act of 2001: Implications for special education policy and practice.* Retrieved March from http://www.cec.sped.org/AM/Template.cfm?Section=Search&section=Public_Policy&template=/CM/ContentDisplay.cfm&ContentFileID=1366

(05) Crawford, L., Helwig, R., & Tindal, G. (2004). Writing performance assessments: How important is extended time? *Journal of Learning Disabilities, 37*(2), 132 - 142.

Crawford, L., & Tindal, G. (2006). Policy and practice: Knowledge and beliefs of education professionals related to the inclusion of students with disabilities in a state assessment. *Remedial and Special Education, 27*(4), 208 - 217.

(06) Dempsey, K. M. (2003). *The impact of additional time on LSAT scores: Does time really matter? The efficacy of making decisions on a case-by-case basis.* (Doctoral dissertation, La Salle University, 2004). Retrieved from ProQuest Dissertations & Theses. (AAT 3108290)

Dillon, E. (July 17, 2007). *Charts you can trust: Labeled: The students behind NCLB's 'disabilities' designation.* Retrieved from http://www.educationsector.org/analysis/analysis_show.htm?doc_id=509392

Dominus, S. (2011, April 20). The crash and burn of an autism guru. *The New York Times*. Retrieved from http://www.nytimes.com/2011/04/24/magazine/mag-24Autism-t.html/?pagewanted=all

(07) Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *The Journal of Special Education, 40*(4), 218 - 229.

(08) Elbaum, B., Arguelles, M. E., Cambpell, Y., & Saleh, M. B. (2004). Effects of a student-read aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality, 12*(2), 71 - 87.

Elliott, J., Erickson, R., Thurlow, M., & Shriner, J. (2000). State-level accountability for the performance of students with disabilities: Five years of change? *The Journal of Special Education, 34* (1), 39 - 47.

Elliott, J., Thurlow, M., Ysseldyke, J., & Erickson, R. (1997). P*roviding assessment accommodations for students with disabilities in state and district assessments (Policy Directions No. 7).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Policy7.html

Elliott, S. N., Kratochwill, T. R., & McKevitt, B. C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 39*(1), 3 - 24.

(09) Engelhard, G., Jr., Fincher, M., & Domaleski, C. S. (2011). Mathematics performance of students with and without disabilities under accommodated conditions using resource guides and calculators on high stakes tests. *Applied Measurement in Education, 24*(1), 22 - 38.

Enriquez, M. (2008). *Examining the effects of linguistic accommodations on the Colorado student assessment program-mathematics.* (Doctoral Dissertation, University of Denver, 2008). Retrieved from ProQuest Dissertations & Theses. (AAT 3337051)

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator, 56*(10), 24-27.

Fuchs, L., Fuchs, D., & Capizzi, A. M. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children, 37*(6), 1 - 8.

(10) Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C., & Karns, K. (2000a). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65 - 86.

(11) Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000b). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67*(1), 67 - 81.

Gilman, C., Thurlow, M., & Ysseldyke, J. (1993). *Responses to working paper 1 conceptual model of educational outcomes for children and youth with disabilities (Synthesis Report No. 3).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/12/ec/d6.pdf

Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage Publications.

Gregg, M., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities, 45*(2), 128 - 138.

Hansen, E., & Mislevy, R. (2008). *Design patterns for improving accessibility for test takers with disabilities (ETS RR-08-49).* Princeton, NJ: Educational Testing Service.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29 - 38). New York, NY: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*, New York, NY: Academic Press.

(12) Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *The Journal of Special Education, 36*(1), 39 - 47.

(13) Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*(2), 211 - 225.

(14) Huesman, R. L. (1999). *The validity of ITBS reading comprehension test scores for learning disabled and non-learning disabled students under extended-time conditions.* (Doctoral dissertation, University of Iowa, 1999). Retrieved from ProQuest Dissertations & Theses. (AAT 304511173)

Higgins, J., & Thompson, S. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23*, 1663 - 1682.

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis: correcting error and bias in research findings.* Newbury Park, CA: SAGE Publications.

Individuals with Disabilities Education Act Amendments of 1997, Pub. L. No. 105-17. (1997). Retrieved from http://www.nectac.org/idea/pl105-17.asp

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446. (2004). Retrieved from http://www.copyright.gov/legislation/pl108-446.pdf

Johnstone, C. J., Altman, J., Thurlow, M., & Thompson, S. J. (2006). *A summary of the research on the effects of tests accommodations: 2002 through 2004 (Technical Report 45).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

(15) Janson, I. B. (2002). *The effects of testing accommodations on students' standardized test scores in a northeast Tennessee school system.* (Doctoral dissertation, East Tennessee State University, 2002). Retrieved from ProQuest Dissertations & Theses. (AAT 3042421)

(16) Johnson, E. (2000). The effects of accommodations on performance assessments. *Remedial and Special Education, 21*, 261 - 267.

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

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168 - 1201.

(18) Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test (Technical Report 28)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

(19) Laitusis, C. C. (2010). Examining the impact of audio presentation on tests of reading comprehension. *Applied Measurement in Education, 23*(2), 153 - 167.

Lam, R. W., & Kennedy, S. H. (2005). Using metaanalysis to evaluate evidence: Practical tips and traps. *Canadian Journal of Psychiatry, 50*(3), 167 - 174.

Lazarus, S., Thurlow, M., Lail, K., Eisnebraun, K., & Kato, K. (2005). *2005 state policies on assessment participation and accommodation for students with disabilities (NCEO Synthesis Report 64).* Retrieved from http://www.cehd.umn.edu/nceo/OnlinePubs/Synthesis64/default.html

(20) Lee, D., & Tindal, G. (2000). *Teacher's perception on students' reading performance and test accommodation outcomes (Attachment 6).* Dover, DE: Delaware Department of Education. Retrieved from http://www.doe.k12.de.us/aab/Report_and_documents/ICAS.shtml

(21) Lesaux, N. K., Pearson, M. R., & Siegel, L. S. (2006). The effects of timed and untimed testing conditions on the reading comprehension performance of adults with reading disabilities. *Reading and Writing, 19*, 21 - 48.

(22) Lewandowski, L. J., Lovett, B. J., Parolin, R., Gordon, M., & Codding, R. S. (2007). Extended time accommodations and the mathematics performance of students with and without ADHD. *Journal of Psychoeducational Assessment, 25*, 17 - 28.

Lewandowski, L. J., Lovett, B. J., & Rogers, C. L. (2008). Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships? *Journal of Psychoeducational Assessment, 26*, 315 - 324.

Lewin, T. (2002, March 18). In testing, one size may not fit all. *The New York Times*, 1 - 3. Retrieved from http://www.nytimes.com/2002/03/18/us/in-testing-one-size-may-not-fit-all.html?scp=9&sq...

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment, 7*(1), 29 - 38.

(23) MacArthur, C., & Cavalier, A. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43 - 58.

McGrew, K., Thurlow, M., Shriner, J., & Spiegel, A. (1992). *Inclusion of students with disabilities in national and stated data collection programs (Technical report 2).* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/24/36/0f.pdf

(24) Marquart, A. M. (2000). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels.* (Doctoral dissertation, University of Wisconsin-Madison, 2000). Retrieved from ProQuest Dissertations & Theses. (AAT 9982212)

(25) Medina, J. (1999). *Classroom testing accommodations for postsecondary students with learning disabilities: The empirical gap.* (Doctoral dissertation, Alfred University, 2000). Retrieved from ProQuest Dissertations & Theses. (AAT 9939751)

(26) Meloy, L., Deville, C., & Frisbie, D. (2002). The effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23*(4), 248 - 255.

Merwin, J. (1993). Inclusion and accommodation: "You can tell what is important to a society by the things it chooses to measure". In National Center on Educational Outcomes (Ed.). *Views on inclusion and testing accommodations for student with disabilities.* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/3a/03.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (3rd Ed.)* (pp. 13 - 103) New York, NY: Macmillan.

Messick, S. (1990). *Validity of test interpretation and use.* ETS Report ETS-RR-90-11, (pp. 1 – 33).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741 - 749.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105 - 125.

Morton, S., Adams, J., Suttorp, M., & Shekelle, P. (2004). *Meta-regression approaches: What, why, when, and how? (Technical Review 8).* Prepared by Southern California–RAND Evidence-based Practice Center, under Contract No 290-97-0001. AHRQ Publication No. 04-0033. Rockville, MD: Agency for Healthcare Research and Quality.

National Association of Special Education Teachers (n.d.). *Introduction to learning disabilities: Definition of learning disabilities.* Retrieved from http://www.naset.org/2522.0.html.

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2002).

(27) Ofiesh, N., Mather, N., & Russell, A. (2005). Using speeded cognitive, reading, and academic measures to determine the need for extended test time among university students with learning disabilities. *Journal of Psychoeducational Assessment, 23*, 35 - 52.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10 - 28.

Pennock-Roman, M., & Rivera, C. (2012, April). *Comparing the performance of focal and reference groups on test accommodations: A new index to evaluate differential boost.* Paper presented at the Annual Meeting of the American Educational Research Associations, Vancouver, BC.

(28) Randall, J., & Engelhard, G., Jr. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education, 23*(3), 286 - 306.

Reschly, D. (1993). Consequences and incentives: Implications for inclusion/exclusion decisions regarding students with disabilities in state and national assessment programs. In National Center on Educational Outcomes (Ed.). *Views on inclusion and testing accommodations for student with disabilities.* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/3a/03.pdf

Reynolds, M. (1993). Inclusion and accommodations in assessment at the margins. In National Center on Educational Outcomes (Ed.). *Views on inclusion and testing accommodations for student with disabilities.* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/3a/03.pdf

(29) Schnirman, R. K. (2005). *The effect of audiocassette presentation on the performance of students with and without learning disabilities on a group standardized math test.* (Doctoral dissertation, Florida Atlantic University, 2005). Retrieved from ProQuest Dissertations & Theses. (66(6))

Section 504 of the Rehabilitation Act, 29 U.S.C. §794 (1973).

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

Sireci, S., & Pitoniak, M. (2007, April). *Assessment accommodations: What have we learned from research?* Paper presented at the Annual Meeting of the American Educational Research Associations, Chicago, IL.

Sireci, S., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457 - 490.

(30) Smith, G. W. (2010). *The impact of a noise-reducing learning accommodation utilized by students with learning disabilities during an independent reading inventory.* (Doctoral dissertation, Clemson University, 2010). Retrieved from ProQuest Dissertations & Theses. (AAT 3402556)

Stanley, T. (2001). Wheat from chaff: Meta-analysis as quantitative literature review, *Journal of Economic Perspectives, 15*, 131 - 150.

Stock, W. A. (1994). Systematic coding for research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 125 - 138). New York, NY: Russell Sage Foundation.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001 (Technical Report 34).* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thompson, S., & Higgins, J. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21,* 1559 - 1573.

Thompson, S., & Sharp, S. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18*(20), 2693 - 2708.

Thurlow, M. (2007, April). *Research impact on state accommodation policies for students with disabilities.* Paper presented at the Annual Meeting of the American Educational Research Associations, Chicago, IL.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy (Synthesis Report 41).* Minneapolis, MN: University of Minnesota National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/nceo/OnlinePubs/Synthesis41.html

Thurlow, M., McGrew, K., Tindal, G., Thompson, S., Ysseldyke, J., & Elliott, J. (2000). *Assessment accommodations research: Considerations for design and analysis (NCEO Technical Report 26).* Minneapolis, MN: University of Minnesota

National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/NCEO/OnlinePubs/Technical26.htm

Thurlow, M., Lazarus, S., Thompson, S., & Blount Morse, A. (2005). State policies on assessment participation and accommodations for students with disabilities. *The Journal of Special Education, 38* (4), 232 - 240.

Thurlow, M., Seyfarth, A., Scott, D., & Ysseldyke, J. (1997). *State assessment policies on participation and accommodations for students with disabilities: 1997 update (NCEO Synthesis Report 29).* Minneapolis, MN: University of Minnesota National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/NCEO/OnlinePubs/Synthesis29.html

(31) Tindal, G. (2002). *Accommodating mathematics testing using a videotaped, read-aloud administration.* Washington, DC: Council of Chief State School Officers. Retrieved from http://www.eric.ed.gov/PDFS/ED473011.pdf

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: Mid-South Regional Resource Center. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/42/59.pdf

Tindal, G., & Ketterlin-Geller, L. (2004). *Research on mathematics test accommodations relevant to NAEP testing.* Washington, DC: National Assessment Governing Board.

U.S. Department of Education, Office for Civil Rights, Free Appropriate Public Education for Students With Disabilities: Requirements Under Section 504 of the Rehabilitation Act of 1973, Washington, D.C., 2007. Retrieved from http://www.ed.gov/about/offices/list/ocr/docs/edlite-FAPE504.html

Van Horn, P., Green, K. E., & Martinussen, M. (2009). Survey response rates and survey administration in counseling and clinical psychology: A meta-analysis. *Educational and Psychological Measurement, 69,* 389 - 403.

(32) Villeneuve, L. C. (2009). *An examination of extended test time. Do students with reading disabilities really need more time?* (Doctoral dissertation, Laurentian University, 2009) Retrieved from ProQuest Dissertations & Theses. (AAT MR48872)

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41 - 53). New York: Russell Sage Foundation.

(33) Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students (Minnesota Report 34).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 26, 2010 from http://www.cehd.umn.edu/NCEO/onlinepubs/archive/AssessmentSeries/MnReport34.html

(34) Weston, T. J. (2002). The validity of oral accommodation in testing (NCES 200306). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2003/200306.pdf

Wilson, D. B. (2006). Analysis.ppt. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Wilson, D. B. (2005). Meta-analysis macros for SAS, SPSS, and Stata. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Ysseldyke, J., & Thurlow, M. (1993). *Developing a model of educational outcomes. Outcomes & Indicators: (NCEO Report, 1).* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/3a/2a.pdf

Ysseldyke, J., Thurlow, M., McGrew, K., & Shriner, J. (1994). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs (NCEO Synthesis Report 15*) Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://cehd.umn.edu/nceo/OnlinePubs/SynthesisReport015.pdf

Ysseldyke, J., Thurlow, M., Kozleski, E., & Reschly, D. (1998). *Accountability for the results of educating students with disabilities: Assessment conference report on the new assessment provisions of the 1997 amendments to the individuals with disabilities act.* Minneapolis, MN: National Center on Educational Outcomes. Retrieved from http://cehd.umn.edu/nceo/OnlinePubs/awgfinal.html

Zenisky, A., & Sireci, S. (2007). *A summary of the research of the effects of test accommodations: 2005 through 2006 (NCEO Technical Report 47).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/nceo/OnlinePubs/Tech47/TechReport47.pdf

Zurcher, R., & Bryant, D. P. (2001). The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities, 34*(5), 462 - 471.

Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13*(1), 99 - 117.

Appendices

Appendix A

Individuals with Disabilities Act reauthorization of 2004, PUBLIC LAW 108–446, 2004

118 STAT.2652

Part A – General Provisions; 20 USC 1401. ‘‘SEC. 602. DEFINITIONS. ‘‘(3) CHILD WITH A DISABILITY.—‘‘(A) IN GENERAL.—The term ‘child with a disability' means a child—‘‘(i) with mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as ‘emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities; and ‘‘(ii) who, by reason thereof, needs special education and related services.

(Part A (SEC. 602) (3) (A) (i), 118 STAT.2652, 2004)

‘‘PART D—NATIONAL ACTIVITIES TO IMPROVE EDUCATION OF CHILDREN WITH DISABILITIES (SEC. 650) FINDINGS. Congress finds the following: ‘‘(4) An effective educational system serving students with disabilities should— ‘‘(A) maintain high academic achievement standards and clear performance goals for children with disabilities, consistent with the standards and expectations for all students in the educational system, and provide for appropriate and effective strategies and methods to ensure that all children with disabilities have the opportunity to achieve those standards and goals; ‘‘(B) clearly define, in objective, measurable terms, the school and post-school results that children with disabilities are expected to achieve; and ‘‘(C) promote transition services and coordinate State and local education, social, health, mental health, and other services, in addressing the full range of student needs, particularly the needs of children with disabilities who need significant levels of support to participate and learn in school and the community.

((SEC. 650) (4) (A), (B), and (C), 118 STAT. 2763, 2004)

Part B – Assistance for Education of All Children with Disabilities (SEC. 611) AUTHORIZATION; ALLOTMENT; USE OF FUNDS; AUTHORIZATION OF APPROPRIATIONS (e) STATE-LEVEL ACTIVITIES.— (2) OTHER STATE-LEVEL ACTIVITIES.— (C) AUTHORIZED ACTIVITIES.—Funds reserved under subparagraph (A) may be used to carry out the following activities: (x) To support the development and provision of appropriate accommodations for children with disabilities, or the development and provision of alternate assessments that are valid and reliable for assessing the performance of children with disabilities, in accordance with sections 1111(b) and 6111 of the Elementary and Secondary Education Act of 1965.

IDEA, Part B (SEC. 611) (e) (2) (C) (x), 118 STAT.2667– 118 STAT.2668)

Appendix B

No Child Left Behind, PUBLIC LAW 107-110, 2002

Improving basic programs operated by local educational agencies (1111) state plans (b) academic standards, academic assessments and accountability, (2) accountability (C) definition – 'Adequate yearly progress' shall be defined by the State in a manner that— (v) includes separate measurable annual objectives for continuous and substantial improvement for each of the following: (II) The achievement of— (cc) students with disabilities (NCLB, 2002, 115 STAT. 1446)

Appendix C

Institutional Review Board (IRB) Exemption

From: Sylk Sotto-Santiago [mailto:Sylk.Sotto-Santiago@du.edu]

Sent: Thursday, January 28, 2010 10:31 AM

To: Kathy Green

Subject: Michelle Vanchu-Orosco

Dear Kathy:

Thanks for checking with me regarding the work by Michelle Vanchu-Orosco. After several emails gathering all the facts. I have determined in consultation with the IRB Chair, Susan Sadler that the study is not under the purview of the IRB based on the following information provided by you:

"Michelle's studies come from published journal articles, published state department reports (from state department of education websites), and conference papers. She has no data other than aggregated data, and no individually identifying information. She is NOT using a database from any institution. For example, she IS using statistics taken from papers presented at AERA available via ERIC, articles from the Journal of Special Education, etc."

Under 45 CFR 46, publicly available data sets that are completely deidentified do not require IRB review. It is safe to say that if the data set has been published in a journal article that it becomes public. In addition, we do not consider this as research with human subjects based on the definition of human subjects: a living individual about whom an investigator conducting research obtains (1) Data through intervention or interaction with the individual, or (2) Identifiable private information. Michelle is not interacting, intervening (she is not the researcher) or obtaining identifiable private information (her data is public and does not contain identifiable information).

Please note that this is NOT a blanket statement. There is the question if data sets are really public, which must be assessed on a case by case basis. DU does not have list of pre-approved public sets, some data sets are limited-data use or restricted access per some agreement, and for example Exempt Category 4 might very well apply in some instances for existing data use.

I hope this helps. Please let me know if you have any questions.

Appendix D

Test Accommodations Meta-Analysis Coding Manual

Coding Manual

This coding manual contains information, such as rejection rules, to be used when reviewing and coding studies using the codebook and coding sheet (see Appendix E for an example of the codebook; see Appendix F for an example of the coding form) for the current meta-analysis. It provides the rationale for inclusion and exclusion of studies from the present meta-analysis analyses.

Those involved with coding a randomly selected group of studies used for inter-rater reliability purposes started coding from *Source Information* onward. Previous sections, *Report Identification*, *Research Study Identification*, and *Study Retrieval*, were verified only.

Many of the coding sections provide one or more criteria for rejecting studies. If a study is rejected based on these criteria *Research Study Information* still must be collected. This *Source Information* was added to Appendix H (Citations for Duplicate and Excluded Studies).

Prior to Coding

To aid in the aggregation and comparison of the research findings for each of the research studies identified only quantitative empirical research focusing on the use of testing accommodations for students with disabilities on large-scale and/or high-stakes tests (assessments), the domain of interest for the researcher, will be included in the meta-analysis.

When collecting the research studies to be used in the meta-analysis the title of the research study, abstract or executive summary information, and research questions or purpose provided are to be reviewed to determine if the study was potentially eligible for inclusion in present meta-analysis.

The following rules are to be used to exclude studies prior to coding. Citation information was not collected for these studies as the focus of these studies did not align with the research hypotheses for the present study.

> ***REJECTION RULE:*** If the research was conducted prior to 1999, or 1999 and later and cited in Chui and Pearson (1999) – do not code study and do not count as an eligible study
>
> ***REJECTION RULE:*** If the research is not reported in English, or an English translation is not available – do not code study and do not count as an eligible study
>
> ***REJECTION RULE:*** If the research examines assessment modifications and not assessment accommodations – do not code study and do not count as an eligible study (See the ***Glossary*** located at the end of this Coding Manual for definitions of assessment accommodation and assessment modification).
>
> ***REJECTION RULE:*** If the research uses alternative assessments or tests (also called alternate assessments) – do not code study and do not count as an eligible study.
>
> ***REJECTION RULE:*** If the research does not include students with disabilities – do not code study and do not count as an eligible study.
>
> ***REJECTION RULE:*** If the research only focuses on English language learners (ELL) or English language learners with disabilities – do not code study and do not count as an eligible study.
>
> ***REJECTION RULE:*** If the research focuses on computer-based testing (CBT) – do not code study and do not count as an eligible study.
>
> ***REJECTION RULE:*** If the research focuses on the comparison of computer-based assessments to paper and pencil assessments – do not code study and do not count as an eligible study.

REJECTION RULE: If the research focuses on individual test items (e.g., not results for the entire assessment or sub-section of the assessment) – do not code study and do not count as an eligible study.

REJECTION RULE: If the research only uses survey methodology (e.g., surveys students, parents, teachers, or administrative staff) – do not code study and do not count as an eligible study.

REJECTION RULE: If the research focuses on policy analysis (e.g., investigation of accommodation decision making and administrative practices though (i) policy presentation, (ii) policy interpretation, (iii) test accommodation implementation analysis) – do not code study and do not count as an eligible study.

REJECTION RULE: If the research only uses qualitative methodology (e.g., ethnography) – do not code study and do not count as an eligible study.

REJECTION RULE: If the purpose of the research focuses a secondary analysis of existing studies, (e.g., literature review or meta-analysis) – do not code study and do not count as an eligible study.

REJECTION RULE: If the research uses factor analysis, structural equation modeling (SEM), item response theory (IRT), or differential item functioning (DIF) – do not code study and do not count as an eligible study.

REJECTION RULE: If the research uses a single-subject design – do not code study and do not count as an eligible study.

REJECTION RULE: If the purpose of the research does not conform to the following:

> determination of the effect of the assessment accommodation on the scores of (i) students with disabilities (ii) students with disabilities as compared to typically developing peers - do not code study and do not count as an eligible study.

The following rules are to be used to exclude studies prior to coding. Citation information was collected for these studies as the focus of these studies aligned with the research hypotheses for the present study although the type of assessment used (e.g., low-stakes) or the method to analyze the information (e.g., correlation) was not applicable to the present study.

> ***REJECTION RULE:*** If the research only provides correlational information – enter citation information and discontinue coding.

> ***REJECTION RULE:*** If the research does not use high-stakes or large-scale assessments, or their proxies – enter citation information and discontinue coding.

> ***REJECTION RULE:*** If the same research was found in multiple sources (e.g., dissertations, papers, journals, etc.), assign a single ID # to the studies with an alpha character appended (e.g., 01A, 01B, 01C, etc). Select the study with the most information from the group for coding then – enter citation information for the remaining studies and discontinue coding.

Coding Eligible Studies

Note 1: Zero (0) is used as the initial code for each variable coded with 0 = N/A, Not Reported, or No

Note 2: Not all studies considered eligible for coding will be included in the present meta-analysis. Citation information and reason for excluding the study will be collected during the coding phase and added to Appendix H (Citations for Duplicate and Excluded Studies).

Note 3:
For research containing more than one research study (e.g., date for grade 3 and grade 6 that was reported separately), create a new record and complete coding form for each 'sub-study'.

Report Identification

Note: This coding is to be completed by primary researcher.

- Enter a two-digit code, starting with 01.

Research Study Identification

Research study information contains citation information for each research study (unit of analysis) located. There are two sections related to study identification (i) citation information and (ii) publisher information.

- Research Study Citation

- Publisher (use Publisher information to track the type of publication and the publication source)
  - ◊ Type of publication refers to the method used to report results; for example, journal
    - ♦ Enter ***(0) uncategorized*** if unable to place research study within the context of the categories provided, or if unable to provide a classificatory name to the method used to disseminate the information in the research study
    - ♦ Enter ***(1) journal*** if the research was reported in a journal then enter the publication source code. If the name of the journal is not listed, enter it under 'other' in the blank space provided.
    - ♦ Enter ***(2) conference proceedings (paper)*** if the research was reported at a conference (e.g., paper, symposia …) then enter the publication source code. If the name of the organization sponsoring the conference is not listed, enter it under 'other' in the blank space provided.
    - ♦ Enter ***(3) organization (report)*** if the research was reported on an organizational website (e.g., report) then enter the publication source code. If the name of the organization sponsoring the website is not listed, enter it under 'other' in the blank space provided.
    - ♦ Enter ***(4) dissertation*** if the research was reported in a dissertation then enter the publication source code.
    - ♦ Enter ***(5) manuscript*** if the research was reported in an unpublished manuscript.

- ♦ Enter ***(6) other*** if the research was reported in a source other than those listed in categories 1 – 5 and there is a descriptor provided for the method of disseminating the research study.

- ◊ Publication source refers to citation information for the journal, report, dissertation, or paper located. This information does not include date as it is tracked at an earlier point in the coding.

Note 1: There are no rejection rules for *Research Study Identification* information.

Note 2: This coding must be completed by primary researcher.

Study Retrieval

Method used to locate the research study by the database used is tracked in this section. This information will be used to provide demographic information regarding the method used to retrieve studies.

- Method to locate study

Enter the code for the method used to locate the research study; for example, ***2C*** (2 = references in eligible studies and C = ERIC)

Note 1: There are no rejection rules for Study Retrieval information.

Note 2: This coding must be completed by primary researcher

Research Quality

A proxy value for research quality, reviewed versus not reviewed, will be used to provide information on the number of reviewed studies versus not reviewed studies.

Enter the code for the quality of the research study; for example, ***2*** (2 = published dissertation)

Note 1: There are no rejection rules for Research Quality information.

Note 2: This coding must be completed by primary researcher

Research Participant Information

There are six sections related to research participant information. Select ***(0) not reported*** for sections when information related to the specific section cannot be found.

Note: Typically developing students are often referred to as students without disabilities in the research literature.

Research Participant Information sections

- Participant data source

  Select the source for the research participant sample from the list provided. If the researchers collected data from subjects involved in the study select 'Primary data collection'. If the researchers used a database (e.g., NAEP), select 'Secondary/archival data collection'. If it is not apparent as to where the data presented are from, select 'not reported'

- Participant sampling method

  Select the source for the research participant sample from the list provided. For example, research participants may have been ***randomly selected*** from a ***single school district***. See the ***Glossary*** located at the end of this Coding Manual for definitions of each sampling method.

  Note: using the entire population is not considered a sampling method. It is included in the sampling method section solely for purposes of tracking the data for the meta-analysis.

- Participant sampling method (additional information)

Enter information regarding the number of schools, districts, states, that participated in the research study.

- Participant assignment

  Select the method used to assign participants to conditions from the list provided. For example, if researchers randomly assigned intact classes to non-accommodated and accommodated conditions select (3) and (B) for random assignment at the classroom level.

- Participant grade level(s)

  Select the grade level from the list provided. Multiple grade levels may be selected if research participants were from multiple grades.

- Participant sample composition

  For each group of participants (students with disabilities; typically developing students), enter the number of participants completing the study (i.e., students who took the tests).

  REJECTION RULE: If participant group information (participant group size and/or type of participants) is not available, ensure citation information is entered and discontinue coding. Add the following note to the citation:
  'missing participant sample/group size'
  <<or>>
  'missing information on type of participants'

  ◊ Total number of research participants

  ♦ Enter the total number of research participants completing the study.

Note: The number of research participants may not be the same as the final number of participants used in the research analysis/analyses. Ensure the number entered is the total number of students completing the study (i.e., students who took the tests).

- Participant disability classification

  For each disability group listed in the research study, select the appropriate disability classification using the Special Education Taxonomy (Appendix E, codebook) on the first line. Then enter the number of participants for that disability classification completing the study (i.e., students who took the tests) on the second line. If the disability group is not present in the Special Education Taxonomy, enter it under 'other' in the blank space provided.

Note: 'Other' is also used to track overall group; students requiring special education services. If the primary research study refers to students/individuals requiring special education services, select 'other' then select either 'representative sample' or 'not representative sample'

Assessment Information

Nine sections will be coded to capture the salient characteristics of the assessment tool used for the research conducted. Select ***(0) not reported*** for sections when information related to the specific section cannot be found.

Note that assessment, measure, test, instrument, and scale are often used interchangeably. However, these terms are not synonymous. See the ***Glossary*** located at the end of this Coding Manual for definitions of each term.

Assessment Information sections

- Citation Information

  Citation information such as name and publication date for the assessment instrument will be collected in the assessment citation section.

- Assessment Classification

  ◊ Type of Assessment

Select the type of assessment used from the list provided. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of assessments. If the type of assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

◊ Assessment Descriptors

Select descriptors for the assessment from the list provided. See the ***Glossary*** located at the end of this Coding Manual for descriptor definitions. If an appropriate descriptor for the assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

> ***REJECTION RULE:*** If the assessment/test is not considered 'high-stakes', 'large-scale' or 'standardized' ensure citation information is entered and discontinue coding. Assessments used in the study research coded should be part of the decision-making process and have prominent educational/financial/social impact. For example, coding a research study which uses a criterion-referenced test to inform class instruction would be discontinued after ensuring citation information had been recorded. Add the following note to the citation:
>
> **'**research uses low-stakes (classroom, etc) assessment'

◊ Assessment Categorization

Select the category for the assessment from the list provided. See the ***Glossary*** located at the end of this Coding Manual for a definition of each category. If an appropriate category for the assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

If the assessment is used to measure achievement, aptitude, and/or performance, proceed to the ***Assessment Content/Construct*** subsection and select the content/construct area measured by the assessment. Otherwise, discontinue coding and add the following note to the citation:

'research uses assessment other than achievement/aptitude/performance assessment'

◊ Assessment Content/Construct

Select the content/construct measured by the assessment from the list provided. See the ***Glossary*** located at the end of this Coding Manual for a definition of each category. If an appropriate content area/construct for the assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

REJECTION RULE: If the assessment/test examines physical skills or other non-academic areas ensure citation information is entered and discontinue coding. Add the following note to citation:
'physical skills (attitudes, etc.) assessed'

REJECTION RULE: If the assessment/test examines psychomotor skills or aptitudes (of or pertaining to a response involving both motor and psychological components) ensure citation information is entered and discontinue coding. Add the following note to citation:
'psychomotor skills (psychomotor aptitudes) assessed'

REJECTION RULE: If the assessment/test examines personality (e.g., individual traits and characteristics), attitude, affect or interest ensure citation information is entered and discontinue coding. Add the following note to citation:
'personality (attitude, affect, interests, etc.) assessed'

◊ Assessment Format

Select the assessment format; e.g., the format for the questions used on the assessment, from the list provided. If an appropriate format for the assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

◊ Number of Assessment Forms

Select the number of assessment forms used in the study; i.e., the number of forms used to collect data, from the list provided.

◊ Reliability

Indicate whether or not reliability information was provided for the assessment used. If reliability information was provided for the assessment, select the type of reliability reported from the list provided in the ‘Reliability Type’ subsection and provide the reliability index value in the blank provided. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of reliability.

◊ Validity

Indicate whether or not validity information was provided for the assessment used. If validity information was provided for the assessment, select the type of validity reported from the list provided in the ‘Validity’ Type subsection and provide the validity index value in the blank provided. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of validity.

Accommodation Information

Test accommodation information is captured in a single section. Categories coded in this section include n/a, not reported, four major test accommodation types (presentation, response, setting, and timing/scheduling), multiple accommodations and other. Additionally, each test accommodation type is further refined to provide more granular information regarding the test accommodation type.

Although, test accommodation information may be thought of as a test administration procedure, such as group administration, this type of coding is considered to be redundant for purposes of coding studies located for the current meta-analysis.

Note: Test accommodations will be rolled up to the four test accommodation types if there are less than 5 studies in a subsection. If there are less than 5 studies for the test accommodation type that test accommodation will be dropped from the results analysis. Select ***(0) n/a*** if the research examines test modification(s) and/or does not include one or more test accommodations as part of the study. Ensure citation information has been entered and discontinue coding.

- Accommodation Information

  Select ***(1) not reported*** when information related to testing accommodation(s) cannot be found. Ensure citation information has been entered and discontinue coding.

  Select ***(2) Presentation Accommodation*** through ***(5) Response Accommodation***, at the level of granularity found in the research under consideration. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of testing accommodations.

  In the case of multiple test accommodations, select ***(6) Multiple Accommodations / Accommodation Packages*** and list each test accommodation in the space provided. Use the codes provided under *Accommodation Information* (e.g., if the research examined administering the assessment in a separate location for each student and the provision of frequent breaks during testing you would select ***(3A) individual administration in a separate location*** and ***(4B) allow frequent breaks during testing***).

  > ***REJECTION RULE:*** If the test accommodation under investigation is an assessment accommodation package (i.e., more than one accommodation per individual) ensure citation information is entered and discontinue coding. Add the following note to the citation:

‘ assessment accommodation package’

REJECTION RULE: If the test accommodation(s) under investigation are not specified (e.g., listed generically as ‘test accommodation’ or ‘assessment accommodation’) ensure citation information is entered and discontinue coding. Add the following note to the citation:

‘ assessment accommodation(s) not specified’

REJECTION RULE: If the test accommodation under investigation utilized an interpreter for purposes of translating directions from one language into another language (e.g., the study focuses on English language learners) ensure citation information is entered and discontinue coding. Add the following note to the citation:

‘language interpretation accommodation’

REJECTION RULE: If the test accommodation under investigation is a computerized accommodation ensure citation information is entered and discontinue coding. Add the following note to the citation:

‘ computerized accommodation’

Note: this would only occur if the study title, abstract, and research purpose did not focus on the use of a computerized accommodation.

Select ***(7) other*** if test accommodation information found in the research under consideration cannot be categorized using codes found in the Accommodation Information section (e.g., technological aid). Enter the test accommodation listed in the research in the blank provided.

Research Study Design Information

Four sections will be coded to capture the salient characteristics of the research design used for the research conducted. Select ***(0) not reported*** for sections when information related to the specific section cannot be found.

- Study Type

Select the type of study; e.g., Experimental, from the list provided.

REJECTION RULE: If the research methodology is not reported, non-experimental, or observational ensure citation information was entered and discontinue coding. Add the following note to the citation (based on type of research method; e.g.,):

'not applicable research method – not experimental'

REJECTION RULE: If the research methodology is descriptive/quantitative (e.g., logical analyses of the difficulties associated with disabilities are conducted to determine what accommodations are considered helpful for students to be able to demonstrate knowledge and skills on a test) ensure citation information was entered and discontinue coding. Add the following note to the citation (based on type of research method; e.g.,):

'not applicable research method – descriptive'

REJECTION RULE: If the research methodology is individual diagnosis (e.g., uses a set procedure for determining which accommodations an individual student should receive) ensure citation information was entered and discontinue coding. Add the following note to the citation (based on type of research method; e.g.,):

'not applicable research method – individual diagnosis'

REJECTION RULE: If the research methodology is something other than those listed in the code book (e.g., not comparative, quasi-experimental, or experimental) ensure citation information was entered and discontinue coding. Add the following note to the citation (based on type of research method; e.g.,):

'not applicable research method – other'

Methodology

Three sections, nested, capture the salient characteristics regarding the research design used for the study. Select ***(0) not reported*** for sections when information related to the specific section cannot be found. See the ***Glossary*** located at the end of this Coding Manual for unfamiliar terms.

- Methodology

  ◊ Research Approach

  Select the type of research approach used from the list provided. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of assessments. If the type of assessment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

> ***REJECTION RULE:*** If the primary study does not use comparison, boost, or differential boost/interaction hypothesis research approaches ensure citation information is entered and discontinue coding. Add the following note to the citation:

◊ Research Design

Select the type of research design used from the list provided. See the ***Glossary*** located at the end of this Coding Manual for definitions of the different types of assessments.

◊ Research Design Variation

Select the research design variation used from the list provided. If a specific design variant used is not listed, select a similar design variation and make a note regarding the differences between the two design variations.

- Accommodation Order

Select the research design variation used from the list provided.

Statistical Method

Select the statistical method from the list provided. If the statistical method used is not listed or cannot be determined, enter it under 'other' in the blank space provided.

> ***REJECTION RULE:*** If research employed a methodology that did not include means, standard deviations, and number of research participants, or some equivalent which can be used to estimate the effect size for the study *and* was not previously eliminated, ensure citation information is entered and discontinue coding. Add the following note to the citation:
>
> 'not applicable statistical method - <<name of statistical method>>'

Results Information

Results information collected will consist of information on participant assignment to condition and statistics used for research conducted. These statistics will be used to calculated the standardized mean effect size.

Results are recoded for each participant group (e.g., students with disabilities by type of disability, and typically developing students). This is to be completed for each relevant participant group that is found in the research study.

- Participant Assignment

  Select the research design variation used from the list provided. If the type of participant assignment is not listed or cannot be determined, enter it under 'other' in the blank space provided.

- Condition = No accommodation

  Provide results for the selected participant group under the 'not accommodated' condition.

- Condition = Accommodation

  Provide results for the selected participant group under the 'accommodated' condition.

> ***REJECTION RULE:*** If reported results provide information at the individual, not aggregate, level, coding will be stopped and the study will not be included in the analysis. For example, if the study has five participants and results are reported for each participant and not at the aggregate level (i.e., across all five participants) it will not be included in the analysis. Add the following note:
>
> 'individual results reported'

> ***REJECTION RULE:*** If reported results only provide 'other' results, coding will be stopped and the study will not be included in the analysis. Add one of the following notes (or a similar note) to the citation:
>
> 'correlational reported'

‘no results reported’

Glossaries

Keyword Glossary

Accommodation

Accommodations provide support for students/students with disabilities and involve adjustments to the assessment setting, timing, scheduling, presentation, or response; accommodations are generally dependant on the disability involved. Accommodations should not provide any advantages to individuals taking the test in question. Test accommodations change in the way a test is administered under standard conditions to facilitate the measurement goals for the assessment (Bolt & Thurlow, 2004).

Words used as synonyms for accommodation (in the context of assessment): modification, adaptation, change, test modifications, test adaptation, or test changes

Questions used to determine if a change to the assessment process is an accommodation or modification are:

1. Will alterations in testing conditions change the skill being measured?
2. Will taking the examination under altered conditions change the meaning of the resulting scores?
3. Would typically developing examinees benefit if allowed the same accommodation?

   Phillips, 1994, p. 104

See Zuriff (2000)

Modification

Modifications (in the context of assessment) change the construct being measured, thus test scores for students taking tests using modifications are considered invalid and student participation is not included in aggregated results for the assessment under consideration

Assessment

Assessment is a multi-stage process involving planning, collecting data, evaluating results and formulating hypotheses, developing recommendations, communicating results and recommendations, conducting re-evaluations, and following up; reference is often made to formative and evaluative assessments (Sattler, 2001)

Words used as synonyms for assessment: test/testing

Test

"… standard procedure for obtaining a sample of behavior from a specified domain" (Crocker & Algina, 1986, p. 4)

Words used as synonyms for test: scale, measure, instrument

Note: The following terms are often used interchangeably: assessment, instrument, measure, test, and scale. It must be noted that testing and assessment are not synonymous as assessment is a process and testing is not (Kubiszyn & Borich, 2003, p. 2). Based on definitions provided by Kubiszyn and Borich (2003) assessment, a multi-stage process, envelops tests and measurement instruments. Scales are viewed as a subset of tests and measures. To avoid confusion when referring to measures and measurement the term measurement instrument has been adopted.

High-stakes assessment

generally refer assessment results tied to important decisions which may significantly impact the lives of students and educational professionals (Reschly, 1993). Statewide assessment programs as part of the accountability structure for states since NCLB (2001) are considered to be high stakes assessments.

Large-scale assessment

Large-scale assessment refers to "… tests are administered to large numbers of students, such as those in a district or state," (Montana Office of Public Instruction, 2001)

Words used as synonyms for large-scale assessment: large-scale testing, large-scale measurement

Standardized assessment

A standardized test, as opposed to a teacher-made test, is designed to be administered and scored under uniform testing conditions (Principles of Educational Measurement, Sax, 1974), has important consequences for the individual examinee, and may be referred to as a high stakes test

Research Participant Information Glossary

Primary data collection

data collection initiated and carried out by the study researcher(s)

Secondary/archival data collection

available data set was collected for a purpose other than the research question posed by the study researcher(s)

Students with disabilities

thirteen legislative special education categories are used to identify students with disabilities; disabilities delineated in IDEA (2004) are

> mental retardation, hearing impairments (including deafness), speech or language impairments, visual impairments (including blindness), serious emotional disturbance (referred to in this title as 'emotional disturbance'), orthopedic impairments, autism, traumatic brain injury, other health impairments, or specific learning disabilities (Part A (SEC. 602) (3) (A) (i), 118 STAT.2652, 2004)

Sampling method

Population

the population contains all individuals within the group under consideration; this is not considered a sampling method

Random sample

random drawing a sample from a population; random samples may be drawn using a numbered list, random number generator, or simply drawing numbers from a hat

Stratified sample

participants are drawn from various strata in the population of subjects being sampled; e.g., 52% of the participants drawn from the population will be female and 48% of the participants drawn from the population will be male; participants may be selected based on random or systematic sampling methods

Systematic sample

participants are drawn from a ‘list’ using a pre-specified method; e.g., every 100th person on a list of 100,000,000 people

Available sample (sample of convenience)

the researcher uses an available pool of research participants; technically, this is not generally considered a sampling method

Assessment Information Glossary

Measurement

operation performed on the physical world by an observer[1] with the assignment of numbers to objects or events according to rules[2] where measurement applies to the properties of said objects and not to the objects themselves[3]. Measurement “of the psychological attribute occurs when a quantitative value is assigned to a behavioral sample collected by using a test” (Crocker & Algina, 1986, p.5)

[1] Weitzenhoffer, 1951

[2] Stevens, 1956

[3] Lord & Novick, 1968; Torgerson, 1958

Measurement Terms

Construct

A psychological characteristic (e.g., numerical ability, spatial ability, introversion, anxiety) considered to vary or differ across individuals. A construct (sometimes called a latent variable) is not directly observable; rather it is a theoretical concept derived from research and other experience that has been constructed to explain observable behavior patterns. When test scores are interpreted by using a construct, the scores are placed in a conceptual framework (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Content domain

A body of knowledge, skills, and abilities defined so that items of knowledge or particular tasks can be clearly identified as included or excluded from the domain (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Criterion

An indicator of the accepted value of outcome performance, such as grade-point average, productivity rate, accident rate, performance rate, absenteeism rate, reject rate and so forth. It is usually a standard against which a predictive measure is evaluated (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Reliability

The degree to which test scores are consistent, dependable, or repeatable, that is, the degree to which they are free of errors of measurement (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Reliability coefficients (Principles of Educational Measurement, Sax, 1974)

- stability: correlation of a set of measurements with themselves over a specified time period (e.g., test-retest)
- equivalence: correlation between score on two or more forms of a test with no time interval between testings (e.g., alternate form)

- stability and equivalence: correlation obtained from testing individuals on two or more forms of a test over specified periods of time
- internal consistency or homogeneity: the extent to which items correlate among themselves

Validity

Validity is an overall evaluative judgment of the degree to which empirical evident and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1995, p. 741)

Types of measurement instruments/tests

Ability test

A test that measure the current performance or estimates future performance of a person in some defined domain of cognitive, psychomotor, or physical functioning (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Achievement test

A test that measure the extent to which a person commands a certain body of information or possesses a certain skill, usually in a field where training or instruction has been received (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999); measures formal or "school taught" learning. Achievement tests measure past performance. Potential synonyms include ability, performance, proficiency, and mastery. An achievement battery is used to assess skills in several different content areas.

Aptitude test

A test that estimates future performance on other tasks not necessarily having evident similarity to the test tasks. Aptitude tests are often aimed at indicating an individual's readiness to learn or to develop proficiency in some particular area if education or training is provided. Aptitude tests sometimes do not differ in form or substance from achievement tests, but may differ in use and interpretation (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999). A cognitive test designed to predict achievement prior to instruction or selection is an example of an aptitude test.

Alternative assessment (also called alternate assessment)

usually designed for a specific subgroup of students; most frequently used to assess students having significant cognitive disabilities who would otherwise not be able to access the assessment, even with accommodations.

Diagnostic test

A test used to point out specific strengths and weaknesses of individuals. Standardized diagnostic tests are available in mathematics and reading (Principles of Educational Measurement, Sax, 1974)

Performance test

A test that requires examinees to perform a task rather than to answer questions. The performance subtests of the WISC include such tasks as assembling objects in puzzle form, etc. (Principles of Educational Measurement, Sax, 1974)

Placement test

A test designed to predict the optimal program or course of study for an individual. For example, a placement test might be used to help a student determine which curriculum is best suited for the student (Principles of Educational Measurement, Sax, 1974)

Readiness test

A test deigned to predict performance, especially in reading or arithmetic; any aptitude measure designed for primary and elementary school children (Principles of Educational Measurement, Sax, 1974)

Screening test

a relatively brief test given to identify students/children who (a) are eligible for certain programs (b) have a disorder or disability needing remediation or rehabilitation (c) require a more comprehensive assessment

Standardized/non-standardized tests

Criterion-referenced test

A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999); test designed to measure content as specified by behavioral objectives/generally, any test having a specified minimum level of attainment and not designed to measure individual differences

Note that criterion-referenced and standards-based tests are terms that, in some cases, are used interchangeably. For coding purposes these two terms will not be used

Domain-referenced test (objectives-referenced test)

A test that allows users to estimate the amount of a specified contain domain that an individual has learned. For example, domains may be based on sets of instructional objectives (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999)

Norm-referenced test

An instrument for which interpretation is based on the comparison of a test taker's performance to the performance of other people in a specified group (Standards for Educational and Psychological Testing, AERA/APA/NCME, 1999); test designed to measure individual differences on some trait or ability

Standards-based test

A test which allows the tester (e.g., states) to incorporate elements of norm-referenced and criterion-referenced testing; standards-based tests are both normed to a reference group and aligned to a set of performance standards

Note that criterion-referenced and standards-based tests are terms that, in some cases, are used interchangeably. For coding purposes these two terms will not be used

Standardized test

A test, carefully prepare over several years, with standardized items and procedures designed to minimize error within the test, error in test administration, and clerical errors in scoring.

Teacher-made test

A test prepared by the teacher for intragroup comparison. If norm-referenced, the test is designed to measure differences among individuals composing the class or group. A criterion-referenced test is a teacher-made test that specifies minimum levels of acceptable performance (Principles of Educational Measurement, Sax, 1974)

Accommodation Information Glossary

Note that the list of accommodations provided, while containing most commonly used assessment accommodations, is not exhaustive.

Assessment accommodation: see *Keyword Glossary*

Assessment modification: see *Keyword Glossary*

Presentation Accommodations

Page layout (for directions/questions/prompts) is different than for test administered without accommodations

Braille edition: page is laid out in Braille only large-type edition/large print: page is laid out using larger font

increase spacing between items: page is laid out with more spacing between each item (e.g., between characters, between words, between each question, etc.)

reduce items/page-line: page is laid out with fewer items on each page, with fewer lines per page

increase size of answer bubbles: item bubbles are larger

reading passages with one complete sentence per line

multiple-choice

answers follow questions down bubbles to right: page is laid out so answer/distractors are below the question and bubbles located to the right of the answer/each distractor

graphic items in the test are given through tactile representation (tactile graphics)

Omit questions which cannot be revised, prorate credit

questions which cannot be changed to accommodate students without modification of the construct being assessed are omitted from the assessment and credit for the question is prorated

Teacher helps student understand prompt

teacher/proctor provides information to help student understand the prompt (answer/distractors) without altering the question construct or modifying the prompts such that it provides an unfair advantage for the student receiving help (makes the answer to the question obvious to the student taking the test)

Student can ask for clarification

student is allowed to ask for clarification of the question

Highlight key words/phrases in directions

key words or phrases in the direction are highlighted (e.g., using color highlighting, **using bold/larger font**)

if the test was not highlighted by the test publisher the teacher/test administrator or student may highlight key words/phrases in the test directions

Simplified language

language used in the instructions/question/prompts (answer and distractors) is simplified without altering the construct being assessed

Oral administration/presentation/read aloud: contents of test are presented in oral format

computer reads paper to student: test is read aloud to the student (directions, questions, and prompts)

prompts available on tape: prompts (answer, distractors) are provided on a tape recorder and presented when the student is ready to answer the question

Interpreter

interpreter is provided to the student to ensure they are able to understand the test content

sign language interpreter: a sign language interpreter is provided to deaf/hard of hearing students

language interpreter: a language interpreter is provided to students whose first language is not the same as the language used on the test. For example, Many states in the United States require that English Language Learners are provided with Language interpreters when they participate in federally-mandated high-stakes assessment. For purposes of the meta-analysis, if research focuses on the use of language interpreters it will not be included.

Verbal encouragement

proctor/teacher provides verbal encouragement to the student while the student is taking the test. It is believed that this type of accommodation provides the student with incentive to continue rather than being discouraged by the perceived difficulty of the test.

Clarify directions

directions may be clarified through restatement (e.g., simplification, paraphrasing) for the student

Provide cues on answer form

additional visual cues are provided for students, such as arrows or stickers

Assistive devices/supports (for directions/questions/prompts)

amanuents/amanuensis (scribe, one who writes from dictation or copies from manuscript, literary assistant)

amplification equipment

equipment that increases the level of sound during the test (e.g. hearing aids)

assistive devices

e.g., speech synthesis

audio-taped administration of sections

auditory amplification device, hearing aid or noise buffers

calculator

standard calculator and special function calculator

dark heavy or raised lines or pencil grips

graphic organizers

graphic organizers created before or during the testing situation

masks or markers to maintain place

questions signed to pupil

questions read aloud to student

e.g., using (1) video (e.g., video cassette), (2) tape-recorder, or (3) computer: questions are read aloud to the student using an assistance device such as a video, tape-recorder or computer

secure papers to work area with tape/magnets

templates

to reduce visible print

to mark location of focus on the test

visual magnification devices

equipment that enlarges the print size of the test

Setting Accommodation

individual administration in a separate location individual assessed separately from other students

small group administration in a separate location

student assessed in small group separate from other students

small group administration using study carrels

student assessed while seated in a study carrel

administer test in location with minimal distractions

student is assessed in a quiet environment

Response Accommodation

Test Format (for responses)

allow student to mark responses in booklet instead of answer sheet

graph paper

increase spacing

paper in alternative format (word processed, Braille, etc.)

wider lines and/or wider margins

Assistive Devices/Supports (for responses)

abacus

alternative response such as oral, sign, typed, pointing

responses may be given by sign language to a sign language interpreter

student points to response and staff member translates this onto an answer sheet

Brailler

device or computer that generates responses in Braille calculator, arithmetic tables

copy assistance between drafts

dictated response

where student provides verbally, response may be tape recorded for later verbatim transcription

interpreter

where interpreter translates response from student; interpreter may be a (1) sign language interpreter for students who are deaf or hard of hearing or a (2) language interpreter for students whose first language is not the language of the test. For purposes of the meta-analysis, if research focuses on the use of language interpreters it will not be included.

large diameter, special grip pencil

proctor/scribe

student responds verbally and a proctor or scribe then translates this to an answer sheet; for writing extended responses, specific instructions about how spelling and punctuation may be included

provide additional examples

slant-board or wedge

spelling dictionary or spell check

spell checker as a separate device or within a word-processing program

tape recorder

word processor

Scheduling/Timing Accommodation

extended time

student may take longer than the time typically allowed.

breaks

time away from test allowed during tests typically administered without breaks, sometimes with conditions about when this can occur (e.g., not within subtests) and how long they can be

time beneficial to student

administered at a time that is most advantageous to the student

multiple sessions

assessments generally given in a single session can be broken into multiple sessions

over multiple days

administered over several days when the assessment is normally administered in one day.

flexible scheduling

the order of subtests may vary from the typical order of subtests

Research Design Information Glossary

Research Design

Comparison

using test accommodation research as an example, researchers examine how the scores for accommodated students with disabilities compare to those of other students with disabilities or those of typically developing students using existing data, generally, a post hoc comparison

Experimental

random assignment of research participants to at least one experimental condition (manipulation of a variable)

Non-experimental

any study that is not an experiment

Quasi-experimental study

an experiment that does not use random assignment of units (research participants)

Research Approach

comparative study

study examining how the scores for accommodated students with disabilities compare to those of other students with disabilities or those of typically developing students

boost study

study examining whether students with disabilities score significantly higher under the accommodated condition; significance of an interaction between disability status and condition is not tested; uses a (i) within subjects (ii) random-independent-groups design; control group not receiving accommodations differential boost study

study where accommodation is expected to boost the scores of students with disabilities significantly more than those of typically developing students (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000)

interaction hypothesis

research, whereby researchers examine the interaction of testing condition and disability status where students with disabilities and typically developing peers are tested under both accommodated and unaccommodated conditions

measurement comparability

study examining tests to determine whether the tests have similar internal characteristics (e.g., factor structure, limited item bias as measured by differential item functioning [DIF]) among accommodated and unaccommodated administrations

Statistical Method Glossary

post hoc test

tests run after the analyses, as a final test; when the overall (omnibus) statistic, such as an F-test, is found to be statistically significant post hoc tests help identify where the significance occurs (e.g., using the Tukey Honestly Significant Difference test analyzing every possible comparison of groups, two at time, to determine which groups are statistically significantly different from one another)

ANOVA

An analysis of variance (ANOVA) is a statistical procedure that compares the amount of between-groups variance in individuals' scores with the amount of within-groups variance (Gall et al., 1996). A general linear model (GLM) univariate procedure, which is more powerful than simple factorial ANOVA, was used (SPSS Base 9.0, 1999).

References

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace Jovanovich College Publishers.

Kubiszyn, T. & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice (7th Edition).* John Wiley & Sons, Inc./Jossey-Bass Publishers, San Francisco, CA.

Montana Office of Public Instruction (2001, August). Release of IOWA test scores, MontCAS memo, p. 5. Retrieved from: www.opi.state.mt.us/PDF/Assessment/MontCas.pdf

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741 - 749.

Sattler, J. (2001). *Assessment of Children: Cognitive applications (4th Edition).* Jerome M. Sattler Publisher, Inc. San Diego, CA.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001 (Technical Report 34)*. Minneapolis, MN: National Center on Educational Outcomes. Retrieved from: http://www.cehd.umn.edu/NCEO/OnlinePubs/Technical34.htm

Thurlow, M., McGrew, K., Tindal, G., Thompson, S., Ysseldyke, J., & Elliott, J. (2000). *Assessment accommodations research: Considerations for design and analysis (NCEO Technical Report 26).* Retrieved from: http://www.cehd.umn.edu/NCEO/OnlinePubs/Technical26.htm

Zenisky, A., & Sireci, S. (2007). *A summary of the research of the effects of test accommodations: 2005 through 2006 (NCEO Technical Report 47).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from: http://www.cehd.umn.edu/nceo/OnlinePubs/Tech47/TechReport47.pdf

Zucker, S. (2003). *Fundamentals of standardized testing.* Pearson Education, Inc, San Antonio, TX.

Appendix E

Test Accommodations Meta-analysis Codebook

Note 1: 0 = N/A, Not Reported, or No for all coding categories
Note 2: Typically developing students are often referred to as students without disabilities in the research literature.

Report Identification

ID code # (start with 01)

- append upper case alpha (A, B, …)
  - for all duplicates (e.g., same study presented in different venue)
- append lower case roman numerals (i, ii, iii, …)
  - for all sub-studies

Research Study Identification (Citation)
Research Study Citation

Author(s) (author's names – last name, first name)
Year of Publication
State/Province of Publication
Country of Publication
Publisher/Publication Type (see below for codes)
Publisher/Publication Source (see below for codes)

Publisher: (i) Publication Type – numeric (ii) Publication Source - alpha

(0) uncategorized
(1) journal
 (A) Applied Measurement in Education
 (B) American Educational Research Journal
 (C) BC Journal of Special Education
 (D) British Journal of Special Services
 (E) Diagnostique
 (F) Educational Assessment
 (G) Educational and Psychological Measurement
 (H) Educational Measurement: Issues and Practice
 (I) Educational Psychologist
 (J) Educational Psychology
 (K) Educational Researcher
 (L) Exceptional Children
 (M) Exceptionality
 (N) Journal of Educational Measurement
 (O) Journal of Learning Disabilities
 (P) Journal of Psychoeducational Assessment
 (Q) Journal of School Psychology
 (R) Journal of Special Education
 (S) Learning Disability Quarterly
 (T) Practical Assessment, Research, and Evaluation
 (U) RE:view
 (V) Remedial and Special Education
 (W) Review of Educational Research
 (X) School Psychology Review
 (Y) other (type in name of journal)
(2) conference proceedings (paper)

- (A) American Educational Research Association
- (B) Council for Chief State School Officers Large-Scale Assessment Conference
- (C) National Council on Measurement in Education
- (D) National Association of School Psychologists
- (E) other (type in name of conference paper was presented at)

(3) organization (report)
- (A) National Center on Educational Outcomes (NCEO) http://www.cehd.umn.edu/nceo/
- (B) Wisconsin Center for Education Research (WCER) http://www.wcer.wisc.edu/
- (C) Center for Research on Evaluation, Standards, and Student Testing (CRESST) http://www.cse.ucla.edu/
- (D) College Entrance Examination Board (College Board) http://www.collegeboard.com/
- (E) Behavioral Research and Teaching at the University of Oregon (BRT) http://www.brtprojects.org/
- (F) National Assessment Governing Board (NAGB) http://www.nagb.org/
- (G) Center for the Study of Assessment Validity and Evaluation (CSAVE) (http://www.c-save.umd.edu/index.html)
- (H) Educational Policy Analysis Achives (EPAA) http://epaa.asu.edu
- (I) Fraiser Institute (http://www.fraserinstitute.org/)
- (J) Council of Chief State School Officers (CCSS) http://www.ccsso.org/
- (K) American Institutes for Research (AIR) http://www.air.org/
- (L) National Center for Education Statistics (NCES) http://nces.ed.gov/
- (M) other (type in name of organization & abbreviation)

(4) dissertation
- (A) Dissertation Abstracts International (DAI)
- (B) Dissertation Abstracts Online (DAO)
- (C) Dissertations & Theses @ University of Denver
- (D) ProQuest Dissertations & Theses (PQDT)
- (E) UMI Dissertations & Theses

(5) manuscript (unpublished)

(6) other (type in name of category of publication)

Study Retrieval
Method to locate study

(0) n/a

(1) review articles (e.g., research synthesis, review of the literature, meta-analysis, …)
- (A) Academic Search Complete
- (B) Applied Social Sciences Index and Abstracts (ASSIA)
- (C) British Periodicals
- (D) Dissertations & Theses @ University of Denver
- (E) ERIC
- (F) ERIC web portal @www.eric.ed.gov
- (G) Google Scholar
- (H) JSTOR
- (I) ProQuest Dissertations & Theses (PQDT)
- (J) ProQuest Education Journals
- (K) PsycINFO
- (L) PsycARTICLES
- (M) Sociological Abstracts

(2) references in eligible studies (e.g., bibliographic reference)
- (A) Academic Search Complete
- (B) Applied Social Sciences Index and Abstracts (ASSIA)

(C) British Periodicals
(D) Dissertations & Theses @ University of Denver
(E) ERIC
(F) ERIC web portal @www.eric.ed.gov
(G) Google Scholar
(H) JSTOR
(I) ProQuest Dissertations & Theses (PQDT)
(J) ProQuest Education Journals
(K) PsycINFO
(L) PsycARTICLES
(M) Sociological Abstracts

(3) computerized bibliographic database search
(A) Academic Search Complete
(B) Applied Social Sciences Index and Abstracts (ASSIA)
(C) British Periodicals
(D) Dissertations & Theses @ University of Denver
(E) ERIC
(F) ERIC web portal @www.eric.ed.gov
(G) Google Scholar
(H) JSTOR
(I) ProQuest Dissertations & Theses (PQDT)
(J) ProQuest Education Journals
(K) PsycINFO
(L) PsycARTICLES
(M) Sociological Abstracts

(4) organizational web site search
(A) National Center on Educational Outcomes (NCEO) http://www.cehd.umn.edu/nceo/
(B) Wisconsin Center for Education Research (WCER) http://www.wcer.wisc.edu/
(C) Center for Research on Evaluation, Standards, and Student Testing (CRESST) http://www.cse.ucla.edu/
(D) College Entrance Examination Board (College Board) http://www.collegeboard.com/
(E) Behavioral Research and Teaching at the University of Oregon (BRT) http://www.brtprojects.org/
(F) National Assessment Governing Board (NAGB) http://www.nagb.org/
(G) Center for the Study of Assessment Validity and Evaluation (CSAVE) (http://www.c-save.umd.edu/index.html)
(H) Educational Policy Analysis Achives (EPAA) http://epaa.asu.edu
(I) Fraiser Institute (http://www.fraserinstitute.org/)
(J) Council of Chief State School Officers (CCSS) http://www.ccsso.org/
(K) American Institutes for Research (AIR) http://www.air.org/
(L) National Center for Education Statistics (NCES) http://nces.ed.gov/
(M) other (type in name of organization & abbreviation)

Research Quality

Proxy for study quality
(0) not reported
(1) Peer-reviewed
(2) Published dissertation
(3) Not peer-reviewed
(4) Unpublished dissertation

Research Participant Information

Participant Data Source

(0) not reported
(1) Primary data collection
(2) Secondary/archival data collection

Participant Sampling Method

(0) not reported
(1) population
    (A) federal population
    (B) state population
    (C) school district population
    (D) local (school) population
(2) simple random selection
    (A) federal population
    (B) state population
    (C) school district population
    (D) local (school) population
(3) stratified random selection
    (A) federal population
    (B) state population
    (C) school district population
    (D) local (school) population
(4) systematic selection
    (A) federal population
    (B) state population
    (C) school district population
    (D) local (school) population
(5) available (sample of convenience)
    (A) federal population
    (B) state population
    (C) school district population
    (D) local (school) population

Participant Sampling Method (additional information)

<u>(list information regarding sample participants (e.g., district – sample of convenience – “2 schools from the district participated”)</u>

Participant Grade Level(s)

(0) not reported
(1) Prekindergarten
(2) Kindergarten
(3) Grade 1
(4) Grade 2
(5) Grade 3
(6) Grade 4
(7) Grade 5
(8) Grade 6
(9) Grade 7
(10) Grade 8
(11) Grade 9
(12) Grade 10

(13) Grade 11
(14) Grade 12
(15) College/University Undergraduate (list college level if available)
(16) Adult

Participant Sample Composition (number)

> ***Note 1:*** if the sample composition was not reported or there were no students with disabilities participating in the study, coding will be stopped and the study will not be included in the analysis.
>
> ***Note 2:*** do ***not*** break out individual disability groups ***unless*** they are used as individual groups in the data analysis.

(0) not reported
(1) students with disabilities (list total number of participants w/ disabilities)
(2) students without disabilities (list total number of participants w/o disabilities)
(3) unclassified

Total number of participants (final sample) (list number of participants in final sample)

> ***Note:*** The total number of participants is based on the number of participants completing the study, broken out by group (i.e., students who took the test(s) whose data is included in the final analysis/analyses under investigation). The number of research participants listed in the participant demographics section of a study may not be the same as the final number of participants used in the research analysis/analyses.

Participant Disability Classification
(number of participants by disability, using Special Education Taxonomy for disability classification)
(0) not reported
(1) visually impaired
(2) hearing impaired
(3) cognitively impaired
(4) physically/orthopedically impaired
(5) speech or language impaired/communication disability
(6) seriously emotionally disturbed/emotional or behavioral disability
(7) autistic
(8) traumatically brain injured
(9) other health impairments
(10) specific learning disability
    (A) reading
    (B) math
    (C) reading & math
    (D) other
    (E) not classified (select if LD students are 'undifferentiated')
(11) other disability (list category of disability)
    (A) representative sample (homogeneous)
    (B) not representative sample (heterogeneous)

Assessment Information
Assessment Citation
Name of Assessment (list name, if not listed use 'state' or similar name)
Version of Assessment (list version of the assessment, ***if provided***)

Author(s) (list authors, ***if provided***)
Publisher (list category, ***if provided***)
Date of Publication (list category, ***if provided***)

Assessment Classification

Note: The rule of thumb for state tests – select '(D) standards-based' unless the study specifically refers to a different category

Type of Assessment

(0) not reported
(1) standardized/published
(A) norm-referenced *
(B) criterion-referenced
(C) domain-referenced
(D) standards-based
(E) curriculum-based
(2) state
(A) norm-referenced
(B) criterion-referenced *
(C) domain-referenced
(D) standards-based *
(E) curriculum-based (e.g., aligned to state curriculum)
(3) researcher or professionally developed (for research purposes)
(A) not reported
(B) not based on state or standardized assessment
(C) based on state or standardized assessment *
(4) other (list category for assessment)
* more commonly found

Assessment Descriptors

(0) not reported
(1) standardized assessment
(2) large-scale assessment
(3) high-stakes assessment
(4) large-scale and high-stakes assessment
(5) other (list category for assessment descriptor)

Assessment Categorization

(0) not reported
(1) achievement test
(2) aptitude test
(3) performance test
(4) placement test
(5) selection test
(6) screening test
(7) diagnostic assessment
(8) other (list category for assessment classification)

Assessment Content / Construct

(0) not reported
(1) mathematics
(2) reading/language arts
(3) science
(4) writing

(5) social studies
(6) physical education
(7) multiple content areas <u>(list content areas by numeric separated by commas (e.g., use **1, 3, 4** for mathematics, science, writing)</u>
(8) not specified/no specific content area
(9) cognition (e.g., intelligence assessment)
(10) psychomotor skills
(11) personality
(12) affect
(13) interest
(14) other <u>(list category for assessment content / construct)</u>

Assessment Format
(0) not reported
(1) multiple choice
(2) fill in the blanks
(3) short answer questions (constructed responses)
(4) open-ended (long answer) questions
(5) mixture <u>(list mix of formats by numeric separated by commas (e.g., use **1, 2** for multiple choice, short answer question)</u>
(6) other <u>(list category for assessment format)</u>

Number of Assessment Forms
(0) not reported
(1) 1 form
(2) 2 forms
(3) multiple forms (>2 forms)

Assessment Reliability reported?
(0) no
(1) yes
(2) published test/can find online

Reliability Type:

Note: Use multiple fields if more than one type of reliability is reported

(0) not reported
(1) coefficient of stability (test-retest)
(2) coefficient of equivalence (alternate form)
(3) coefficient of stability and equivalence
(4) internal consistency or homogeneity
   (A) Cronbach's alpha
   (B) Spearman rho
   (C) Split-half
(5) criterion reliability
(6) other <u>(list category for type of reliability)</u>

Reliability Index (value) <u>(list value)</u>

Note: if more than one reliability index, list separately using the 'Reliability Type' codes and Reliability Index (value).

Assessment Validity reported?

Note: Use multiple fields if more than one type of validity is reported

(0) no

- (1) yes
- (2) published test/can find online

Validity Type:

- (0) not reported
- (1) Cronbach's alpha
- (2) Spearman rho
- (3) Split-half
- (4) Factor Analysis
- (5) Correlational (e.g., with other published test measuring the same construct/content)
- (6) other (list category for type of validity)

Validity Index (value) (list value)

> ***Note:*** if more than one validity index, list separately using the 'Validity Type' codes and Validity Index (value).

Accommodation Information

- (0) not reported
- (1) Presentation Accommodation
  - (A) Presentation
    - (i) page layout (for directions/questions/prompts)
      - (1) Braille edition
      - (2) large-type edition/large print
      - (3) increase spacing between items
      - (4) reduce items/page-line
      - (5) increase size of answer bubbles
      - (6) reading passages with one complete sentence/line
      - (7) multiple-choice, answers follow questions down bubbles to right
      - (8) other (list other page layout)
    - (ii) omit questions which cannot be revised, prorate credit
    - (iii) teacher helps student understand prompt
    - (iv) student can ask for clarification
    - (v) highlight key words/phrases in directions
    - (vi) simplified language
    - (vii) oral administration/presentation/read-aloud (reads 'entire' test)
      - (1) computer reads paper to student
      - (2) prompts available on tape
      - (3) other (list other oral administration)
    - (viii) cueing
    - (ix) interpreter
      - (1) sign language interpreter
      - (2) language interpreter
    - (x) verbal encouragement
    - (xi) other (list other presentation accommodation)
  - (B) Test directions
    - (i) typewriter
    - (ii) dictation to a proctor/scribe
    - (iii) communication device
    - (iv) signing directions to students (sign language interpreter)
    - (v) simplify language in directions or problems
    - (vi) page layout (for directions)
      - (1) highlight verbs in instructions by underlining
    - (vii) clarify directions

(viii) provide cues on answer form
(ix) oral administration/presentation
(1) read directions to students
(2) reread (repeat) directions (e.g., for each page of questions)
(x) other (list other test directions accommodation)
(C) Assistive devices/supports (for directions/questions/prompts)
(i) visual magnification devices
(ii) templates to reduce visible print
(iii) auditory amplification device, hearing aid or noise buffers
(iv) audio-taped administration of sections
(v) secure papers to work area with tape/magnets
(vi) questions read-aloud to student
(1) video (e.g., video cassette)
(2) tape-recorder
(3) computer (e.g., computer-read text)
(4) other (list method used to read questions aloud)
(vii) masks or markers to maintain place
(viii) questions signed to pupil
(ix) dark heavy or raised lines or pencil grips
(x) assistive devices – speech synthesis
(xi) amanuents/amanuensis (scribe, one who writes from dictation or copies from manuscript, literary assistant)
(xii) other (list other assistive device/supports accommodation)
(2) Setting Accommodation
(A) individual administration in a separate location
(B) small group administration in a separate location
(C) small group administration using study carrels
(D) provide adaptive or special furniture
(E) administer test in location with minimal distractions
(F) provide special acoustics
(G) other (list other setting accommodation)
(3) Timing/Scheduling Accommodation
(A) use of flexible schedule
(B) allow frequent breaks during testing
(C) extend the time allotted to complete the test
(D) administer the test in several sessions, specify duration
(E) provide special lighting
(F) time of day
(G) administer the test over several days, specify duration
(H) other (list other timing/scheduling accommodation)
(4) Response Accommodation
(A) Test Format (for responses)
(i) increase spacing
(ii) wider lines and/or wider margins
(iii) graph paper
(iv) paper in alternative format (word processed, Braille, etc.)
(v) allow student to mark responses in booklet instead of answer sheet
(vi) other (list other test format accommodation)
(B) Assistive Devices/Supports (for responses)
(i) word processor
(ii) calculator, arithmetic tables
(iii) spelling dictionary or spell check
(iv) alternative response such as oral, sign, typed, pointing

(v) Brailler
(vi) large diameter, special grip pencil
(vii) copy assistance between drafts
(viii) slant-board or wedge
(ix) tape recorder
(x) abacus
(xi) provide additional examples
(xii) dictated response (e.g., scribe)
(1) student tapes response for later verbatim transcription
(xiii) interpreter
(1) sign language interpreter
(2) language interpreter
(xiv) other (list other assistive devices/supports accommodation)

(5) Multiple Accommodations/Accommodation Packages – DISCONTINUE CODING
(6) Other (list category for accommodation)

Research Study Design Information
Methodology

Note: Research Study Design is broken into Methodology/Study Type, Methodology/Research Approach, Methodology/Design, and Accommodation Order

Study Type
(0) not reported
(1) Post hoc
(examines existing database; comparison between groups without random assignment)
(2) Quasi-Experiment
('experimental'; group comparison without random assignment)
(3) Experiment
(experiment; group comparison with random assignment)

Use the following for Research Design and Research Design Variation
Research Approach – numeric (e.g., 1, 2,)
Research Design – alpha (e.g., A, B, …)
Research Design Variation – numeric (e.g., i, ii, …)

The following abbreviations are used for Research Design Variation:

Abbreviation	Meaning	Subscripting
swd	students with disabilities	subscripted to represent different groups of students with disabilities (#)
sw/od	students without disabilities	subscripted to represent different groups of students without disabilities (#)
accomm	condition = accommodated	
n/accomm	condition = not accommodated	
swd	students with disabilities	
A	test form A	
B	test form B	

Research Approach/Research Design/Research Design Variation
(0) not reported
(1) Comparison

(2) Boost

(A) Repeated Measures
(e.g., pre- & post-assessment using the same group of participants)

(i) Variation 1

Time 1		**Time 2**	
swd$_1$	n/accomm	swd$_1$	accomm

(ii) Variation 2

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (A)

(iii) Variation 3

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (A)
swd$_2$	n/accomm (A)	swd$_2$	accomm (A)

(iv) Variation 4

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	n/accomm (A)	swd$_2$	accomm (B)

(v) Variation 5

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)

(vi) Variation 6

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (A)
swd$_2$	accomm (A)	swd$_2$	accomm (A)

(vii) Variation 7

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	accomm (A)	swd$_2$	accomm (B)

(viii) Variation 8

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	n/accomm (B)	swd$_2$	accomm (A)

(ix) Variation 9

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (A)
swd$_2$	accomm (A)	swd$_2$	n/accomm (A)

(x) Variation 10

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	accomm (A)	swd$_2$	n/accomm (B)

(xi) Variation 11

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	accomm (B)	swd$_2$	n/accomm (A)

(xii) Variation 12

Time 1		**Time 2**	
swd$_1$	accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	accomm (B)	swd$_2$	n/accomm (A)
swd$_3$	n/accomm (A)	swd$_3$	accomm (B)
swd$_4$	n/accomm (B)	swd$_4$	accomm (A)

(xiii) Variation 13

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (A)
swd$_2$	n/accomm (B)	swd$_2$	n/accomm (B)
swd$_3$	accomm (A)	swd$_3$	accomm (A)
swd$_4$	accomm (B)	swd$_4$	accomm (B)

(xiv) Variation 14

Time 1		**Time 2**	
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	n/accomm (B)	swd$_2$	n/accomm (A)
swd$_3$	accomm (A)	swd$_3$	accomm (B)
swd$_4$	accomm (B)	swd$_4$	accomm (A)

(B) Independent Groups (matched)

(i) Variation 1

Group		**Group**	
swd$_1$	accomm	swd$_2$ or sw/od$_1$	accomm

(ii) Variation 2

Group		**Group**	
swd$_1$	accomm	swd$_2$ or sw/od$_1$	accomm

(iii) Variation 3

Group		**Group**	
swd$_1$	n/accomm	swd$_2$	accomm

(iv) Variation 4

Group		**Group**	
swd$_1$	n/accomm (A)	swd$_2$	accomm (A)

(v) Variation 5

Group		**Group**	
swd$_1$	n/accomm (A)	swd$_2$	accomm (B)

(3) Boost/Differential Boost

Select Design Variation code from Boost or Differential Boost, dependent upon research question and data structure

(4) Differential Boost

(A) Repeated Measures

(e.g., pre- & post-assessment using the same group of participants)

(i) Variation 1

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (A)
swd$_1$	n/accomm (A)	swd$_1$	accomm (A)

(ii) Variation 2

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)

(iii) Variation 3

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	n/accomm (A)
sw/od$_2$	accomm (A)	sw/od$_2$	accomm (A)
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (A)
swd$_2$	accomm (A)	swd$_2$	accomm (A)

(iv) Variation 4

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	n/accomm (B)
sw/od$_2$	n/accomm (A)	sw/od$_2$	accomm (B)
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	n/accomm (A)	swd$_2$	accomm (B)

(v) Variation 5

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	n/accomm (B)
sw/od$_2$	accomm (A)	sw/od$_2$	accomm (B)
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	accomm (A)	swd$_2$	accomm (B)

(vi) Variation 6

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	n/accomm (A)
sw/od$_2$	n/accomm (B)	sw/od$_2$	n/accomm (B)
sw/od$_3$	accomm (A)	sw/od$_3$	accomm (A)
sw/od$_4$	accomm (B)	sw/od$_4$	accomm (B)
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (A)
swd$_2$	n/accomm (B)	swd$_2$	n/accomm (B)
swd$_3$	accomm (A)	swd$_3$	accomm (A)
swd$_4$	accomm (B)	swd$_4$	accomm (B)

(vii) Variation 7

Time 1		Time 2	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (A)
sw/od$_2$	accomm (A)	sw/od$_2$	n/accomm

			(A)
swd$_1$	n/accomm (A)	swd$_1$	accomm (A)
swd$_2$	accomm (A)	swd$_2$	n/accomm (A)

(viii) Variation 8

Time 1		**Time 2**	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
sw/od$_2$	accomm (A)	sw/od$_2$	n/acomm (B)
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	accomm (A)	swd$_2$	n/accomm (B)

(ix) Variation 9

Time 1		**Time 2**	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
sw/od$_2$	n/accomm (B)	sw/od$_2$	accomm (A)
swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	n/accomm (B)	swd$_2$	accomm (A)

(x) Variation 10

Time 1		**Time 2**	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
sw/od$_2$	n/accomm (B)	sw/od$_2$	accomm (A)
sw/od$_3$	accomm (A)	sw/od$_3$	n/accomm (B)
sw/od$_4$	accomm (B)	sw/od$_4$	n/accomm (A)
swd$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
swd$_2$	n/accomm (B)	sw/od$_2$	accomm (A)
swd$_3$	accomm (A)	sw/od$_3$	n/accomm (B)
swd$_4$	accomm (B)	sw/od$_4$	n/accomm (A)

(xi) Variation 11

Time 1		**Time 2**	
sw/od$_1$	n/accomm (A)	sw/od$_1$	n/accomm (B)
sw/od$_2$	n/accomm (B)	sw/od$_2$	n/accomm (A)
sw/od$_3$	accomm (A)	sw/od$_3$	accomm (B)
sw/od$_4$	accomm (B)	sw/od$_4$	accomm (A)
swd$_1$	n/accomm (A)	swd$_1$	n/accomm (B)
swd$_2$	n/accomm (B)	swd$_2$	n/accomm (A)
swd$_3$	accomm (A)	swd$_3$	accomm (B)
swd$_4$	accomm (B)	swd$_4$	accomm (A)

(xii) Variation 12

Time 1		**Time 2**	
sw/od$_1$	n/accomm (A)	sw/od$_1$	accomm (B)
sw/od$_2$	accomm (B)	sw/od$_2$	n/accomm (A)

swd$_1$	n/accomm (A)	swd$_1$	accomm (B)
swd$_2$	accomm (B)	swd$_2$	n/accomm (A)

(B) Independent Groups (matched)

(i) Variation 1

Group		Group	
sw/od$_1$	n/accomm (A)	sw/od$_2$	accomm (A)
swd$_1$	n/accomm (A)	swd$_2$	accomm (A)

(ii) Variation 2

Group		Group	
sw/od1	n/accomm (A)	sw/od$_2$	accomm (B)
swd1	n/accomm (A)	swd$_2$	accomm (B)

(5) Independent Groups (not matched)
Use the ‘matched’ independent group design variation codes for ‘not matched’

Accommodation Order

(0) not reported
(1) Not accommodated – Accommodated
(2) Accommodated – Not Accommodated
(3) Counter-balanced
(4) n/a (e.g., matched @ student/class/school/district/state level so order is not a consideration; used a covariate to make ‘equivalent’ so matching is not necessary)

Statistical Method

(0) Descriptive Statistics (mean, s.d., n)
(1) t-test
(2) F-test
(3) chi-square
(4) ANOVA
(5) ANCOVA (use adjusted means)
(6) Multiple Regression (use unstandardized regression coefficient, β)
(7) Proportions (frequencies)
(8) Other <u>(list other statistical method used)</u>

Results Information
(Results information is reported by group – students with disabilities, students without disabilities)

Note: Use information in the Research Study Design Information section to determine which groups were included in the research study and have results information. If the research study only provides ‘other’ results, coding will be stopped and the study will not be included in the analysis.

Participant Assignment

(0) not reported
(1) not random assignment
(2) all conditions and/or all forms
(A) school level
(B) class level

(C) student level
(3) random assignment
(A) school level
(B) class level
(C) student level

Condition = No accommodation
Enter information for the appropriate group(s) studied selecting from the following:

- Group (students with disabilities)
- Group (students without disabilities)

Enter type of statistic with values for appropriate group(s). For example:
((list statistic)) (list value)
((list statistic)) (list value)
n / df (list value)
((list statistic)) (list value)

Note: Use last 'statistic for correlation between not accommodated & accommodated conditions with pre- post design (e.g., repeated measures)

Condition = Accommodation
Enter information for the appropriate group(s) studied selecting from the following:

- Group (students with disabilities)
- Group (students without disabilities)

Enter type of statistic with values for appropriate group(s). For example:
((list statistic)) (list value)
((list statistic)) (list value)
n / df (list value)
((list statistic)) (list value)

Note: Use last 'statistic for correlation between not accommodated & accommodated conditions with pre- post design (e.g., repeated measures)

Appendix F

Test Accommodations Meta-analysis Coding Form

Note 1: 0 = N/A, Not Reported, or No for all coding categories
Note 2: Typically developing students are often referred to as students without disabilities in the research literature.

Report Identification
ID code #: ____________________

Research Study Identification (Citation)
Research Study Citation
Author(s) ____
Year of Publication ____
State/Province of Publication ____
Country of Publication ____
*Publisher/*Publication Type ____
*Publisher/*Publication Source ____

Study Retrieval Information
Method to locate study ____

Research Question(s)/Research Purpose
Note: Include page #, question/paragraph #, first 3 to 4 words of question

Research question(s) selected
____

Research Quality (proxy)
Research Quality ____

Research Participant Information
Participant data source _________________________________
Participant sampling method _________________________________
Participant sampling method (additional) ______________________________
Participant grade level(s) _________________________________
Total number of research participants _________________________________

Participant Sample Composition
Note 1: Sample size is for analysis(es) run may be different from the original participant sample size, enter sample size for analysis(es) run

Note 2: if the sample composition was not reported or there were no students with disabilities participating in the study, coding will be stopped and the study will not be included in the analysis

Students with disabilities (n) ____
Students without disabilities (n) ____
Unclassified students (n) ____

Participant Disability Classification

Note: enter number of participants by disability used in analysis(es) in the study (use Special Education Taxonomy for disability classification)

Participant Disability Classification 1 ________________________________
Disability Classification 1 (n) ________________________________
Participant Disability Classification 2 ________________________________
Disability Classification 2 (n) ________________________________

Assessment (Measure) Information

Assessment Citation

Name of Assessment ____
Version(s) of Assessment ____
Author(s) ____
Publisher ____
Date of Publication ____

Assessment Classification

Type of Assessment ____
Assessment Descriptors ____
Assessment Categorization ____
Assessment Content/Construct ____
Assessment Format ____

Assessment Reliability reported? (0) no (1) yes

Reliability Type ______________________________________
Reliability Index (value) ______________________________________

Assessment Validity reported? (0) no (1) yes

Validity Type ______________________________________
Validity Index (value) ______________________________________

Accommodation Information ______________________________________

Research Study Design Information

Note: See Codebook for information on Research Approach Variations

Methodology/Study Type ______________________________
Methodology/Research Approach ______________________________
Methodology/Research Design ______________________________
Methodology/Research Design Variation ______________________________
Accommodation Order ______________________________

Statistical Method (select one)

Note 1: if 'Other' was selected, coding will be stopped and the study will not be included in the analysis

Note 2: if data for the statistical method is not available make a note of the author(s) name(s) and contact information and discontinue coding. If the information can be tracked down the study will be included in the analysis, otherwise the study will be dropped the analysis

Statistical Method ______________________________

Results

Note 1: For statistic enter the type of statistic (e.g., mean) and the value (e.g., 14.01) in the space provided

Note 2: If correlation coefficient for testing condition between time 1 and time 2, or group 1 and group 2, is available enter information the last line for statistic (e.g., correlation) and value (e.g., 0.75)

Participant Assignment ______________________________

Condition = No accommodation

Group (students with disabilities)
(statistic/value) ______________________________
(statistic/value) ______________________________
n/df ______________________________
(statistic/value) ______________________________

Group (students without disabilities)
(statistic/value) ______________________________
(statistic/value) ______________________________
n/df ______________________________
(statistic/value) ______________________________

Condition = Accommodation

Group (students with disabilities)
(statistic/value) ______________________________

(statistic/value) ______________________________

n/df ______________________________

(statistic/value) ______________________________

Group (students without disabilities)

(statistic/value) ______________________________

(statistic/value) ______________________________

n / df ______________________________

(statistic/value) ______________________________

Appendix G

Keyword Search Terms

A sequence of search terms was used within each research database. Wildcarding was used to ensure maximal coverage during the search process. Date criteria–1999 through 2011–were used to limit the searches to dates matching the meta-analysis inclusion criteria.

The following search criteria were used:

- assess*and accomm*
- assess*and accomm* and disabil*
- high-stake* and accomm*
- high-stake* and accomm* and disabil*
- large-scale* and accomm*
- large-scale* and accomm* and disabil*
- standard*and accomm*
- test*and accomm*
- test*and accomm* and disabil*

Where wildcard search terms were equivalent to:

- accomm* (also = accommodate, accommodated, accommodates, accommodating, accommodation, accommodations)
- assess* (also = assessed, assesses, assessing, assessment, assessments)
- disabil* (also = disability, disable, disabled, disables, disabilities)
- high-stake* (also = high-stakes)

- large-scale* (also = large-scaled)
- standard* (also = standards, standardized)
- test* (also = tested, testing, tests)

Databases searched included:

- Academic Search Complete
- Applied Social Sciences Index and Abstracts (ASSIA)
- British Periodicals
- Dissertations & Theses @ University of Denver
- ERIC
- Google Scholar
- JSTOR
- ProQuest Dissertations & Theses (PQDT)
- ProQuest Education Journals
- PsycINFO
- PsycARTICLES
- Sociological Abstracts

Number of potentially eligible studies found = 242

Note: This is the total number of studies found and includes duplicates that were not deleted until after the primary research studies were evaluated.

Appendix H

Citations for Duplicate and Excluded Studies

Citations for Duplicate Studies

Abedi, J., Kao, J. C., Leon, S., Sullivan, L., Herman, J. L., Pope, R., Nambiar, V., & Mastergeorge, A.M. (2008). *Exploring factors that affect the accessibility of reading comprehension assessments for students with disabilities: A study of segmented text (CRESST Report 746).* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/R746.pdf
Duplicate of Abedi et al., 2010 (1)
Decision: duplicate – exclude study/do not count as separate study

Cahalan-Laitusis, C. (2006). *Impact of read aloud on test of reading comprehension (An examination of the validity of a read aloud accommodation for a standardized reading assessment using differential boost and predictive validity as criteria).* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
Duplicate of Laitusis, 2010 (19)
Decision: duplicate – exclude study/do not count as separate study

Cahalan-Laitusis, C., Cook, L., Cline, F., & King, T. (2006). *Examining differential boost from read aloud on a test of reading comprehension at grades 4 and 8.* Paper presented at the annual meeting of the Council for Exceptional Children, Salt Lake City, UT.
Duplicate of Laitusis, 2010 (19)
Decision: duplicate – exclude study/do not count as separate study

Elliott, S. N., & Marquart, A. M. (2003). *Extended time as an accommodation on a standardized mathematics test: An investigation of its effects on scores and perceived consequences for students with varying mathematical skills.* Madison, WI: University of Wisconsin- Madison, Wisconsin Center for Education Research. Retrieved from http://www.wcer.wisc.edu/publications/workingPapers/Working_Paper_No_2003_1.pdf
Duplicate of Marquart, 2000 (24)
Decision: duplicate – exclude study/do not count as separate study

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children, 70*(3), 349–367.
Duplicate of Marquart, 2000 (24)
Decision: duplicate – exclude study/do not count as separate study

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2000). *An oral versus standard administration of a large-scale mathematics test (Attachment 7).* Dover, DE: Delaware Department of Education. Retrieved from http://www.doe.k12.de.us/aab/Report_and_documents/ICAS.shtml
Duplicate of Helwig et al., 2002 (12)
Decision: duplicate – exclude study/do not count as separate study

Huesman, R. L., & Frisbie, D. (2000). *The validity of ITBS reading comprehension test scores for learning disabled and non–learning disabled students under extended-time conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
Duplicate of Huesman, 1999 (14)
Decision: duplicate – exclude study/do not count as separate study

MacArthur, C. A., & Cavalier, A. R. (2000). *Dictation and speech recognition technology as accommodations in large-scale assessments for students with learning disabilities (Attachment 11).* Dover, DE: Delaware Department of Education. Retrieved from http://www.doe.k12.de.us/aab/Report_and_documents/ICAS.shtml
Duplicate of MacArthur & Cavalier, 2004 (23)
Decision: duplicate – exclude study/do not count as separate study

Marquart, A. M. (2000). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels.* Paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird, UT.
Duplicate of Marquart, 2000 (24)
Decision: duplicate – exclude study/do not count as separate study

Meloy, L., Deville, C., & Frisbie, D. (2000). *The effects of a reading accommodation on standardized test scores of learning disabled and non learning disabled students.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA
Duplicate of Meloy et al., 2002 (26)
Decision: duplicate – exclude study/do not count as separate study

Citations for Excluded Studies

Dropped during initial review of primary studies

Beddow, P. A. (2011). *Effects of testing accommodations and item modifications on students' performance: An experimental investigation of test accessibility strategies.* (Doctoral Dissertation, Vanderbilt University, 2011). (AAT 3479839)

Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education, 17*(1), 25 - 37.

Bruins, S. K. (2006). *Investigating how students with disabilities receiving special education services affect the school's ability to meet adequate yearly progress. .* (Doctoral Dissertation, University of Idaho, 2006) (AAT 3185562)

Corn, A. L., Wall, R. S., Jose, R.T., Bell, J. K., Wilcox, K., & Perez, A. (2002). An initial study of reading and comprehension rates for students who received optical devices. *Journal of Visual Impairment and Blindness, 96*, 322 - 334.

DiCerbo, K., Stanley, E., Roberts, M., & Blanchard, J. (2001). *Attention and standardized reading test performance: Implications for accommodation.* Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.

DiRosa, F. (2007). *The impact of testing accommodations on individual postsecondary student test outcomes.* (Doctoral Dissertation, Temple University, 2007). Retrieved from Proquest Digital Dissertations. (AAT 3268142)

Elliott, S., Kratochwill, T., & McKevitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology, 39*(1)*,* 3 - 24.

Elliott, S. N., Kratochwill, T. R., McKevitt, B. C., & Malecki, C. K. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly, 24*(4), 224 - 239.

Feldman, E.; Kim, J.; & Elliott, S. N. (2011). The effects of accommodations on adolescents' self-efficacy and test performance. *Journal of Special Education, 45*(2), 77 - 88.

Fletcher, J. M., Francis, D. J., Boudousquie, A. & Copeland, K. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children, 72*(2), 136 - 150.

Hall, S. E. H. (2002). *The impact of test accommodations on the performance of students with disabilities.* (Doctoral Dissertation, The George Washington University, 2002). Retrieved from Proquest Digital Dissertations. (ATT 3045478)

Hanson, K., Brown, B., Levine, R., & Garcia, T. (2001). Should standard calculators be provided in testing situations? An investigation of performance and preference differences. *Applied Measurement in Education, 14* (1), 59-72.

Harris, L. W. (2008). *Comparison of student performance between teacher read and CD-ROM delivered modes of test administration of English language arts tests.* (Doctoral Dissertation, University of South Carolina, 2008). Retrieved from Proquest Digital Dissertations. (AAT 3321402)

Jackson, L. M. (2003). *The effects of testing adaptations on students' standardized test scores for students with visual impairments in Arizona.* (Doctoral Dissertation, The University of Arizona, 2003). Retrieved from Proquest Digital Dissertations. (AAT 3108915)

Jones, A. K. F. (2006). *The effects of accommodations on standardized test scores in mathematics for students with special needs and English language learners.* (Doctoral Dissertation, University of California, Irvine and University of California, Los Angeles, 2006). Retrieved from Proquest Digital Dissertations. (AAT 3214047)

Kettler, R. J., Niebling, B. C., Mroch, A. A. Feldman, E. S., Newell, M. L., Elliott, S. N., Kratochwill, T., & Bolt, D. M. (2005). Effects of testing accommodations on math and reading scores: An experimental analysis of the performance of students with and without disabilities. *Assessment for Effective Intervention [Special issue: Testing Accommodations: Research to Guide Practice], 31*(1), 37 – 48.

Koretz, D., & Hamilton, L. (1999). Assessing students with disabilities in Kentucky: *The effects of accommodations, format, and subject (CSE Technical Report 498).* Los Angeles, CA: Center for Research on Standards and Student testing.

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

McKevitt, B. C. (2001). *The effects and consequences of using testing accommodations on a standardized reading test.* (Doctoral Dissertation, The University of Wisconsin - Madison, 2001). Retrieved from Proquest Digital Dissertations. (ATT 3020768)

Mandinach, E. B., Bridgeman, B., Cahalan-Laitusis, C., & Trapani, C. (2005). T*he impact of extended time on SAT test performance. Research Report No 2005-8.* New York, NY: The College Board. Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-20.pdf

Schulte, A. A. Gilbertson, (2001). *Experimental analysis of the effects of testing accommodations on the students' standardized mathematics test scores.* (Doctoral Dissertation, The University of Wisconsin - Madison, 2001). Retrieved from Proquest Digital Dissertations. (ATT 9982260)

Sharoni, V., & Vogel, G. (2007). Entrance test accommodations, admission and enrollment of students with learning disabilities in teacher training colleges in Israel. *Assessment & Evaluation in Higher Education, 32*(3), 255 - 270.

Stirling, I. R. (2008). *The use of accommodations for special education students and the reliability of test score achieved.* (Doctoral Dissertation, California State University Dominguez Hills, 2008). Retrieved from Proquest Digital Dissertations. (ATT 1461594)

Trammell, J. K. (2003). The impact of academic accommodations on final grades in a postsecondary setting. *Journal of College Reading and Learning, 34*(1), 76 - 90.

Wainer, H., Bridgeman, B., Najarian, M., & Trapani, C. (2004). How much does extra time on the SAT help? *Chance, 17*(2), 19 - 24.

Zurcher, R. (1999). *The effects of testing accommodations on the admissions test scores of students with learning disabilities.* (Doctoral Dissertation, The University of Texas at Austin, 1999). Retrieved from Proquest Digital Dissertations. (ATT 9947446)

Zentall, S. S., Grskovic, J. A., Javorsky, J., & Hall, A. M. (2000). Effects of noninformational color on the reading test performance of students with and without attentional deficits. *Diagnostique, 25*(2), 129 – 46.

Dropped During Coding (data not useable)

Antalek, E. E. (2005). *The relationships between specific learning disability attributes and written language: A study of the performance of learning disabled high school subjects completing the TOWL-3.* (Doctoral Dissertation, Clark University, 2005). Retrieved from Proquest Digital Dissertations. (ATT 3154958)

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement, 41,* 291 - 310.

Burch, M. (2002). *Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities.* (Doctoral dissertation, Vanderbilt University, 2002). Retrieved from Proquest Digital Dissertations. (ATT 3047429)

Huynh, H., & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education, 19*(1), 21 - 39.

Huynh, H., Meyer, J. P., & Gallant, D. J. (2004). Comparability of student performance between regular and oral administrations for a high-stakes mathematics test. *Applied Measurement in Education, 17,* 39 - 57.

Jerome, M. K. (2007). The state of accommodations for fifth grade students with disabilities on the Virginia SOL reading, writing, and math tests. (Doctoral Dissertation, George Mason University, 2007). Retrieved from Proquest Digital Dissertations. (ATT 3289706)

Maihoff, N. A. (2000). *The effects of administering an ASL signed standardized test via DVD/television and by paper-and-pencil: A pilot study. (Attachment 9).* Dover, DE: Delaware Department of Education. Retrieved from http://www.doe.state.de.us/aab/dstp

Ricketts, C., Brice, J., & Coombes, L. (2010). Are multiple choice tests fair to medical students with specific learning disabilities? *Advances in Health Sciences Education, 15*(2), 265 - 275.

Szarko, J. (2000). *Familiar versus unfamiliar examiners: The effects on test performance and behaviors of children with autism and elated developmental disabilities.* (Doctoral Dissertation, The Pennsylvania State University, 2000). Retrieved from Proquest Digital Dissertations. (ATT 9966904)

Dropped During Coding (missing data for effect size calculation)

Bouck, E. C. (2009). Calculating the value of graphing calculators for seventh-grade students with and without disabilities: A pilot study. *Remedial and Special Education, 30*(4), 207 - 215.

Bouck, E. C., & Bouck, M. K. (2008). Does it add up? Calculators as accommodations for sixth grade students with disabilities. *Journal of Special Education Technology, 23*(2), 17 – 32.

Kappel, A.T. (2002). The effect of testing accommodations on subtypes of students with learning disabilities. (Doctoral Dissertation, University of Pittsburg, 2002). Retrieved from Proquest Digital Dissertations. (ATT 3054293)

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000). *Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test (Technical Report 28).* Minneapolis, MN: University of

Minnesota, National Center on Educational Outcomes. Retrieved from http://education.umn.edu/NCEO/OnlinePubs/Technical28.htm

Dropped During Coding (decided not to include post hoc studies)

Baker, J. S. (2006). *Effect of extended time testing accommodations on grade point averages of college students with learning disabilities.* (Doctoral Dissertation, Capella University, 2000). Retrieved from Proquest Digital Dissertations. (ATT 3205714)

Elliott, J., Bielinski, J., Thurlow, M.L., DeVito, P., & Hedlund, E. (1999). *Accommodations and the performance of all students on Rhode Island's performance assessment (Rhode Island Assessment Report 1).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.eric.ed.gov/PDFS/ED440516.pdf

Koretz, D., & Hamilton, L. (2001). *The performance of students with disabilities on New York's Revised Regents Comprehensive Examination in English (CSE Technical Report 540).* Los Angeles, CA: Center for the Study of Evaluation. Retrieved from http://www.cse.ucla.edu/products/reports/TR540.pdf

Dropped during outlier analysis

Bouck, E. C., & Yadav, A. (2008). Assessing calculators as assessment accommodations for students with disabilities. *Assistive Technology Outcomes and Benefits, 5*(1), 19 - 28.

Lewandowski, L. J., Lovett, B. J., & Rogers, C. L. (2008). Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships? *Journal of Psychoeducational Assessment, 26*, 315 - 324.

Duplicate of dropped studies

Duplicate of Fletcher et al (2006) – dropped primary study

Fletcher, J. M., Francis, D. J., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C. J., Kalinowski, S., Young, V., & Vaughn, S. (2009). Effects of a bundled accommodations package on high-stakes testing for middle school students with reading disabilities. *Exceptional Children, 75*(4), 447 - 463.

Duplicate of Koretz & Hamilton (1999) – dropped primary study

Koretz, D., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22*(3), 255 – 272.

Duplicate of McKevitt (2001) dissertation – dropped primary study

McKevitt, B. C. (2000, June). *The use and effects of testing accommodations on math and science performance assessments.* Paper presented at the annual large-scale assessment conference of the Council for Chief State School Officers, Snowbird, UT.

McKevitt, B. C., & Elliott, S. N. (2003). Effects and perceived consequences of using read-aloud and teacher-recommended accommodations on a reading achievement test. *School Psychology Review, 32*(4), 583 – 600.

McKevitt, B. C., Marquart, A. M., Mroch, A., Gilbertson Schulte, A., Elliott, S., & Kratochwill, T. (2000). *The use and effects of testing accommodations on math and science performance assessments.* Paper presented at the annual meeting of the National Association of School Psychologists, New Orleans, LA.

McKevitt, B. C., Marquart, A. M., Mroch, A., Schulte, A., Elliott, S., & Kratochwill, T. (1999). *Test accommodations for students with disabilities: An empirical analysis.* Poster presented at the annual meeting of the American Psychological Association, Boston, MA.

McKevitt, B. C., Marquart, A. M., Mroch, A., Schulte, A., Elliott, S., & Kratochwill, T. (2000, June). *Understanding the effects of testing accommodations: a single-case approach.* Paper presented at the annual meeting of the CSSO Large-Scale Assessment Conferences, Snowbird, UT.

Duplicate of Schulte (2000) dissertation – dropped primary study
Schulte, A. A., Elliott, S. N., & Kratochwill, T. R. (2001). *Experimental analysis of the effects of testing accommodations on students' standardized mathematics test scores.* Paper presented at the Council of Chief State School Officers Conference Snowbird, Utah.

Schulte, A. A., Elliott, S. N., & Kratochwill, T. R. (2001). Effects of testing accommodations on students' standardized mathematics test scores: An experimental analysis. *School Psychology Review, 30*(4), 527 - 547.

Duplicate of Zurcher (1999) – dropped primary study
Zurcher, R., & Bryant, D. P. (2001). The validity and comparability of entrance examination scores after accommodations are made for students with LD. *Journal of Learning Disabilities, 34*(5), 462 - 471.

Appendix I

Citations for Irretrievable Studies

Bielinski, J. (2001). *Evaluating the effect of read-aloud accommodation on multiple-choice reading tests and math items.* Paper presented at the National Council on Measurement in Education, Seattle, WA:

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

Bolt, S. E., & Diao, Q. (2005, May). *Reading aloud a reading test: Examining reading sub-skill performance.* Paper presented at the American Psychological Association Conference, Washington, DC.

DiCerbo, K., Stanley, E., Roberts, M., & Blanchard, J. (2001). A*ttention and standardized reading test performance: Implications for accommodation.* Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.

Elliott, S. N., & Roach, A. T. (2002, April). *The impact of providing testing accommodations to students with disabilities.* Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research and Department of Educational Psychology. Retrieved from http://www.wcer.wisc.edu/testacc

Huynh, H., Meyer, J. P., & Gallant-Taylor, D. (2002). *Comparability of scores of accommodated and non-accommodated testings for a high school exit examination of mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

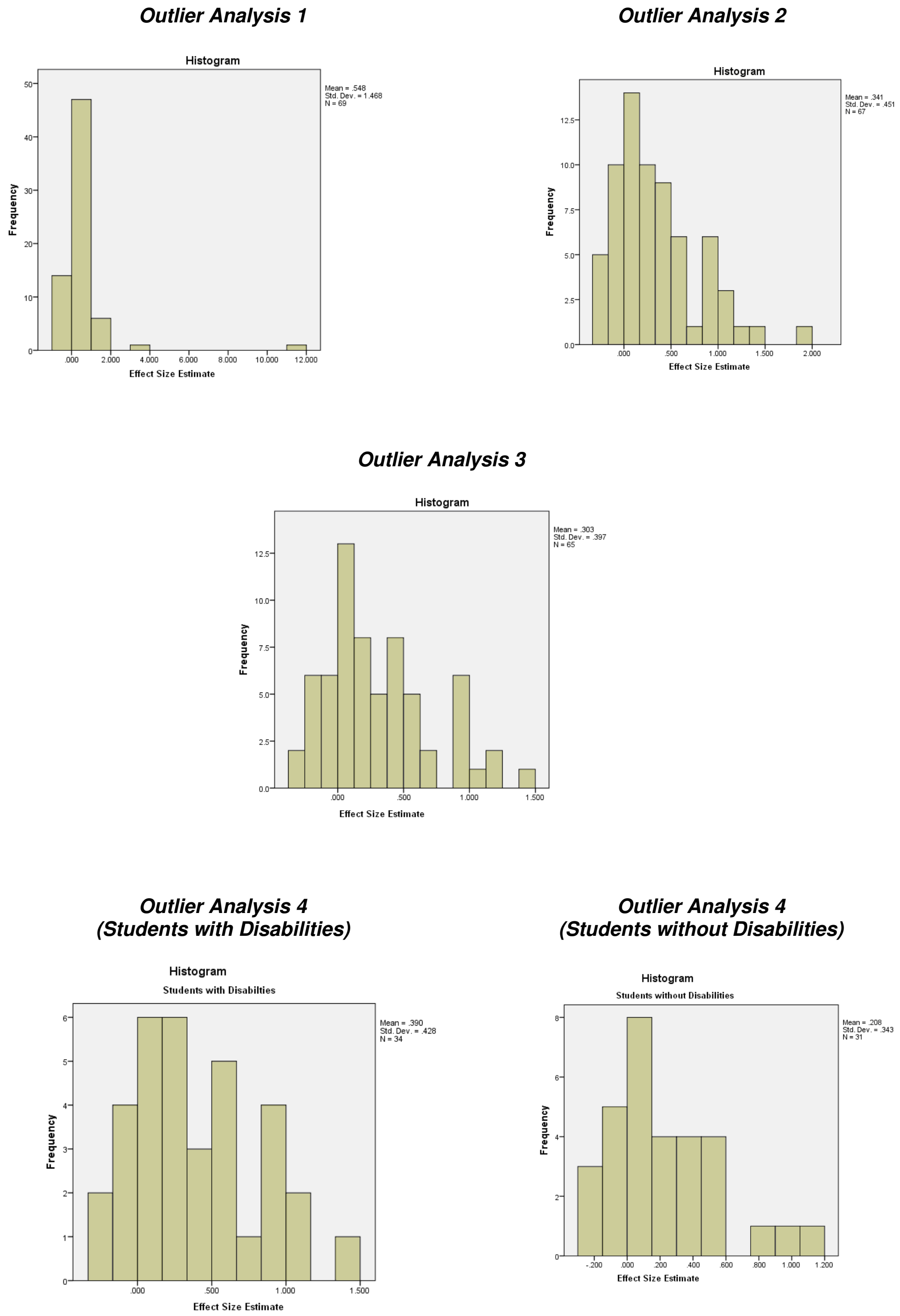
Weston, T. J. (1999). *The validity of oral accommodation in testing (NCES 200306).* Paper presented at the annual meeting of the American Education Research Association. Montreal, QC.

May be duplicate of Weston, 2000 (34)

Note: These studies may have been rejected during the review or coding phase and will not be counted as part of the total number of studies found

Appendix J

Outlier Analyses



Outlier Analysis 1

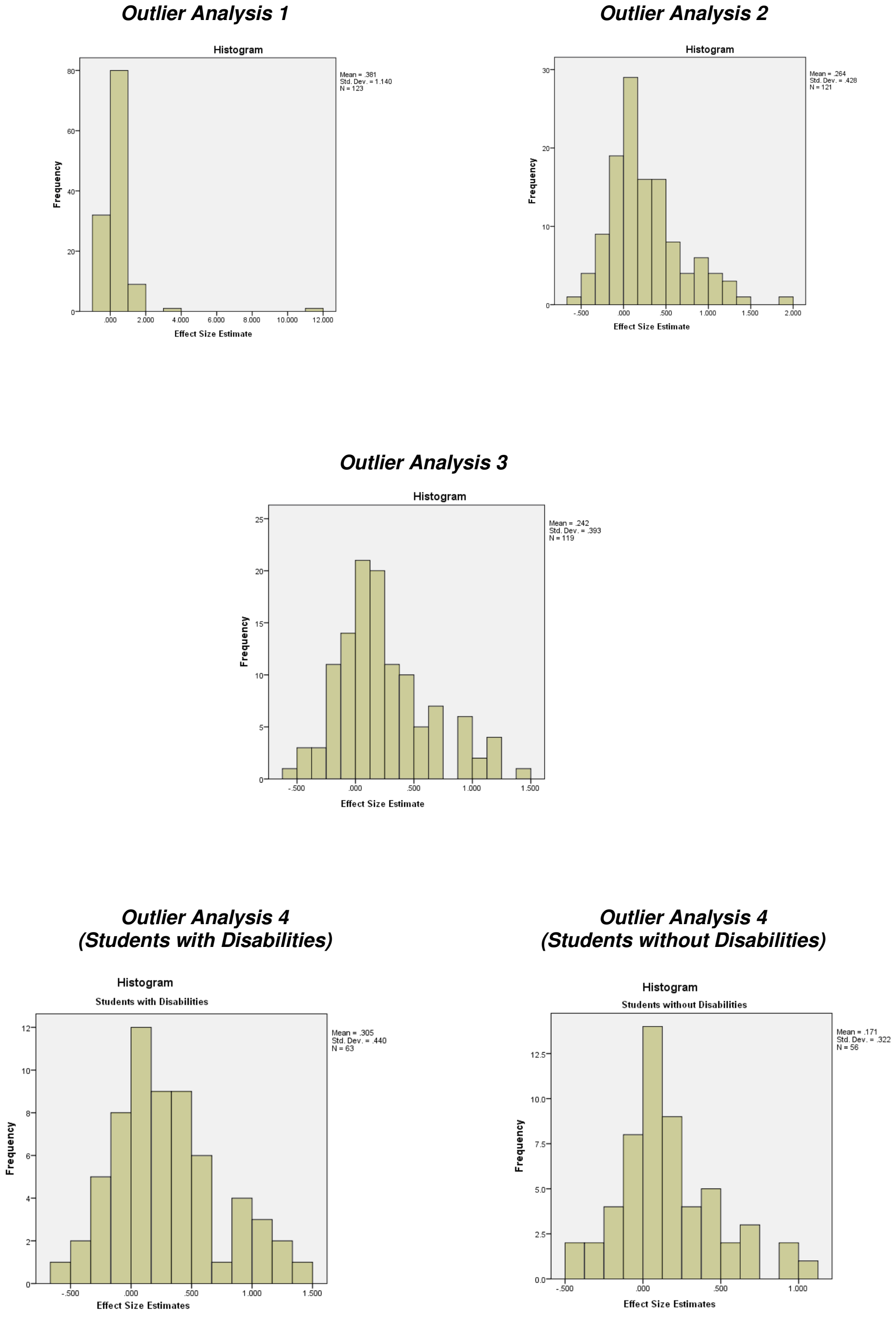
Outlier Analysis 2

Outlier Analysis 3

Outlier Analysis 4
(Students with Disabilities)

Outlier Analysis 4
(Students without Disabilities)

Figure 15: Histograms for Study as the Unit of Analysis


Outlier Analysis 1

Outlier Analysis 2

Outlier Analysis 3

Outlier Analysis 4
(Students with Disabilities)

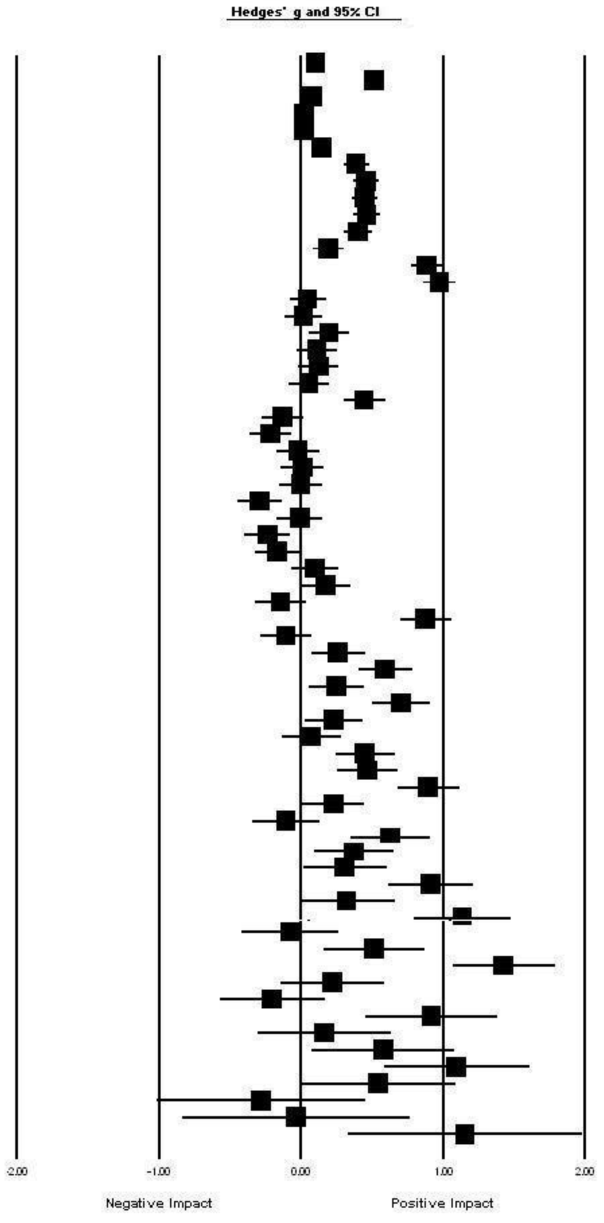
Outlier Analysis 4
(Students without Disabilities)

Figure 16: Histograms for Substudy as the Unit of Analysis

Appendix K

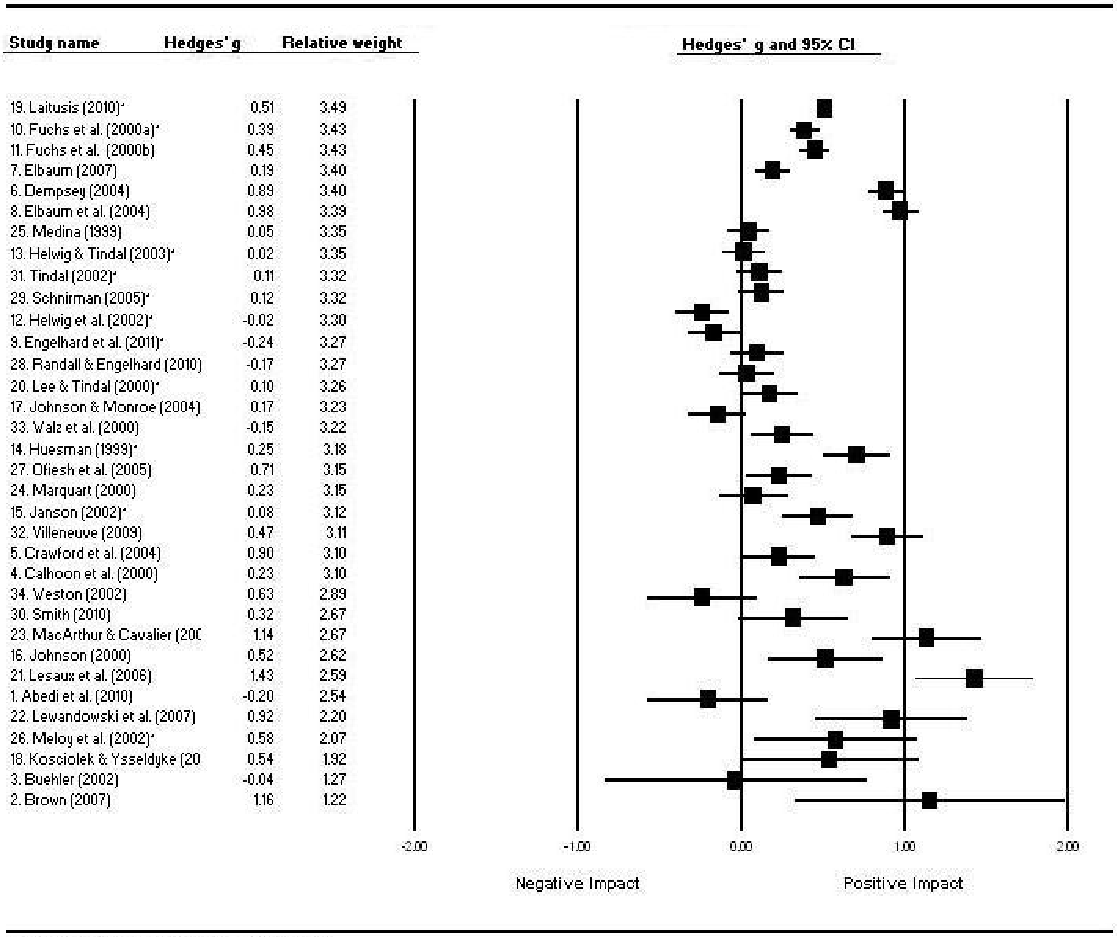
Publication Bias Analysis – Effect Sizes by Weights

Study name	Group	Hedges' g	Relative weight
19. Laitusis (2010)*	swod	0.10	1.86
19. Laitusis (2010)*	swd	0.51	1.86
25. Medina (1999)	swod	0.08	1.86
31. Tindal (2002)*	swod	0.02	1.84
13. Helwig & Tindal (2003)*	swod	0.03	1.83
14. Huesman (1999)*	swod	0.15	1.83
10. Fuchs et al. (2000a)*	swd	0.39	1.81
11. Fuchs et al. (2000b)	swod	0.46	1.81
11. Fuchs et al. (2000b)	swd	0.45	1.81
10. Fuchs et al. (2000a)*	swod	0.46	1.80
5. Crawford et al. (2004)	swod	0.40	1.80
7. Elbaum (2007)	swd	0.19	1.79
6. Dempsey (2004)	swd	0.89	1.79
8. Elbaum et al. (2004)	swd	0.98	1.78
25. Medina (1999)	swd	0.05	1.75
13. Helwig & Tindal (2003)*	swd	0.02	1.75
24. Marquart (2000)	swod	0.20	1.74
31. Tindal (2002)*	swd	0.11	1.73
29. Schnirman (2005)*	swd	0.12	1.73
20. Lee & Tindal (2000)*	swod	0.05	1.73
7. Elbaum (2007)	swod	0.45	1.73
9. Engelhard et al. (2011)*	swod	-0.13	1.72
28. Randall & Engelhard (2010)*	swod	-0.22	1.72
12. Helwig et al. (2002)*	swd	-0.02	1.72
12. Helwig et al. (2002)*	swod	0.01	1.71
30. Smith (2010)	swod	0.00	1.71
33. Walz et al. (2000)	swod	-0.29	1.70
1. Abedi et al. (2010)	swod	-0.01	1.70
9. Engelhard et al. (2011)*	swd	-0.24	1.70
28. Randall & Engelhard (2010)*	swd	-0.17	1.70
20. Lee & Tindal (2000)*	swd	0.10	1.69
17. Johnson & Monroe (2004)	swd	0.17	1.67
33. Walz et al. (2000)	swd	-0.15	1.66
8. Elbaum et al. (2004)	swod	0.88	1.66
17. Johnson & Monroe (2004)	swod	-0.11	1.66
27. Ofiesh et al. (2005)	swod	0.26	1.64
2. Brown (2007)	swod	0.59	1.64
14. Huesman (1999)*	swd	0.25	1.63
27. Ofiesh et al. (2005)	swd	0.71	1.61
24. Marquart (2000)	swd	0.23	1.61
15. Janson (2002)*	swd	0.08	1.60
32. Villeneuve (2009)	swod	0.45	1.59
32. Villeneuve (2009)	swd	0.47	1.58
5. Crawford et al. (2004)	swd	0.90	1.58
4. Calhoon et al. (2000)	swd	0.23	1.58
29. Schnirman (2005)*	swod	-0.11	1.54
34. Weston (2002)	swd	0.63	1.44
26. Meloy et al. (2002)*	swod	0.37	1.44
34. Weston (2002)	swod	0.31	1.40
21. Lesaux et al. (2006)	swod	0.91	1.39
30. Smith (2010)	swd	0.32	1.30
23. MacArthur & Cavalier (2004)	swd	1.14	1.30
16. Johnson (2000)	swod	-0.08	1.29
16. Johnson (2000)	swd	0.52	1.27
21. Lesaux et al. (2006)	swd	1.43	1.25
23. MacArthur & Cavalier (2004)	swod	0.22	1.24
1. Abedi et al. (2010)	swd	-0.20	1.22
22. Lewandowski et al. (2007)	swd	0.92	1.02
18. Kosciolek & Ysseldyke (2000)	swod	0.16	1.02
26. Meloy et al. (2002)*	swd	0.58	0.95
22. Lewandowski et al. (2007)	swod	1.10	0.93
18. Kosciolek & Ysseldyke (2000)	swd	0.54	0.87
3. Buehler (2002)	swod	-0.28	0.61
3. Buehler (2002)	swd	-0.04	0.54
2. Brown (2007)	swd	1.16	0.52



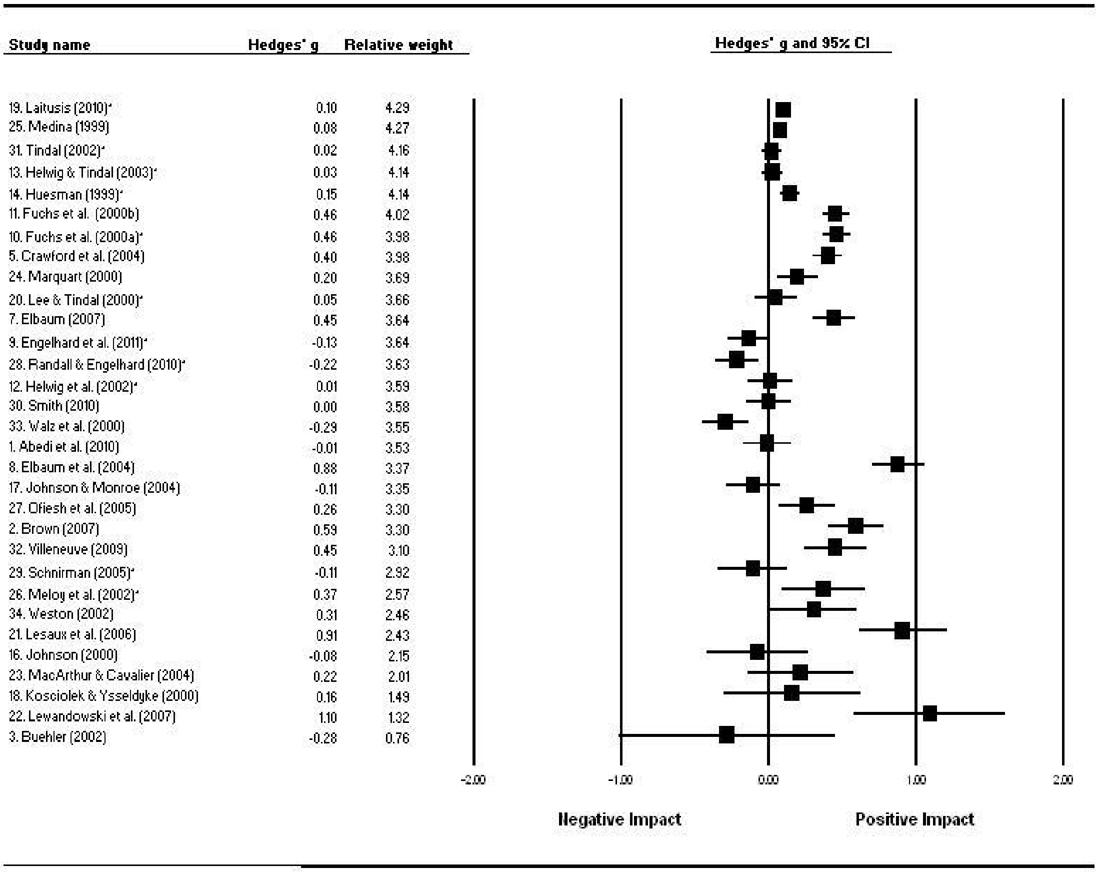
* effect size computed used combined substudies

Figure 17: Effect Size Estimates by Weights - Study Level (all data)



* effect size computed used combined substudies

Figure 18: Effect Size Estimates by Weights - Study Level (students with disabilities)

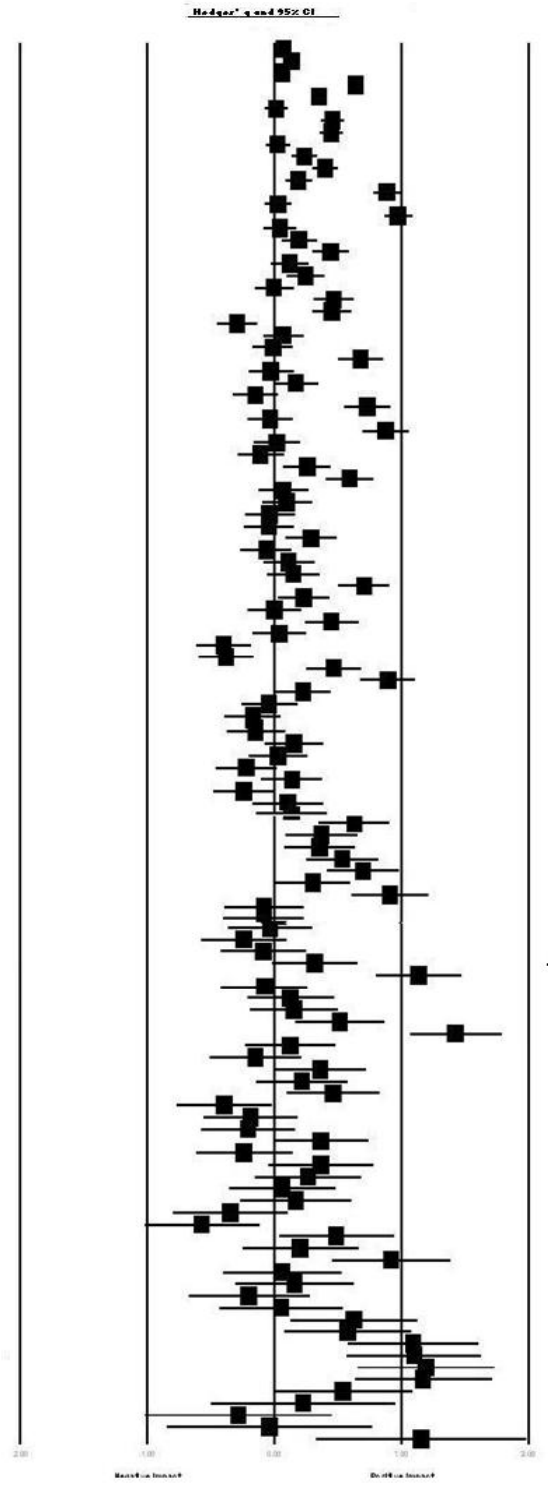


Study name	Hedges' g	Relative weight
19. Laitusis (2010)*	0.10	4.29
25. Medina (1999)	0.08	4.27
31. Tindal (2002)*	0.02	4.16
13. Helwig & Tindal (2003)*	0.03	4.14
14. Huesman (1999)*	0.15	4.14
11. Fuchs et al. (2000b)	0.46	4.02
10. Fuchs et al. (2000a)*	0.46	3.98
5. Crawford et al. (2004)	0.40	3.98
24. Marquart (2000)	0.20	3.69
20. Lee & Tindal (2000)*	0.05	3.66
7. Elbaum (2007)	0.45	3.64
9. Engelhard et al. (2011)*	-0.13	3.64
28. Randall & Engelhard (2010)*	-0.22	3.63
12. Helwig et al. (2002)*	0.01	3.59
30. Smith (2010)	0.00	3.58
33. Walz et al. (2000)	-0.29	3.55
1. Abedi et al. (2010)	-0.01	3.53
8. Elbaum et al. (2004)	0.88	3.37
17. Johnson & Monroe (2004)	-0.11	3.35
27. Ofiesh et al. (2005)	0.26	3.30
2. Brown (2007)	0.59	3.30
32. Villeneuve (2009)	0.45	3.10
29. Schnirman (2005)*	-0.11	2.92
26. Meloy et al. (2002)*	0.37	2.57
34. Weston (2002)	0.31	2.46
21. Lesaux et al. (2006)	0.91	2.43
16. Johnson (2000)	-0.08	2.15
23. MacArthur & Cavalier (2004)	0.22	2.01
18. Kosciolek & Ysseldyke (2000)	0.16	1.49
22. Lewandowski et al. (2007)	1.10	1.32
3. Buehler (2002)	-0.28	0.76

* effect size computed used combined substudies

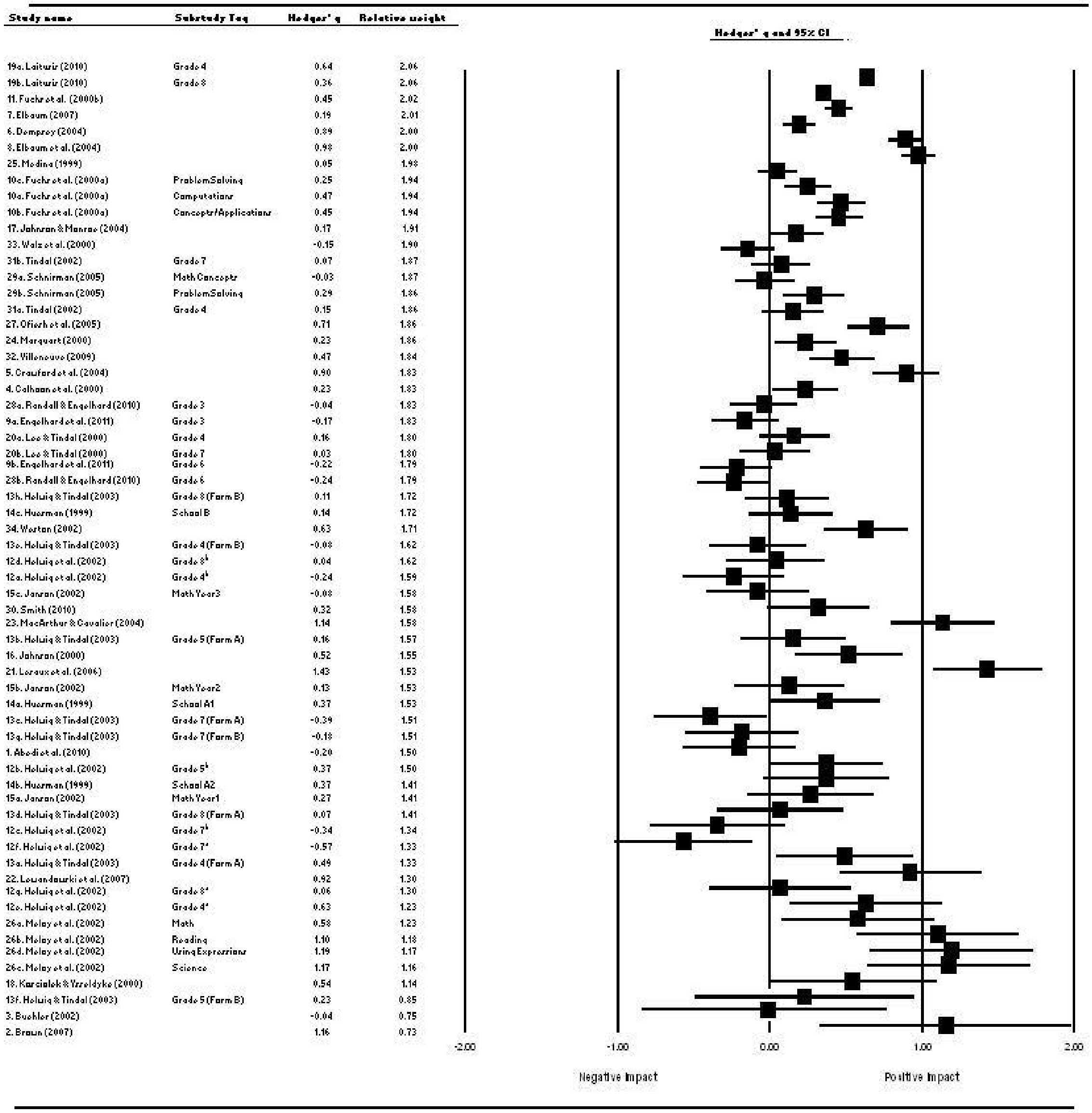
Figure 19: Effect Size Estimates by Weights - Study Level (typically developing students)

Study name	Substudy Tag	Group	Hedges' g	Relative weight
25. Medina (1999)		nswd	0.00	1.08
19a. Laitusis (2010)	Grade 4	nswd	0.14	1.08
19b. Laitusis (2010))	Grade 8	nswd	0.06	1.07
19a. Laitusis (2010)	Grade 4	swd	0.64	1.07
19b. Laitusis (2010)	Grade 8	swd	0.56	1.07
31a. Tindal (2002)	Grade 4	nswd	0.02	1.05
11. Fuchs et al. (2000b)		nswd	0.46	1.05
11. Fuchs et al. (2000b)		swd	0.45	1.05
31b. Tindal (2002)	Grade 7	nswd	0.03	1.05
14a. Harmon (1999)	School A1	nswd	0.24	1.05
5. Crawford et al. (2004)		nswd	0.40	1.04
7. Elbaum (2007)		swd	0.19	1.04
6. Dempsey (2004)		swd	0.03	1.04
14c. Harmon (1999)	School B	nswd	0.03	1.04
8. Elbaum et al. (2004)		swd	0.30	1.04
25. Medina (1999)		swd	0.05	1.02
24. Marquart (2000)		nswd	0.20	1.01
7. Elbaum (2007)		nswd	0.45	1.01
19c. Helwig & Tindal (2003)	Grade 4 (Form B)	nswd	0.12	1.00
18c. Fuchs et al. (2000a)	ProblemSolving	swd	0.25	1.00
30. Smith (2010)		nswd	0.00	1.00
18a. Fuchs et al. (2000a)	Computations	swd	0.47	0.99
18b. Fuchs et al. (2000a)	Concepts/Applications	swd	0.45	0.99
33. Walz et al. (2000)		nswd	-0.29	0.99
18c. Fuchs et al. (2000a)	ProblemSolving	nswd	0.00	0.99
1. Abedi et al. (2010)		nswd	-0.01	0.99
18b. Fuchs et al. (2000a)	Concepts/Applications	nswd	0.60	0.97
19b. Helwig & Tindal (2003)	Grade 5 (Form A)	nswd	-0.02	0.97
17. Johnson & Monroe (2004)		swd	0.17	0.97
33. Walz et al. (2000)		swd	-0.15	0.97
18a. Fuchs et al. (2000a)	Computations	nswd	0.79	0.97
19a. Helwig & Tindal (2003)	Grade 4 (Form A)	nswd	-0.03	0.97
8. Elbaum et al. (2004)		nswd	0.00	0.97
19e. Helwig & Tindal (2003)	Grade 7 (Form A)	nswd	0.02	0.97
17. Johnson & Monroe (2004)		nswd	-0.11	0.97
27. Ofiesh et al. (2005)		nswd	0.26	0.96
2. Brown (2007)		nswd	0.59	0.96
31b. Tindal (2002)	Grade 7	swd	0.07	0.95
28a. Lee & Tindal (2000)	Grade 4	nswd	0.10	0.95
29a. Schnirman (2005)	Math Concepts	swd	-0.03	0.95
19g. Helwig & Tindal (2003)	Grade 7 (Form B)	nswd	-0.04	0.95
29b. Schnirman (2005)	ProblemSolving	swd	0.29	0.95
20b. Randall & Engelhard (2010)	Grade 6	nswd	-0.06	0.95
9a. Engelhard et al. (2011)	Grade 3	nswd	0.12	0.94
31a. Tindal (2002)	Grade 4	swd	0.15	0.94
27. Ofiesh et al. (2005)		swd	0.71	0.94
24. Marquart (2000)		swd	0.29	0.94
28b. Lee & Tindal (2000)	Grade 7	nswd	0.00	0.93
32. Villeneuve (2003)		nswd	0.45	0.93
19d. Helwig & Tindal (2003)	Grade 8 (Form A)	nswd	0.04	0.93
9b. Engelhard et al. (2011)	Grade 6	nswd	-0.39	0.93
20a. Randall & Engelhard (2010)	Grade 3	nswd	-0.30	0.93
32. Villeneuve (2003)		swd	0.47	0.93
5. Crawford et al. (2004)		swd	0.30	0.93
4. Calhoon et al. (2000)		swd	0.29	0.92
20a. Randall & Engelhard (2010)	Grade 3	swd	-0.04	0.92
9a. Engelhard et al. (2011)	Grade 3	swd	-0.17	0.92
19h. Helwig & Tindal (2003)	Grade 8 (Form B)	nswd	-0.14	0.91
28b. Lee & Tindal (2000)	Grade 7	nswd	0.00	0.93
32. Villeneuve (2003)		nswd	0.45	0.93
19d. Helwig & Tindal (2003)	Grade 8 (Form A)	nswd	0.04	0.93
9b. Engelhard et al. (2011)	Grade 6	nswd	-0.39	0.93
20a. Randall & Engelhard (2010)	Grade 3	nswd	-0.30	0.93
32. Villeneuve (2003)		swd	0.47	0.93
5. Crawford et al. (2004)		swd	0.30	0.93
4. Calhoon et al. (2000)		swd	0.29	0.92
20a. Randall & Engelhard (2010)	Grade 3	swd	-0.04	0.92
9a. Engelhard et al. (2011)	Grade 3	swd	-0.17	0.92
19h. Helwig & Tindal (2003)	Grade 8 (Form B)	nswd	-0.14	0.91
28a. Lee & Tindal (2000)	Grade 4	swd	0.16	0.91
28b. Lee & Tindal (2000)	Grade 7	swd	0.03	0.91
9b. Engelhard et al. (2011)	Grade 6	swd	-0.22	0.90
14b. Harmon (1999)	School A2	nswd	0.14	0.90
20b. Randall & Engelhard (2010)	Grade 6	swd	-0.24	0.90
19h. Helwig & Tindal (2003)	Grade 8 (Form B)	swd	0.11	0.85
14c. Harmon (1999)	School B	swd	0.14	0.85
34. Weston (2002)		swd	0.69	0.85
26a. Meloy et al. (2002)	Math	nswd	0.97	0.85
26c. Meloy et al. (2002)	Science	nswd	0.96	0.85
26d. Meloy et al. (2002)	Using Expressions	nswd	0.54	0.84
26b. Meloy et al. (2002)	Reading	nswd	0.70	0.84
34. Weston (2002)		nswd	0.91	0.83
21. Lesaux et al. (2006)		nswd	0.91	0.82
29b. Schnirman (2005)	ProblemSolving	nswd	-0.00	0.81
19c. Helwig & Tindal (2003)	Grade 4 (Form B)	swd	-0.00	0.80
12d. Helwig et al. (2002)	Grade 8'	swd	0.04	0.79
12d. Helwig et al. (2002)	Grade 8'	nswd	-0.03	0.78
12a. Helwig et al. (2002)	Grade 4'	swd	-0.24	0.77
15a. Janson (2002)	Math Year3	swd	-0.00	0.77
30. Smith (2010)		swd	0.92	0.77
23. MacArthur & Cavalier (2004)		swd	1.14	0.77
16. Johnson (2000)		nswd	-0.00	0.77
12a. Helwig et al. (2002)	Grade 4'	nswd	0.19	0.76
19b. Helwig & Tindal (2003)	Grade 5 (Form A)	swd	0.16	0.76
16. Johnson (2000)		swd	0.52	0.75
21. Lesaux et al. (2006)		swd	1.49	0.74
15b. Janson (2002)	Math Year2	swd	0.19	0.74
29a. Schnirman (2005)	Math Concepts	nswd	-0.14	0.74
14a. Harmon (1999)	School A1	swd	0.97	0.74
23. MacArthur & Cavalier (2004)		nswd	0.22	0.74
19f. Helwig & Tindal (2003)	Grade 5 (Form B)	nswd	0.47	0.74
19e. Helwig & Tindal (2003)	Grade 7 (Form A)	swd	-0.39	0.73
19g. Helwig & Tindal (2003)	Grade 7 (Form B)	swd	-0.10	0.73
1. Abedi et al. (2010)		swd	-0.20	0.73
12b. Helwig et al. (2002)	Grade 5'	swd	0.97	0.72
12b. Helwig et al. (2002)	Grade 5'	nswd	-0.29	0.71
14b. Harmon (1999)	School A2	swd	0.97	0.67
15a. Janson (2002)	Math Year1	swd	0.27	0.67
19d. Helwig & Tindal (2003)	Grade 8 (Form A)	swd	0.07	0.67
12f. Helwig et al. (2002)	Grade 7'	nswd	0.17	0.65
12e. Helwig et al. (2002)	Grade 7'	swd	-0.94	0.63
12f. Helwig et al. (2002)	Grade 7'	swd	-0.57	0.63
19a. Helwig & Tindal (2003)	Grade 4 (Form A)	swd	0.49	0.63
12e. Helwig et al. (2002)	Grade 7'	nswd	0.21	0.62
22. Lewandowski et al. (2007)		swd	0.92	0.61
12g. Helwig et al. (2002)	Grade 8'	swd	0.06	0.61
10. Kosciolek & Ysseldyke (2000)		nswd	0.16	0.61
12c. Helwig et al. (2002)	Grade 4'	nswd	-0.20	0.60
12g. Helwig et al. (2002)	Grade 8'	nswd	0.06	0.59
12c. Helwig et al. (2002)	Grade 4'	swd	0.69	0.57
26a. Meloy et al. (2002)	Math	swd	0.50	0.57
22. Lewandowski et al. (2007)		nswd	1.10	0.56
26b. Meloy et al. (2002)	Reading	swd	1.10	0.54
26d. Meloy et al. (2002)	Using Expressions	swd	1.19	0.54
26c. Meloy et al. (2002)	Science	swd	1.17	0.53
10. Kosciolek & Ysseldyke (2000)		swd	0.54	0.52
19f. Helwig & Tindal (2003)	Grade 5 (Form B)	swd	0.29	0.50
3. Dackler (2002)		nswd	-0.20	0.37
3. Dackler (2002)		swd	-0.04	0.33
2. Brown (2007)		swd	1.16	0.32



' Condition order: not accommodated - accommodated
' Condition order: accommodated - not accommodated

Figure 20: Effect Size Estimates by Weights - Substudy Level (all data)

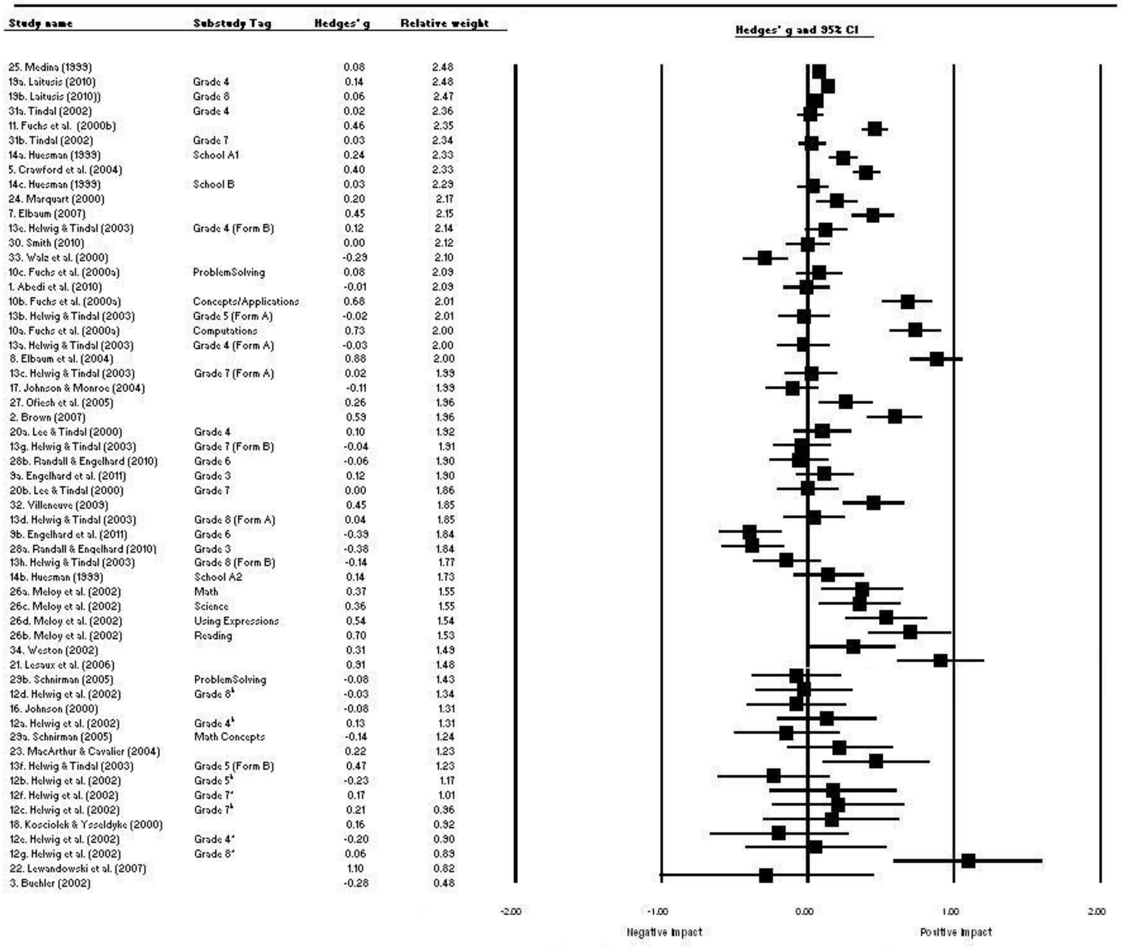


Study name	Substudy Tag	Hedges' g	Relative weight
19a. Laitusis (2010)	Grade 4	0.64	2.06
19b. Laitusis (2010)	Grade 8	0.36	2.06
11. Fuchs et al. (2000b)		0.45	2.02
7. Elbaum (2007)		0.19	2.01
6. Dempsey (2004)		0.89	2.00
8. Elbaum et al. (2004)		0.98	2.00
25. Medina (1999)		0.05	1.98
10c. Fuchs et al. (2000a)	ProblemSolving	0.25	1.94
10a. Fuchs et al. (2000a)	Computations	0.47	1.94
10b. Fuchs et al. (2000a)	Concepts/Applications	0.45	1.94
17. Johnson & Monroe (2004)		0.17	1.91
33. Walz et al. (2000)		-0.15	1.90
31b. Tindal (2002)	Grade 7	0.07	1.87
29a. Schnirman (2005)	Math Concepts	-0.03	1.87
29b. Schnirman (2005)	ProblemSolving	0.29	1.86
31a. Tindal (2002)	Grade 4	0.15	1.86
27. Ofiesh et al. (2005)		0.71	1.86
24. Marquart (2000)		0.23	1.86
32. Villeneuve (2009)		0.47	1.84
5. Crawford et al. (2004)		0.90	1.83
4. Calhoon et al. (2000)		0.23	1.83
28a. Randall & Engelhard (2010)	Grade 3	-0.04	1.83
9a. Engelhard et al. (2011)	Grade 3	-0.17	1.83
20a. Lee & Tindal (2000)	Grade 4	0.16	1.80
20b. Lee & Tindal (2000)	Grade 7	0.03	1.80
9b. Engelhard et al. (2011)	Grade 6	-0.22	1.79
28b. Randall & Engelhard (2010)	Grade 6	-0.24	1.79
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	0.11	1.72
14c. Huesman (1999)	School B	0.14	1.72
34. Weston (2002)		0.63	1.71
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	-0.08	1.62
12d. Helwig et al. (2002)	Grade 8[b]	0.04	1.62
12a. Helwig et al. (2002)	Grade 4[b]	-0.24	1.59
15c. Janson (2002)	Math Year3	-0.08	1.58
30. Smith (2010)		0.32	1.58
23. MacArthur & Cavalier (2004)		1.14	1.58
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	0.16	1.57
16. Johnson (2000)		0.52	1.55
21. Lesaux et al. (2006)		1.43	1.53
15b. Janson (2002)	Math Year2	0.13	1.53
14a. Huesman (1999)	School A1	0.37	1.53
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	-0.39	1.51
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.18	1.51
1. Abedi et al. (2010)		-0.20	1.50
12b. Helwig et al. (2002)	Grade 5[b]	0.37	1.50
14b. Huesman (1999)	School A2	0.37	1.41
15a. Janson (2002)	Math Year1	0.27	1.41
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.07	1.41
12c. Helwig et al. (2002)	Grade 7[b]	-0.34	1.34
12f. Helwig et al. (2002)	Grade 7[a]	-0.57	1.33
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	0.49	1.33
22. Lewandowski et al. (2007)		0.92	1.30
12g. Helwig et al. (2002)	Grade 8[a]	0.06	1.30
12e. Helwig et al. (2002)	Grade 4[a]	0.63	1.23
26a. Meloy et al. (2002)	Math	0.58	1.23
26b. Meloy et al. (2002)	Reading	1.10	1.18
26d. Meloy et al. (2002)	Using Expressions	1.19	1.17
26c. Meloy et al. (2002)	Science	1.17	1.16
18. Kosciolek & Ysseldyke (2000)		0.54	1.14
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.23	0.85
3. Buehler (2002)		-0.04	0.75
2. Brown (2007)		1.16	0.73

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

Figure 21: Effect Size Estimates by Weights - Substudy Level (students with disabilities)



Study name	Substudy Tag	Hedges' g	Relative weight
25. Medina (1999)		0.08	2.48
19a. Laitusis (2010)	Grade 4	0.14	2.48
19b. Laitusis (2010))	Grade 8	0.06	2.47
31a. Tindal (2002)	Grade 4	0.02	2.36
11. Fuchs et al. (2000b)		0.46	2.35
31b. Tindal (2002)	Grade 7	0.03	2.34
14a. Huesman (1999)	School A1	0.24	2.33
5. Crawford et al. (2004)		0.40	2.33
14c. Huesman (1999)	School B	0.03	2.29
24. Marquart (2000)		0.20	2.17
7. Elbaum (2007)		0.45	2.15
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	0.12	2.14
30. Smith (2010)		0.00	2.12
33. Walz et al. (2000)		-0.29	2.10
10c. Fuchs et al. (2000a)	ProblemSolving	0.08	2.09
1. Abedi et al. (2010)		-0.01	2.09
10b. Fuchs et al. (2000a)	Concepts/Applications	0.68	2.01
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	-0.02	2.01
10a. Fuchs et al. (2000a)	Computations	0.73	2.00
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	-0.03	2.00
8. Elbaum et al. (2004)		0.88	2.00
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	0.02	1.99
17. Johnson & Monroe (2004)		-0.11	1.99
27. Ofiesh et al. (2005)		0.26	1.96
2. Brown (2007)		0.59	1.96
20a. Lee & Tindal (2000)	Grade 4	0.10	1.92
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.04	1.91
28b. Randall & Engelhard (2010)	Grade 6	-0.06	1.90
9a. Engelhard et al. (2011)	Grade 3	0.12	1.90
20b. Lee & Tindal (2000)	Grade 7	0.00	1.86
32. Villeneuve (2009)		0.45	1.85
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.04	1.85
9b. Engelhard et al. (2011)	Grade 6	-0.39	1.84
28a. Randall & Engelhard (2010)	Grade 3	-0.38	1.84
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	-0.14	1.77
14b. Huesman (1999)	School A2	0.14	1.73
26a. Meloy et al. (2002)	Math	0.37	1.55
26c. Meloy et al. (2002)	Science	0.36	1.55
26d. Meloy et al. (2002)	Using Expressions	0.54	1.54
26b. Meloy et al. (2002)	Reading	0.70	1.53
34. Weston (2002)		0.31	1.49
21. Lesaux et al. (2006)		0.91	1.48
29b. Schnirman (2005)	ProblemSolving	-0.08	1.43
12d. Helwig et al. (2002)	Grade 8[b]	-0.03	1.34
16. Johnson (2000)		-0.08	1.31
12a. Helwig et al. (2002)	Grade 4[b]	0.13	1.31
29a. Schnirman (2005)	Math Concepts	-0.14	1.24
23. MacArthur & Cavalier (2004)		0.22	1.23
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.47	1.23
12b. Helwig et al. (2002)	Grade 5[b]	-0.23	1.17
12f. Helwig et al. (2002)	Grade 7[a]	0.17	1.01
12c. Helwig et al. (2002)	Grade 7[b]	0.21	0.96
18. Kosciolek & Ysseldyke (2000)		0.16	0.92
12e. Helwig et al. (2002)	Grade 4[a]	-0.20	0.90
12g. Helwig et al. (2002)	Grade 8[a]	0.06	0.89
22. Lewandowski et al. (2007)		1.10	0.82
3. Buehler (2002)		-0.28	0.48

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

Figure 22: Effect Size Estimates by Weights - Substudy Level (typically developing students)

Appendix L

Effect Sizes and Standard Errors for Students with Disabilities: Study as the Unit of Analysis

			95% Confidence Interval		
Study name	**ES**[b]	**Std Err**[b]	**LL**[b]	**UL**[b]	***p*(ES)**
21. Lesaux et al. (2006)	1.43	0.18	1.07	1.79	< 0.001
2. Brown (2007)	1.16	0.42	0.33	1.98	0.006
23. MacArthur & Cavalier (2004)	1.14	0.17	0.80	1.48	< 0.001
8. Elbaum et al. (2004)	0.98	0.06	0.86	1.09	< 0.001
22. Lewandowski et al. (2007)	0.92	0.24	0.45	1.39	< 0.001
5. Crawford et al. (2004)	0.90	0.11	0.68	1.11	< 0.001
6. Dempsey (2004)	0.89	0.05	0.78	1.00	< 0.001
27. Ofiesh et al. (2005)	0.71	0.10	0.51	0.91	< 0.001
34. Weston (2002)	0.63	0.14	0.35	0.91	< 0.001
26. Meloy et al. (2002)[a]	0.58	0.26	0.08	1.08	0.024
18. Kosciolek & Ysseldyke (2000)	0.54	0.28	-0.01	1.09	0.053
16. Johnson (2000)	0.52	0.18	0.16	0.87	0.004
19. Laitusis (2010)[a]	0.51	0.02	0.48	0.54	< 0.001
32. Villeneuve (2009)	0.47	0.11	0.25	0.69	< 0.001
11. Fuchs et al. (2000b)	0.45	0.05	0.36	0.54	< 0.001
10. Fuchs et al. (2000a)[a]	0.39	0.05	0.30	0.48	< 0.001
30. Smith (2010)	0.32	0.17	-0.02	0.66	0.061
14. Huesman (1999)[a]	0.25	0.10	0.06	0.45	0.011
24. Marquart (2000)	0.23	0.10	0.03	0.44	0.025
4. Calhoon et al. (2000)	0.23	0.11	0.01	0.45	0.039
7. Elbaum (2007)	0.19	0.05	0.09	0.30	0.000
17. Johnson & Monroe (2004)	0.17	0.09	0.00	0.35	0.057
29. Schnirman (2005)[a]	0.12	0.07	-0.02	0.27	0.083
31. Tindal (2002)[a]	0.11	0.07	-0.03	0.25	0.123
20. Lee & Tindal (2000)[a]	0.10	0.08	-0.07	0.26	0.252
15. Janson (2002)[a]	0.08	0.11	-0.14	0.29	0.484
25. Medina (1999)	0.05	0.07	-0.08	0.18	0.474
13. Helwig & Tindal (2003)[a]	0.02	0.07	-0.11	0.15	0.796
12. Helwig et al. (2002)[a]	-0.02	0.08	-0.17	0.13	0.801
3. Buehler (2002)	-0.04	0.41	-0.84	0.77	0.931
33. Walz et al. (2000)	-0.15	0.09	-0.32	0.03	0.112
28. Randall & Engelhard (2010)[a]	-0.17	0.08	-0.33	-0.01	0.043
1. Abedi et al. (2010)	-0.20	0.19	-0.58	0.17	0.285
9. Engelhard et al. (2011)[a]	-0.24	0.08	-0.40	-0.08	0.004
Overall (random effects)	**0.36**	**0.06**	**0.25**	**0.48**	**< 0.001**

[a] effect size computed used combined substudies (students with disabilities)

[b] ES is Hedges' g effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Appendix M

Effect Sizes and Standard Errors for Students without Disabilities: Study as the Unit of Analysis

			95% Confidence Interval		
Study name	**ES[b]**	**Std Err[b]**	**LL[b]**	**UL[b]**	***p*(ES)**
22. Lewandowski et al. (2007)	1.10	0.26	0.58	1.61	< 0.001
21. Lesaux et al. (2006)	0.91	0.15	0.61	1.21	< 0.001
8. Elbaum et al. (2004)	0.88	0.09	0.70	1.06	< 0.001
2. Brown (2007)	0.59	0.10	0.40	0.78	< 0.001
10. Fuchs et al. (2000a)[a]	0.46	0.05	0.37	0.56	< 0.001
11. Fuchs et al. (2000b)	0.46	0.05	0.37	0.55	< 0.001
32. Villeneuve (2009)	0.45	0.11	0.24	0.67	< 0.001
7. Elbaum (2007)	0.45	0.07	0.30	0.59	< 0.001
5. Crawford et al. (2004)	0.40	0.05	0.30	0.50	< 0.001
26. Meloy et al. (2002)[a]	0.37	0.14	0.09	0.65	0.009
34. Weston (2002)	0.31	0.15	0.01	0.60	0.041
27. Ofiesh et al. (2005)	0.26	0.10	0.07	0.45	0.007
23. MacArthur & Cavalier (2004)	0.22	0.19	-0.14	0.58	0.237
24. Marquart (2000)	0.20	0.07	0.06	0.34	0.005
18. Kosciolek & Ysseldyke (2000)	0.16	0.24	-0.31	0.63	0.500
14. Huesman (1999)[a]	0.15	0.04	0.08	0.22	< 0.001
19. Laitusis (2010)[a]	0.10	0.01	0.08	0.13	< 0.001
25. Medina (1999)	0.08	0.02	0.05	0.11	< 0.001
20. Lee & Tindal (2000)[a]	0.05	0.07	-0.09	0.20	0.460
13. Helwig & Tindal (2003)[a]	0.03	0.04	-0.04	0.09	0.475
31. Tindal (2002)[a]	0.02	0.03	-0.04	0.09	0.490
12. Helwig et al. (2002)[a]	0.01	0.08	-0.14	0.16	0.883
30. Smith (2010)	0.00	0.08	-0.15	0.15	1.000
1. Abedi et al. (2010)	-0.01	0.08	-0.17	0.15	0.916
16. Johnson (2000)	-0.08	0.17	-0.42	0.27	0.666
17. Johnson & Monroe (2004)	-0.11	0.09	-0.29	0.08	0.257
29. Schnirman (2005)[a]	-0.11	0.12	-0.34	0.13	0.377
9. Engelhard et al. (2011)[a]	-0.13	0.07	-0.28	0.01	0.078
28. Randall & Engelhard (2010)[a]	-0.22	0.07	-0.36	-0.07	0.004
3. Buehler (2002)	-0.28	0.38	-1.02	0.45	0.453
33. Walz et al. (2000)	-0.29	0.08	-0.45	-0.13	< 0.001
Overall (random effects)	**0.19**	**0.04**	**0.12**	**0.26**	**< 0.001**

[a] effect size computed used combined substudies (students with disabilities)

[b] ES is Hedges' g effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Appendix N

Effect Sizes and Standard Errors for Students with Disabilities: Substudy as the Unit of Analysis

				95% Confidence Interval		
Study name	**Study subgroup**	**ES[c]**	**Std Err[c]**	**LL[c]**	**Ul[c]**	***p* (ES)**
21. Lesaux et al. (2006)		1.43	0.18	1.07	1.79	< 0.001
26d. Meloy et al. (2002)	Using Expressions	1.19	0.27	0.66	1.73	< 0.001
26c. Meloy et al. (2002)	Science	1.17	0.27	0.63	1.71	< 0.001
2. Brown (2007)		1.16	0.42	0.33	1.98	0.006
23. MacArthur & Cavalier (2004)		1.14	0.17	0.80	1.48	< 0.001
26b. Meloy et al. (2002)	Reading	1.10	0.27	0.57	1.63	< 0.001
8. Elbaum et al. (2004)		0.98	0.06	0.86	1.09	< 0.001
22. Lewandowski et al. (2007)		0.92	0.24	0.45	1.39	< 0.001
5. Crawford et al. (2004)		0.90	0.11	0.68	1.11	< 0.001
6. Dempsey (2004)		0.89	0.05	0.78	1.00	< 0.001
27. Ofiesh et al. (2005)		0.71	0.10	0.51	0.91	< 0.001
19a. Laitusis (2010)	Grade 4	0.64	0.02	0.60	0.68	< 0.001
34. Weston (2002)		0.63	0.14	0.35	0.91	< 0.001
12e. Helwig et al. (2002)	Grade 4[a]	0.63	0.26	0.13	1.13	0.014
26a. Meloy et al. (2002)	Math	0.58	0.26	0.08	1.08	0.024
18. Kosciolek & Ysseldyke (2000)		0.54	0.28	-0.01	1.09	0.053
16. Johnson (2000)		0.52	0.18	0.16	0.87	0.004
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	0.49	0.23	0.04	0.94	0.034
32. Villeneuve (2009)		0.47	0.11	0.25	0.69	< 0.001
10a. Fuchs et al. (2000a)	Computations	0.47	0.08	0.31	0.63	< 0.001
10b. Fuchs et al. (2000a)	Concepts/Applications	0.45	0.08	0.30	0.61	< 0.001
11. Fuchs et al. (2000b)		0.45	0.05	0.36	0.54	< 0.001
12b. Helwig et al. (2002)	Grade 5[b]	0.37	0.19	-0.01	0.75	0.054
14b. Huesman (1999)	School A2	0.37	0.21	-0.05	0.78	0.084
14a. Huesman (1999)	School A1	0.37	0.18	0.00	0.73	0.048
19b. Laitusis (2010)	Grade 8	0.36	0.02	0.31	0.40	< 0.001
30. Smith (2010)		0.32	0.17	-0.02	0.66	0.061
29b. Schnirman (2005)	ProblemSolving	0.29	0.10	0.09	0.49	0.005
15a. Janson (2002)	Math Year1	0.27	0.21	-0.15	0.68	0.209
10c. Fuchs et al. (2000a)	ProblemSolving	0.25	0.08	0.09	0.40	0.001
24. Marquart (2000)		0.23	0.10	0.03	0.44	0.025
4. Calhoon et al. (2000)		0.23	0.11	0.01	0.45	0.039
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.23	0.37	-0.50	0.95	0.540
7. Elbaum (2007)		0.19	0.05	0.09	0.30	< 0.001
17. Johnson & Monroe (2004)		0.17	0.09	0.00	0.35	0.057
20a. Lee & Tindal (2000)	Grade 4	0.16	0.12	-0.07	0.39	0.181
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	0.16	0.18	-0.19	0.50	0.378
31a. Tindal (2002)	Grade 4	0.15	0.10	-0.05	0.35	0.144
14c. Huesman (1999)	School B	0.14	0.14	-0.14	0.41	0.329
15b. Janson (2002)	Math Year2	0.13	0.18	-0.23	0.49	0.487
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	0.11	0.14	-0.17	0.38	0.436
31b. Tindal (2002)	Grade 7	0.07	0.10	-0.12	0.27	0.462
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.07	0.21	-0.35	0.49	0.753
12g. Helwig et al. (2002)	Grade 8[a]	0.06	0.24	-0.40	0.53	0.787
25. Medina (1999)		0.05	0.07	-0.08	0.18	0.474

				95% Confidence Interval		
Study name	**Study subgroup**	**ES[c]**	**Std Err[c]**	**LL[c]**	**Ul[c]**	***p* (ES)**
12d. Helwig et al. (2002)	Grade 8[b]	0.04	0.16	-0.28	0.36	0.817
20b. Lee & Tindal (2000)	Grade 7	0.03	0.12	-0.20	0.27	0.779
29a. Schnirman (2005)	Math Concepts	-0.03	0.10	-0.23	0.16	0.738
3. Buehler (2002)		-0.04	0.41	-0.84	0.77	0.931
28a. Randall & Engelhard (2010)	Grade 3	-0.04	0.11	-0.26	0.18	0.730
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	-0.08	0.16	-0.40	0.24	0.622
15c. Janson (2002)	Math Year3	-0.08	0.17	-0.42	0.25	0.632
33. Walz et al. (2000)		-0.15	0.09	-0.32	0.03	0.112
9a. Engelhard et al. (2011)	Grade 3	-0.17	0.11	-0.39	0.06	0.142
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.18	0.19	-0.56	0.19	0.335
1. Abedi et al. (2010)		-0.20	0.19	-0.58	0.17	0.285
9b. Engelhard et al. (2011)	Grade 6	-0.22	0.12	-0.46	0.02	0.071
28b. Randall & Engelhard (2010)	Grade 6	-0.24	0.12	-0.48	0.01	0.055
12a. Helwig et al. (2002)	Grade 4[b]	-0.24	0.17	-0.57	0.10	0.162
12c. Helwig et al. (2002)	Grade 7[b]	-0.34	0.23	-0.79	0.10	0.133
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	-0.39	0.19	-0.76	-0.02	0.039
12f. Helwig et al. (2002)	Grade 7[a]	-0.57	0.23	-1.02	-0.12	0.014
Overall (random effects)		**0.30**	**0.04**	**0.21**	**0.38**	**< 0.001**

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

[c] ES is Hedges' g effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Appendix O

Effect Sizes and Standard Errors for Students without Disabilities: Substudy as the Unit of Analysis

				95% Confidence Interval		
Study name	**Study subgroup**	**ES[c]**	**Std Err[c]**	**LL[c]**	**Ul[c]**	***p* (ES)**
22. Lewandowski et al. (2007)		1.10	0.26	0.58	1.61	< 0.001
21. Lesaux et al. (2006)		0.91	0.15	0.61	1.21	< 0.001
8. Elbaum et al. (2004)		0.88	0.09	0.70	1.06	< 0.001
10a. Fuchs et al. (2000a)	Computations	0.73	0.09	0.55	0.91	< 0.001
26b. Meloy et al. (2002)	Reading	0.70	0.15	0.41	0.98	< 0.001
10b. Fuchs et al. (2000a)	Concepts/Applications	0.68	0.09	0.50	0.86	< 0.001
2. Brown (2007)		0.59	0.10	0.40	0.78	< 0.001
26d. Meloy et al. (2002)	Using Expressions	0.54	0.14	0.26	0.82	< 0.001
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	0.47	0.19	0.10	0.83	0.012
11. Fuchs et al. (2000b)		0.46	0.05	0.37	0.55	< 0.001
32. Villeneuve (2009)		0.45	0.11	0.24	0.67	< 0.001
7. Elbaum (2007)		0.45	0.07	0.30	0.59	< 0.001
5. Crawford et al. (2004)		0.40	0.05	0.30	0.50	< 0.001
26a. Meloy et al. (2002)	Math	0.37	0.14	0.09	0.65	0.009
26c. Meloy et al. (2002)	Science	0.36	0.14	0.08	0.64	0.013
34. Weston (2002)		0.31	0.15	0.01	0.60	0.041
27. Ofiesh et al. (2005)		0.26	0.10	0.07	0.45	0.007
14a. Huesman (1999)	School A1	0.24	0.05	0.14	0.34	< 0.001
23. MacArthur & Cavalier (2004)		0.22	0.19	-0.14	0.58	0.237
12c. Helwig et al. (2002)	Grade 7[b]	0.21	0.23	-0.25	0.66	0.376
24. Marquart (2000)		0.20	0.07	0.06	0.34	0.005
12f. Helwig et al. (2002)	Grade 7[a]	0.17	0.22	-0.27	0.61	0.443
18. Kosciolek & Ysseldyke (2000)		0.16	0.24	-0.31	0.63	0.500
14b. Huesman (1999)	School A2	0.14	0.12	-0.10	0.38	0.246
19a. Laitusis (2010)	Grade 4	0.14	0.02	0.10	0.17	< 0.001
12a. Helwig et al. (2002)	Grade 4[b]	0.13	0.17	-0.21	0.47	0.459
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	0.12	0.08	-0.02	0.27	0.102
9a. Engelhard et al. (2011)	Grade 3	0.12	0.10	-0.09	0.32	0.264
20a. Lee & Tindal (2000)	Grade 4	0.10	0.10	-0.10	0.30	0.322
25. Medina (1999)		0.08	0.02	0.05	0.11	< 0.001
10c. Fuchs et al. (2000a)	ProblemSolving	0.08	0.08	-0.08	0.23	0.352
19b. Laitusis (2010))	Grade 8	0.06	0.02	0.02	0.10	0.004
12g. Helwig et al. (2002)	Grade 8[a]	0.06	0.25	-0.43	0.54	0.820
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	0.04	0.11	-0.17	0.25	0.714
14c. Huesman (1999)	School B	0.03	0.06	-0.08	0.14	0.548
31b. Tindal (2002)	Grade 7	0.03	0.05	-0.07	0.12	0.556
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	0.02	0.09	-0.16	0.21	0.793
31a. Tindal (2002)	Grade 4	0.02	0.05	-0.07	0.11	0.693
20b. Lee & Tindal (2000)	Grade 7	0.00	0.11	-0.21	0.21	0.984
30. Smith (2010)		0.00	0.08	-0.15	0.15	1.000
1. Abedi et al. (2010)		-0.01	0.08	-0.17	0.15	0.916
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	-0.02	0.09	-0.20	0.15	0.797
12d. Helwig et al. (2002)	Grade 8[b]	-0.03	0.17	-0.36	0.31	0.870
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	-0.03	0.09	-0.21	0.15	0.745
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	-0.04	0.10	-0.24	0.16	0.681

				95% Confidence Interval		
Study name	**Study subgroup**	**ES[c]**	**Std Err[c]**	**LL[c]**	**Ul[c]**	***p* (ES)**
28b. Randall & Engelhard (2010)	Grade 6	-0.06	0.10	-0.26	0.14	0.553
16. Johnson (2000)		-0.08	0.17	-0.42	0.27	0.666
29b. Schnirman (2005)	ProblemSolving	-0.08	0.16	-0.39	0.23	0.621
17. Johnson & Monroe (2004)		-0.11	0.09	-0.29	0.08	0.257
29a. Schnirman (2005)	Math Concepts	-0.14	0.18	-0.50	0.22	0.435
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	-0.14	0.12	-0.37	0.09	0.221
12e. Helwig et al. (2002)	Grade 4[a]	-0.20	0.24	-0.67	0.28	0.415
12b. Helwig et al. (2002)	Grade 5[b]	-0.23	0.19	-0.62	0.15	0.228
3. Buehler (2002)		-0.28	0.38	-1.02	0.45	0.453
33. Walz et al. (2000)		-0.29	0.08	-0.45	-0.13	< 0.001
28a. Randall & Engelhard (2010)	Grade 3	-0.38	0.11	-0.59	-0.16	0.001
9b. Engelhard et al. (2011)	Grade 6	-0.39	0.11	-0.61	-0.18	< 0.001
Overall (random effects)		**0.17**	**0.03**	**0.11**	**0.22**	**< 0.001**

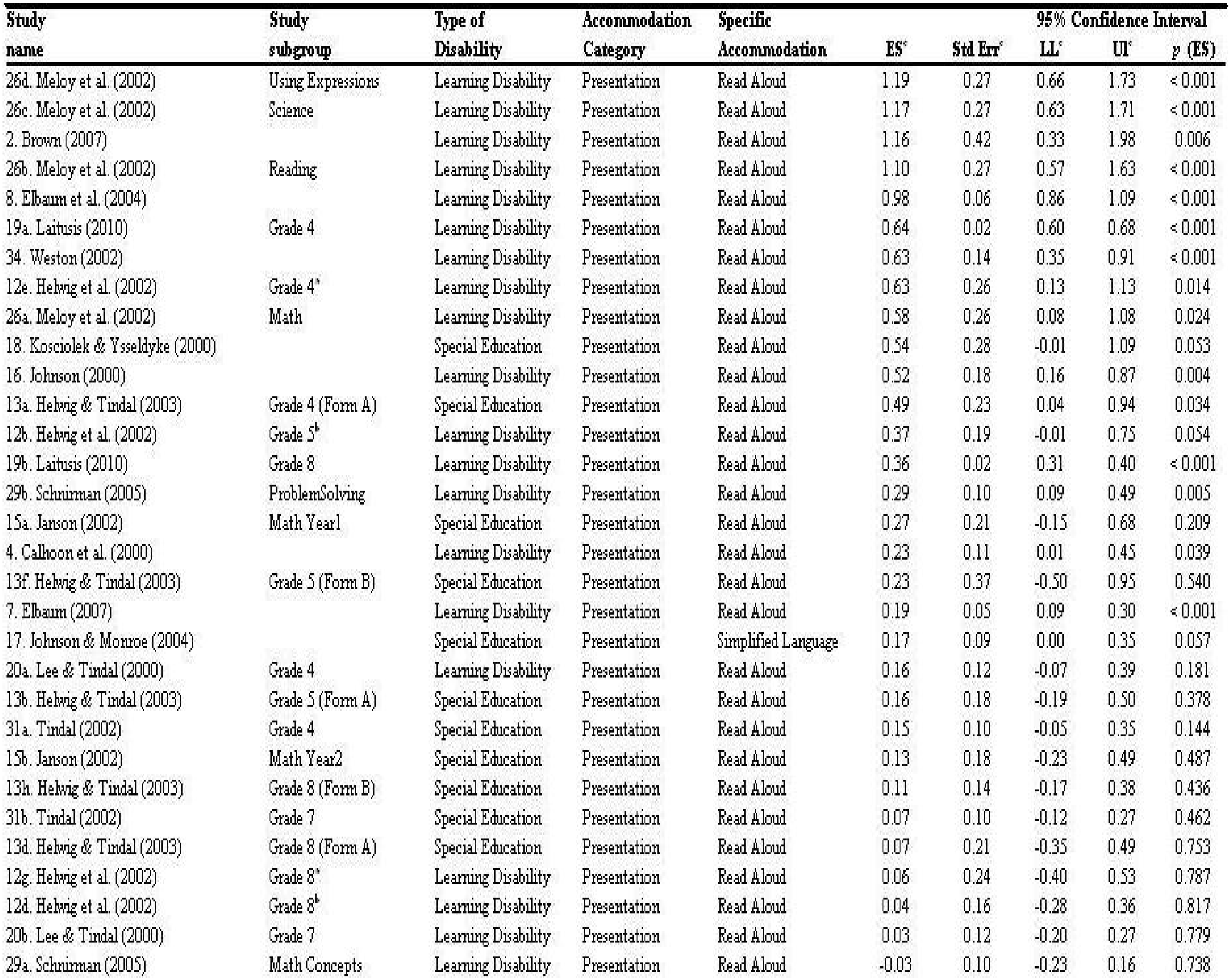
[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

[c] ES is Hedges' g effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Appendix P

Effect Sizes and Standard Errors for Students with Disabilities by Type of Disability, Accommodation Category, and Specific Accommodation: Substudy as the Unit of Analysis



Study	Study	Type of	Accommodation	Specific			95% Confidence Interval		
name	subgroup	Disability	Category	Accommodation	ES[c]	Std Err[c]	LL[c]	UI[c]	p (ES)
26d. Meloy et al. (2002)	Using Expressions	Learning Disability	Presentation	Read Aloud	1.19	0.27	0.66	1.73	< 0.001
26c. Meloy et al. (2002)	Science	Learning Disability	Presentation	Read Aloud	1.17	0.27	0.63	1.71	< 0.001
2. Brown (2007)		Learning Disability	Presentation	Read Aloud	1.16	0.42	0.33	1.98	0.006
26b. Meloy et al. (2002)	Reading	Learning Disability	Presentation	Read Aloud	1.10	0.27	0.57	1.63	< 0.001
8. Elbaum et al. (2004)		Learning Disability	Presentation	Read Aloud	0.98	0.06	0.86	1.09	< 0.001
19a. Laitusis (2010)	Grade 4	Learning Disability	Presentation	Read Aloud	0.64	0.02	0.60	0.68	< 0.001
34. Weston (2002)		Learning Disability	Presentation	Read Aloud	0.63	0.14	0.35	0.91	< 0.001
12e. Helwig et al. (2002)	Grade 4[a]	Learning Disability	Presentation	Read Aloud	0.63	0.26	0.13	1.13	0.014
26a. Meloy et al. (2002)	Math	Learning Disability	Presentation	Read Aloud	0.58	0.26	0.08	1.08	0.024
18. Kosciolek & Ysseldyke (2000)		Special Education	Presentation	Read Aloud	0.54	0.28	-0.01	1.09	0.053
16. Johnson (2000)		Learning Disability	Presentation	Read Aloud	0.52	0.18	0.16	0.87	0.004
13a. Helwig & Tindal (2003)	Grade 4 (Form A)	Special Education	Presentation	Read Aloud	0.49	0.23	0.04	0.94	0.034
12b. Helwig et al. (2002)	Grade 5[b]	Learning Disability	Presentation	Read Aloud	0.37	0.19	-0.01	0.75	0.054
19b. Laitusis (2010)	Grade 8	Learning Disability	Presentation	Read Aloud	0.36	0.02	0.31	0.40	< 0.001
29b. Schnirman (2005)	ProblemSolving	Learning Disability	Presentation	Read Aloud	0.29	0.10	0.09	0.49	0.005
15a. Janson (2002)	Math Year1	Special Education	Presentation	Read Aloud	0.27	0.21	-0.15	0.68	0.209
4. Calhoon et al. (2000)		Learning Disability	Presentation	Read Aloud	0.23	0.11	0.01	0.45	0.039
13f. Helwig & Tindal (2003)	Grade 5 (Form B)	Special Education	Presentation	Read Aloud	0.23	0.37	-0.50	0.95	0.540
7. Elbaum (2007)		Learning Disability	Presentation	Read Aloud	0.19	0.05	0.09	0.30	< 0.001
17. Johnson & Monroe (2004)		Special Education	Presentation	Simplified Language	0.17	0.09	0.00	0.35	0.057
20a. Lee & Tindal (2000)	Grade 4	Learning Disability	Presentation	Read Aloud	0.16	0.12	-0.07	0.39	0.181
13b. Helwig & Tindal (2003)	Grade 5 (Form A)	Special Education	Presentation	Read Aloud	0.16	0.18	-0.19	0.50	0.378
31a. Tindal (2002)	Grade 4	Special Education	Presentation	Read Aloud	0.15	0.10	-0.05	0.35	0.144
15b. Janson (2002)	Math Year2	Special Education	Presentation	Read Aloud	0.13	0.18	-0.23	0.49	0.487
13h. Helwig & Tindal (2003)	Grade 8 (Form B)	Special Education	Presentation	Read Aloud	0.11	0.14	-0.17	0.38	0.436
31b. Tindal (2002)	Grade 7	Special Education	Presentation	Read Aloud	0.07	0.10	-0.12	0.27	0.462
13d. Helwig & Tindal (2003)	Grade 8 (Form A)	Special Education	Presentation	Read Aloud	0.07	0.21	-0.35	0.49	0.753
12g. Helwig et al. (2002)	Grade 8[a]	Learning Disability	Presentation	Read Aloud	0.06	0.24	-0.40	0.53	0.787
12d. Helwig et al. (2002)	Grade 8[b]	Learning Disability	Presentation	Read Aloud	0.04	0.16	-0.28	0.36	0.817
20b. Lee & Tindal (2000)	Grade 7	Learning Disability	Presentation	Read Aloud	0.03	0.12	-0.20	0.27	0.779
29a. Schnirman (2005)	Math Concepts	Learning Disability	Presentation	Read Aloud	-0.03	0.10	-0.23	0.16	0.738

Study	Study	Type of	Accommodation	Specific			95% Confidence Interval		
name	subgroup	Disability	Category	Accommodation	ES[c]	Std Err[c]	LL[c]	Ul[c]	p (ES)
28a. Randall & Engelhard (2010)	Grade 3	Special Education	Presentation	Read Aloud	-0.04	0.11	-0.26	0.18	0.730
13e. Helwig & Tindal (2003)	Grade 4 (Form B)	Special Education	Presentation	Read Aloud	-0.08	0.16	-0.40	0.24	0.622
15c. Janson (2002)	Math Year3	Special Education	Presentation	Read Aloud	-0.08	0.17	-0.42	0.25	0.632
13g. Helwig & Tindal (2003)	Grade 7 (Form B)	Special Education	Presentation	Read Aloud	-0.18	0.19	-0.56	0.19	0.335
1. Abedi et al. (2010)		Special Education	Presentation	Segmented Text	-0.20	0.19	-0.58	0.17	0.285
28b. Randall & Engelhard (2010)	Grade 6	Special Education	Presentation	Read Aloud	-0.24	0.12	-0.48	0.01	0.055
12a. Helwig et al. (2002)	Grade 4[b]	Learning Disability	Presentation	Read Aloud	-0.24	0.17	-0.57	0.10	0.162
12c. Helwig et al. (2002)	Grade 7[b]	Learning Disability	Presentation	Read Aloud	-0.34	0.23	-0.79	0.10	0.133
13c. Helwig & Tindal (2003)	Grade 7 (Form A)	Special Education	Presentation	Read Aloud	-0.39	0.19	-0.76	-0.02	0.039
12f. Helwig et al. (2002)	Grade 7[a]	Learning Disability	Presentation	Read Aloud	-0.57	0.23	-1.02	-0.12	0.014
23. MacArthur & Cavalier (2004)		Learning Disability	Response	Dictation (scribe)	1.14	0.17	0.80	1.48	< 0.001
9a. Engelhard et al. (2011)	Grade 3	Special Education	Response	Calculator Use	-0.17	0.11	-0.39	0.06	0.142
9b. Engelhard et al. (2011)	Grade 6	Special Education	Response	Calculator Use	-0.22	0.12	-0.46	0.02	0.071
30. Smith (2010)		Learning Disability	Setting	Special Acoustics	0.32	0.17	-0.02	0.66	0.061
21. Lesaux et al. (2006)		Learning Disability	Timing-Scheduling	Extended Time	1.43	0.18	1.07	1.79	< 0.001
22. Lewandowski et al. (2007)		Other Health Impaired	Timing-Scheduling	Extended Time	0.92	0.24	0.45	1.39	< 0.001
5. Crawford et al. (2004)		Special Education	Timing-Scheduling	Extended Time	0.90	0.11	0.68	1.11	< 0.001
6. Dempsey (2004)		Learning Disability	Timing-Scheduling	Extended Time	0.89	0.05	0.78	1.00	< 0.001
27. Ofiesh et al. (2005)		Learning Disability	Timing-Scheduling	Extended Time	0.71	0.10	0.51	0.91	< 0.001
32. Villeneuve (2009)		Learning Disability	Timing-Scheduling	Extended Time	0.47	0.11	0.25	0.69	< 0.001
10a. Fuchs et al. (2000a)	Computations	Learning Disability	Timing-Scheduling	Extended Time	0.47	0.08	0.31	0.63	< 0.001
10b. Fuchs et al. (2000a)	Concepts/Applications	Learning Disability	Timing-Scheduling	Extended Time	0.45	0.08	0.30	0.61	< 0.001
11. Fuchs et al. (2000b)		Learning Disability	Timing-Scheduling	Extended Time	0.45	0.05	0.36	0.54	< 0.001
14b. Huesman (1999)	School A2	Learning Disability	Timing-Scheduling	Extended Time	0.37	0.21	-0.05	0.78	0.084
14a. Huesman (1999)	School A1	Learning Disability	Timing-Scheduling	Extended Time	0.37	0.18	0.00	0.73	0.048
10c. Fuchs et al. (2000a)	ProblemSolving	Learning Disability	Timing-Scheduling	Extended Time	0.25	0.08	0.09	0.40	0.001
24. Marquart (2000)		Special Education	Timing-Scheduling	Extended Time	0.23	0.10	0.03	0.44	0.025
14c. Huesman (1999)	School B	Learning Disability	Timing-Scheduling	Extended Time	0.14	0.14	-0.14	0.41	0.329
25. Medina (1999)		Learning Disability	Timing-Scheduling	Extended Time	0.05	0.07	-0.08	0.18	0.474
3. Buehler (2002)		Learning Disability	Timing-Scheduling	Extended Time	-0.04	0.41	-0.84	0.77	0.931
33. Walz et al. (2000)		Special Education	Timing-Scheduling	Extended Time	-0.15	0.09	-0.32	0.03	0.112

[a] Condition order: not accommodated - accommodated

[b] Condition order: accommodated - not accommodated

[c] ES is Hedges' g effect size estimate, Std Err is standard error, LL is lower limit, & UL is upper limit

Appendix Q

Random-effects Model – Students with Disabilities (All Data)

Table 30: *Random Effects Model for Students with Disabilities - Statistically Significant Variables Only*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Statistically Significant Variables

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model[a]		0.30	0.30	0.357			
Test Content (Math)	37				-0.70	-0.83	< 0.001
Test Content (Reading/LA)	18				-0.45	-0.40	0.005

[a] overall model for statistically significant variables only

Table 31: *Random Effects Model for Students with Disabilities - Overall Model without Test Accommodation*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance (no researcher-manipulated variables)

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model[a]		0.32	0.42	0.523			
Disability Classification	38[b]				0.16	0.18	0.150
Grade Level/s (Elementary)	27				-0.04	-0.05	0.827
Grade Level/s (Middle school)	29				-0.11	-0.13	0.554
Publication Year	62				0.00	-0.02	0.857
Publication Type (Journal)	40				0.01	0.02	0.940
Publication Type (Dissertation)	15				-0.15	-0.16	0.433
Test Content (Math)	37				-0.64	-0.76	< 0.001
Test Content (Reading/LA)	18				-0.35	-0.39	0.054
Test Format	47[c]				-0.17	-0.17	0.175

[a] overall model does not include researcher-manipulated variable (test accommodation)
[b] total for students with learning disabilities
[c] total for multiple-choice format

Appendix R

Random-effects Model: Students with Disabilities - Timing & Presentation Accommodation Data Only

Table 32: *Random Effects Model for Students with Disabilities - Statistically Significant Variables Only*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Statistically Significant Variables

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model[a]		0.30	0.16	0.322			
Test Content (Math)	35				-0.33	-0.40	0.001

[a] overall model for statistically significant variables only

Table 33: *Random Effects Model for Students with Disabilities - Overall Model without Test Accommodation*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance (no researcher-manipulated variables)

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model		0.30	0.38	0.364			
Disability Classification	36[a]				0.17	0.20	0.119
Grade Level/s (Elementary)	25				-0.02	-0.03	0.902
Grade Level/s (Middle school)	28				-0.09	-0.12	0.592
Publication Year	58				0.00	0.00	0.988
Publication Type (Journal)	37				0.00	0.00	0.986
Publication Type (Dissertation)	14				-0.14	-0.15	0.456
Test Content (Math)	35				-0.63	-0.75	< 0.001
Test Content (Reading/LA)	17				-0.33	-0.37	0.072
Test Format	47[b]				-0.19	-0.18	0.141

[a] total for students with learning disabilities
[b] total for multiple-choice format

Appendix S

Random-effects Model: Students with Learning Disabilities - Timing & Presentation Accommodation Data Only

Table 34: *Random Effects Model for Students with Disabilities - Statistically Significant Variables Only*
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for Statistically Significant Variables

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model[a]		0.41	0.28	0.261			
Test Content (Math)	19				-0.44	-0.53	< 0.001

[a] overall model for statistically significant variables only

Table 35: *Random Effects Model for Students with Learning Disabilities - Timing & Presentation Accommodation Data Only*[a]
Mean Effect Size, R^2, Significance of the Residual Q, Variable Weight and Significance for the Overall Model

Variable	k	Mean ES	R^2	$p(Q_{residual})$	*b*	β	*p*(*b*)
Overall Model		0.40	0.48	0.217			
Grade Level/s (Elementary)	14				-0.03	-0.03	0.878
Grade Level/s (Middle school)	17				-0.02	-0.03	0.899
Publication Year	36				0.02	0.15	0.299
Publication Type (Journal)	23				0.03	0.03	0.916
Publication Type (Dissertation)	10				-0.31	-0.33	0.241
Test Content (Math)	19				-0.81	-0.97	< 0.001
Test Content (Reading/LA)	12				-0.32	-0.37	0.108
Test Format	29[b]				-0.22	-0.21	0.212

[a] for this subset of data test accommodation category data and specific test accommodation data are the same
[b] total for multiple-choice format