Abstract

MURR, NATALIE SIMONA. Examining Options for School-level Disaggregation of Achievement Outcomes for Students with Disabilities Under No Child Left Behind. (Under the direction of Dr. Ann Schulte.)

One of No Child Left Behind's main goals is to increase focus on the accountability of all students, including students with disabilities, by mandating the disaggregation of student scores by student subgroup. However, the disaggregation policy poses problems for schools when applied to the students with disabilities subgroup due to a number of unique characteristics of this subgroup, such as small sample size and low initial achievement levels. As a result, policymakers have implemented a number of policy variants aimed at counteracting the negative effects of the disaggregation policy on this subgroup. Although currently implemented, research has yet to evaluate the impact of these policy variants, either in terms of changes in school results for the students with disabilities subgroup or in terms of the validity of the results obtained. This study evaluated school-level outcomes for the students with disabilities subgroup under three policy variants: (a) including students in the disability subgroup for two years after they exit special education, (b) applying different minimum required subgroup sizes and confidence intervals when determining the percent of students reaching proficiency, and (c) substituting a performance index for percent proficient in Adequate Yearly Progress determinations. Multiple analyses were used to evaluate the effect of each policy on schools' outcomes for the students with disabilities subgroup. Results of these analyses, as well as the practical ramifications of adopting these policies, are discussed.

Examining Options for School-level Disaggregation of Achievement Outcomes for Students with Disabilities Under No Child Left Behind

by
Natalie Simona Murr

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Psychology

Raleigh, North Carolina

2014

APPROVED BY:

Ann Schulte, Ph.D.
Chair of Advisory Committee

Scott Stage, Ph.D.

Jeffery Braden, Ph.D

Biography

Natalie Simona Murr was born on May 31, 1981 in Marbach, Germany. She graduated from H.H. Arnold High School (Wiesbaden, Germany) in 1999 and went on to pursue her Bachelor’s degree in Psychology at the University of Tennessee, Knoxville and her Master’s degree in Comparative Social Policy at the University of Oxford in Oxford, England. Following graduation from Oxford, Natalie worked for three years in London, England before beginning her graduate career at North Carolina State University.

Table of Contents

List of Tables

List of Figures

The No Child Left Behind Act of 2001 (NCLB) has been described as one of the most important pieces of federal legislation of recent decades (Nagle, Yunker, & Malmgren, 2006; R. Simpson, LaCava, & Graner, 2004). The law is built upon a foundation that emphasizes increased school accountability and educational equality, with the overall goals of improving student achievement, eliminating gaps in achievement between groups of students, and ensuring that 100% of students are proficient by 2014. To accomplish these goals, NCLB required that schools establish grade level proficiency standards (applicable to all but the most severely cognitively impaired students), and then report, at the school, district, and state level, the extent to which students are meeting these grade level standards. In addition, the law holds schools accountable for the attainment of annual performance targets, referred to as adequate yearly progress, or AYP. Increasingly harsh sanctions are imposed on schools that consistently fail to meet proficiency targets.

The emphasis on accountability and assessment underlying NCLB is not new; indeed, both have been essential components of educational reform since the 1950s (Linn, 2000). However, NCLB (2001) is distinguishable from previous legislation by its clear emphasis on school-level accountability for student outcomes. The increased focus on accountability is perhaps best illustrated by the NCLB mandate that schools disaggregate and report proficiency data for particular subgroups of students, including economically disadvantaged students, ethnic minorities, English language learners, and students with disabilities. By requiring the disaggregation of student scores, policymakers aimed to ensure that all students, not just those within the general education population, are being held to the same high standards. Requiring schools to separately assess and report on the achievement of individual student groups allows policymakers to refocus schools' attention on learning and academic outcomes for some of the most vulnerable populations, presumably increasing the quality of education, improving achievement outcomes for these students, and reducing the achievement gap between groups.

Despite laudable intentions, the disaggregation policy has proven to be particularly problematic for schools when applied to the students with disabilities subgroup. Research consistently indicates that a smaller percentage of students with disabilities meet AYP proficiency targets than general education students (Cole, 2006; Mintrop & Sunderman, 2009; M. Simpson, Gong, & Marion, 2005; Ysseldyke & Bielinski, 2002), often for reasons outside of the school's direct control. NCLB (2001) states that the failure of any one subgroup to meet annual yearly progress (AYP) will result in an entire school failing; therefore, under current policy, the relatively low performance of the special education subgroup may cause an entire school to fail, even when AYP targets for other subgroups and the general education population have been met (Eckes & Swando, 2009). Thus, concerns about the fairness of the disaggregation policy- especially for schools with significant populations of students with disabilities- have intensified as increasingly large numbers of schools have failed to meet AYP, due to the relatively low performance of the students with disabilities subgroup.

Since implementation of NCLB, many researchers have sought to identify and address some of the problems unique to the application of the disaggregation policy to the students with disabilities subgroup (Allbritten, Mainzer, & Ziegler, 2004; Koretz & Barton, 2003; M. Simpson et al., 2005; Thurlow, 2000; Ysseldyke & Bielinski, 2002), including (a) the impact of reclassification on the composition of the special education subgroup, (b) the effect of small subgroup size, and (c) the effect of lower average test scores for students with disabilities. Each of these issues has proven particularly problematic for the students with disabilities subgroup, as each has the potential to distort student outcomes and jeopardize a school's ability to meet AYP for this population of students (Kane & Staiger, 2002; Linn, 2000).

Since implementation of NCLB, the United States Department of Education (USDE) has recognized many of the issues inherent in the law's disaggregation policies and subsequently passed new regulations that grant states increased flexibility in determining annual yearly progress for the students with disabilities subgroup (Keele, 2004; Stephenson, 2006). For example, states are now allowed to determine their own minimum subgroup size and use confidence intervals in their calculations of AYP. Some states have been allowed to also use index scores when calculating AYP for the students with disabilities subgroup. More recently, the USDE approved a plan submitted by the state of North Carolina to permit more flexible identification of the students with disabilities subgroup by including students who have exited special education services in the disability subgroup.

The implementation of these more relaxed regulations has been met with mixed reactions. Although many researchers have suggested that the new provisions have the potential to increase the reliability and validity of school-level results, others have criticized the federal government for providing opportunities for states to evade NCLB's accountability policies (Stephenson, 2006) and have questioned whether these policies may result in the unintended consequence of enabling schools to once again exclude students with disabilities from their accountability systems, at least in terms of disaggregated reporting and consequences for failing to meet AYP for this subgroup (M. Simpson et al., 2005). Until now, however, these conclusions remain largely conjecture, as only a few researchers have investigated the effect of these policies on achievement outcomes for the students with disabilities subgroup (Eckes & Swando, 2009; M. Simpson et al., 2005). Thus, the purpose of the proposed study is to evaluate the extent to which these revisions in policy alter disaggregated special education outcomes at the school level in the state of North Carolina.

This manuscript begins with a review of the literature, including a summary of special education legislation and an overview of NCLB's key accountability and assessment requirements. The next section outlines key challenges to the disaggregation policy as it relates to the special education subgroup, as well as some of the proposed policy responses. Next, a summary of the purpose of the study and its hypotheses are presented. Finally, this manuscript concludes with the method and data analytic plan used to evaluate the study's key research questions.

Literature Review

An Overview of Education Legislation

Special education law. Prior to the 1950s, federal involvement in education was extremely limited. Without a clear constitutional mandate requiring federal involvement in education, responsibility for the provision of educational services fell exclusively to individual states (Huefner, 2006; Jacob, Decker, & Hartshorne, 2011). In addition, the lack of mention of education within the US Constitution also meant that citizens were not afforded a right to educational services. Thus, states could- and often did- opt to only provide educational services to the most able and educable students, thereby excluding more vulnerable groups of children, including students with disabilities (Jacob et al., 2011).

It was not until the Civil Rights era of the 1950s and 1960s that individuals began to question the legitimacy of states' exclusionary educational practices. Building on the national movement for increased equality and integration, special education advocates pushed for integration and equal access to education for all students, including students with disabilities. What followed was a significant shift, both in legislation and practice, to ensure that students with disabilities were provided equal access to the same free and appropriate education as their non-disabled peers (Huefner, 2006; Jacob et al., 2011). The 1954 Supreme Court ruling Brown v. Board of Education provided legal support to the claim that equality of educational opportunity was a right afforded to all citizens by the U.S. Constitution, and required that states provide equal access to educational services to all students. The passing of key legislation, such as the Education of the Handicapped Act (EHA) of 1970, ensured a focus on the appropriate education of students with disabilities, including their right to an individualized educational program in the least restrictive environment (Gordon, 2006). Subsequent reauthorizations of EHA, including the Education for All Handicapped Children Act (EAHCA) of 1975 and the Individuals with Disabilities Education Act (IDEA) of 1990 and 1997, continued to promote the inclusion of students with disabilities in standards-based reform efforts, including state and district-level assessments (Huefner, 2006).

Although seminal to ensuring equality of access to educational services for students with disabilities, legislation during this time did not require that schools be held responsible for the educational outcomes of these students (McLaughlin & Thurlow, 2003). With the passage of the 2004 reauthorization of the Individuals with Education Act (IDEIA), Congress signaled a clear shift in emphasis from only holding schools accountable for ensuring access to educational services, to specifically outlining new mandates which required school accountability to be based on academic outcomes and achievement (Gordon, 2006). As an example, Section 1412 of IDEIA (2004) clearly outlined the requirement for schools to establish strict performance goals for all students, including those with disabilities. In addition, Section 1414 (IDEIA, 2004) mandated that the individual education program (IEP) of each student with a disability include a description of how performance for each student will be measured, as well as the specific services and accommodations provided to the child to help him or her meet these goals. Both sections also required the mandatory participation of all students in state and district-wide assessment programs. This focus on accountability was not limited to special education law; indeed, similar provisions were outlined in the general education legislation as well. The No Child Left Behind Act of 2001, for example, was unprecedented in its focus on student achievement data for accountability purposes. Since its implementation, schools have been expected to evidence more than just compliance with legal procedures and inclusion mandates. As discussed in the following section, accountability under NCLB is now closely linked to ambitious content and performance standards, the achievement of similarly high educational outcomes, and ensuring that schools meet AYP targets (McLaughlin & Thurlow, 2003).

The No Child Left Behind Act (NCLB) of 2001. The No Child Left Behind Act was officially signed into law in 2002 by former President George Bush, initiating a series of landmark educational reforms aimed at improving the quality of schools and increasing student achievement. Four key principles served as the basis for this legislation, including (a) stronger accountability for student outcomes, (b) greater flexibility in the use of federal funds for states and school districts, (c) more choice for parents in the selection of high-performing schools, and (d) increased emphasis on scientifically based educational programs and practices ("Four Pillars," 2012; USDE, 2002). Of these, NCLB's accountability policies have produced the most significant demands on schools and school districts, requiring that they modify or redevelop accountability plans and practices to align with the law's new mandates ("Four Pillars," 2012; USDE, 2002).

In its goal of increasing student achievement, NCLB required that all states develop and implement strict accountability systems that enable the yearly monitoring of student performance. Specifically, schools were required to identify challenging academic content and performance standards in the areas of math, reading/language arts, and science, and develop assessments that would be used to determine students' performance against these standards. At the elementary level, these assessments are conducted annually during grades 3-8. In addition, schools have been required to establish cut-off scores to define proficiency in each subject and develop annual performance targets to ensure that all students are proficient by 2014. These annual performance targets are referred to as AYP (NCLB, 2001).

Unlike previous legislation, NCLB's mandates extend beyond simply determining how a school should be held accountable for student achievement to also identifying who should be encompassed within these requirements. The law requires that schools not only report outcomes for their student population as a whole, but also to disaggregate proficiency data by subgroups of students that are historically at risk for low achievement, such as students with disabilities, or who are ethnic minorities or economically disadvantaged. Schools must also show that they have met AYP by evidencing (amongst other things) that at least 95% of all students in general education, as well as 95% of each subgroup of students, have participated in assessments. In addition, schools must meet the proficiency targets (referred to as annual measurable objectives, or AMOs) set by the state in each subject. By the time both academic and participation targets have been applied to each student group and for each academic area (i.e., for reading and math), schools may face up to 41 possible targets that they must meet in any given year to pass AYP (AYP in Pennsylvania, 2006). Schools that fail to meet AYP will be identified for improvement or face a number of other federal sanctions.

The Problems Posed by Disaggregation

By requiring the disaggregation of student scores, policymakers signaled a clear commitment to ensuring equal access for all students to the benefits of standards-based reform, including challenging standards and high expectations, increased participation in state and district assessments, and greater transparency amongst teachers and parents as to the performance of vulnerable groups, including students with disabilities (Hardman & Dawson, 2008; McLaughlin & Thurlow, 2003; Mintrop & Sunderman, 2009; Thurlow, 2000). However, the disaggregation policy presents unique challenges when applied to the students with disabilities subgroup. Three of these challenges, including the effect of reclassification, the impact of small subgroup size, and the effect of lower average test scores for students with disabilities entering elementary school, will be addressed in the following sections.

The effect of reclassification. One of the main purposes of NCLB's disaggregation policy is to enable educators and policymakers to monitor the performance of vulnerable students, including students with disabilities. By disaggregating student scores by subgroup, schools may determine whether these students are benefiting from educational practices and policies and achieving required levels of proficiency as mandated by NCLB (Linn, Baker, & Betebenner, 2002). To evaluate the performance of student subgroups, the disaggregation policy requires the use of a single group case study design, in which each subgroup's performance in a given year is compared against the state's AYP targets (Stevens, 2005). The results of these analyses are then used to determine which subgroups have met AYP expectations during that school year. In addition, these results are also compared across years as an indicator of trends in performance across time and subgroup.

As noted earlier, a school's performance (i.e., whether or not all subgroups within a school meet AYP targets) is directly tied to the receipt of federal sanctions; thus, it is crucial that any analysis of student performance be statistically reliable and valid. Unfortunately, the application of the disaggregation policy to the students with disabilities subgroup has proven particularly problematic when viewed from a statistical perspective. As Ysseldyke and Bielinski (2002) noted, comparing the performance of a group of students against a predetermined standard, such as AYP, requires that the comparison group be both clearly defined and consistent. Any variation in the composition of the comparison group may distort results, especially if such results are used to analyze changes in average test scores or trends in student performance from year to year (Kane & Staiger, 2002; Linn, 2000; Linn et al., 2002; Stevens, 2005). However, this requirement is not often met by the special education subgroup, as student reclassification often results in frequent transitions into and out of special education, thereby jeopardizing the stability of the subgroup. In their investigation into the magnitude of the effects of reclassification, Ysseldyke and Bielinski (2002) found that up to 20% of the special education population exited or entered special education in any given year, resulting in significant fluctuations in the composition of the subgroup. Hanushek, Kain, and Rivkin (1998) reported similarly high rates of student transitions in their evaluation of the special education population in elementary schools across Texas, and noted that approximately 10% of students in 4th grade entered special education in 5th grade, while 16% of special education students exited from special education in the same year. Such high fluctuations in subgroup composition have the potential to significantly distort school-level performance results (Linn et al., 2002; Stevens, 2005; Ysseldyke & Bielinski, 2002).

Further complicating the issue of subgroup stability is the fact that transitions into or out of special education are strongly related to academic performance. In their evaluation of the effects of reclassification, Ysseldyke and Bielinski (2002) analyzed the mean reading performance of over 200,000 students on the Texas Assessment of Academic Skills (TAAS) reading test. Average reading performance was calculated and compared in four groups of students over a period of four years: (a) students remaining in general education, (b) students moving from general education to special education, (c) students moving from special education to general education, and (d) students remaining in special education. Results of this analysis indicated that the average reading performance of the special education students (group D) was consistently lower than the average reading score for students leaving special education (group C) across the four years. Thus, students who left special education and returned to general education had higher average reading scores than those who remained in special education. Furthermore, the average reading performance of students entering special education (group B) was lower than the average reading performance of both students in special education (group D) and those transitioning back into general education (group C), indicating that low academic performance was consistently associated with student transitions into special education. In fact, the mean difference between the average reading scores of students leaving special education and those entering was, in some instances, as great as .75 standard deviation units.

Thus, although the composition of the students within the special education subgroup may be in a constant state of flux, proficiency levels, by the very nature of the special education reclassification process, may remain consistently low. Within the current system of special education reclassification, any gains in academic proficiency may result in the child being exited from special education. As a result, these students' higher performance levels are unlikely to be included in subsequent analyses of subgroup performance results. In fact, only the results of students currently in special education, and those who have recently entered, are included in these calculations and are, as Ysseldyke and Bielinski (2002) evidenced, likely to be relatively low. This characteristic of the group of students served in special education is particularly problematic in an accountability system that links school success and sanctions with expected student gains towards proficiency.

Together, the instability of the special education subgroup and the concentration of lower performing students in this group pose significant challenges to a school's ability to meet AYP expectations. By failing to account for student transitions into and out of special education, special education group proficiency results are unable to capture individual student gains in achievement, thereby falsely inflating the apparent achievement gap between general and special education students. In the same study described previously, Ysseldyke and Bielinski (2002) provided evidence for the effect of reclassification on the size of the achievement gap between general and special education students. In this study, the authors contrasted two different ways of defining special education group membership- one which accounted for the effect of reclassification, and one which did not- and evaluated differences in mean achievement scores on the Texas Assessment of Academic Skills (TAAS) reading and math tests between general and special education students. In the cohort-static method, special education group membership was defined according to the special education status of each student in the first year of the study only. Thus, students who received special education services during this first year were defined as special education students, regardless of whether they continued to receive services in subsequent years of the study. As the name suggests, this approach to defining group membership allowed special education status to be held constant, regardless of any actual transitions into or out of special education programs. The cohort-dynamic approach, on the other hand, redefined group membership each year according to whether each student received special education services in that year or not. Unlike the cohort-static approach, transitions into and out of special education were not held constant within the cohort-dynamic approach, as each student's membership was redefined each time he or she is entered or exited special education services.

Results of this comparison confirmed a discrepancy in the performance of special education students depending on whether group membership was defined according to the cohort-static or cohort-dynamic method. When the cohort-dynamic approach was used, mean student performance in both reading and math for the special education group dropped consistently across grades. In addition to comparing students' mean TAAS scores, Ysseldyke and Bielinski (2002) also used effect sizes to evaluate the size of the achievement gap across grades. Effect sizes were calculated by first converting raw scores on the TAAS to scaled scores on the Texas Learning Index (TLI). Like other scaled scores, those reported by the TLI have a fixed mean and standard deviation, thereby enabling the comparison of student performance between grades. In this instance, TLI scaled scores were also used to create a standardized measure of effect by calculating the mean achievement score difference between the reference group (e.g., general education students) and the focal group (e.g., special education students), and then dividing this by the standard deviation of the test scores for the reference group. In reading, effect sizes increased from -.48 in 4th grade to -.93 in 8th grade, a difference of almost .50 standard deviation units. A similar drop was observed for mean student performance in math, and effect sizes increased from -.64 in 4th grade to -1.16 by 8th grade. Thus, the achievement gap between general and special education students when defined using the cohort-dynamic approach increased substantially across grades.

When special education membership was defined using the cohort-static method, however, a more positive trend was observed. For the cohort-static group, mean student performance in both reading and math actually increased slightly from 4th to 8th grade. During this time, effect sizes for the reading test decreased from -.48 to -.42; in math, a similar increase in effect size between 4th and 8th grade was observed.

It is important to note that although differences in performance trends between the cohort-static and cohort-dynamic groups were observed, average performance on the TAAS for both special education groups was lower than for general education students, suggesting that special education students performed consistently lower than their general education peers, regardless of how special education group membership was defined. However, the size of this achievement gap between general and special education students was affected by the approach used to define special education group membership. When transitions into and out of special education were accounted for (the cohort-static approach), slight gains in student performance across time points were observed, and the size of achievement gap between these students and their general education peers did not fluctuate across grades. However, when reclassification was not accounted for and special education status was redefined each year (the cohort-dynamic approach), the size of the achievement gap increased steadily across time points.

Ysseldyke and Bielinski's study was the first to evaluate the effect of reclassification on the size of the achievement gap between general and special education students using large-scale achievement data, and to suggest a new approach to defining special education group membership in a way that accounted for patterns of transitions into and out of special education. More recently, Parker (2011) replicated Ysseldyke and Bielinksi's study using state-level, large-scale assessment data from North Carolina. Results from Parker's analysis found similar differences in the size of the achievement gaps between general and special education students depending on whether the special education subgroup was defined using the cohort-dynamic or the cohort-static method. Parker noted that the achievement gap between general and special education students either stayed consistent (as was the case for comparisons of mean reading achievement), or grew larger (e.g., comparisons of mean math achievement) when special education membership was defined using the cohort-dynamic approach. These findings mirrored those described by Ysseldyke and Bielinski (2002). Similarly, Parker noted that when special education membership was defined using the cohort-static method, comparison of mean math achievement scores indicated that the achievement gap between general education and special education students also remained unchanged. However, Parker noted that when the cohort-static approach was used in comparisons of mean reading scores, the size of the achievement gap between general and special education students grew smaller over time. This pattern was not observed in Ysseldyke and Bielinski's original study, and suggests that a narrowing of the achievement gap between these subgroups of children is possible when the transitory nature of special education is accounted for.

In summary, researchers have clearly recognized the challenges posed by the instability of the special education subgroup, as well as the implications that frequent student transitions into and out of special education may have on a school's ability to meet AYP for the students with disabilities subgroup. In addition, researchers have identified the use of alternate approaches to the definition of special education group membership that account for the inherent instability of the subgroup, thereby providing a more accurate representation of performance for students with disabilities. In recognition of the problem of reclassification, the state of North Carolina recently revised its accountability plan to allow for the more flexible identification of the disaggregated special education subgroup. Specifically, the revised North Carolina Accountability Plan now states that:

> Starting with the 2008-2009 school year, students previously identified as students with disabilities (SWD), who have exited SWD identification during the last two years, were included in the calculations for determining the status of the SWD subgroup for AYP only if that subgroup already met the minimum number of 40 students required for a subgroup (Education, 2010).

Thus, North Carolina schools are now allowed to include the achievement results of students who have exited special education within their calculations of AYP for the students with disabilities subgroup, thereby replicating the cohort-static approach to defining special education group membership.

However, although research supports the notion that redefining special education group membership can substantially affect the size of the achievement gap between general and special education students, no empirical evidence exists to indicate that these definitional approaches have any effect on school-level achievement outcomes. Although the cohort-static approach may enable schools to account for the increased performance of students who recently exited special education, overall achievement for this subgroup may continue to fall short of the school's AYP target. Thus, the purpose of the proposed study is to empirically evaluate the extent to which North Carolina's two-year post exit inclusion policy (described above) affects school-level outcomes for the students with disabilities subgroup. In addition, this study will determine whether the implementation of this policy significantly alters the percent of schools in North Carolina meeting AYP proficiency targets for the students with disabilities subgroup.

The impact of small subgroup sizes. Reclassification and the transitory nature of the special education subgroup have the potential to jeopardize the validity of performance measures for the students with disabilities subgroup. In addition, other sources of volatility on school-level measures of performance, including sampling error and non-persistent effects on student learning (e.g., a particularly disruptive child in the classroom), all have the potential to distort not only the validity, but also the reliability, of measures of school performance (Kane & Staiger, 2002; Linn et al., 2002), especially when sample size is small.

The issue of small subgroup size is particularly problematic for the students with disabilities subgroup, as the most commonly used subgroup size across states is approximately ten students (Eckes & Swando, 2009). When sample sizes are this small, even the low achievement of a few students can significantly distort overall group performance results (Eckes & Swando, 2009; Kane & Staiger, 2002).

As noted by Schulte and Villwock (2004), the problem of small subgroup size poses significant problems in an accountability system which emphasizes school-level disaggregation and reporting of student outcomes. First, the standard error of measurement for many large scale assessments only becomes small enough to allow reliable detection of changes across time when results are cumulated across sample sizes of approximately 25-30 students; however, the maximum number of students within special education classrooms (or within the special education subgroup in a school) is often below this threshold. Furthermore, calculations of special education outcomes may be influenced by a myriad of school-level factors unrelated to students' actual performance, including a particular school's special education classification process, frequent transitions into and out of special education, and differences in special education cohorts from one year to the next (e.g., student ability level). The negative effect of these factors is only confounded by small sample size, and is therefore more detrimental to the students with disabilities subgroup than other student populations within a school (Schulte & Villwock, 2004).

To illustrate the susceptibility of small subgroups to various sources of volatility, Schulte and Villwock (2004) compared general and special education student outcomes using three different measures of school performance: (1) the percent of students meeting grade-level proficiency targets, (2) the percent of students achieving high growth expectations, and (3) longitudinal examination of each student's proficiency across his or her elementary-school years. The first method (percent of students meeting grade-level proficiency targets in a given year) mirrors the approach currently required by NCLB in determining whether each school- and each subgroup within each school- has met AYP targets. Results of this study clearly illustrated considerable variability in the percent of students with disabilities achieving grade level proficiency across schools when the current NCLB method was used to calculate student outcomes. In the first year of the study, the percent of students with disabilities meeting proficiency standards in reading across schools ranged from approximately 37 to 80%, a difference of 43 percentage points. In comparison, the percent of general education students meeting proficiency standards in the first year of the study ranged from approximately 85 to 93%, a difference of only eight percentage points. Similar patterns were observed in years two and three of the study. Thus, greater variability in student outcomes was found between schools' special education groups than between their general education populations. The greater variability exhibited by special education subgroups may be due to the relatively small size of these samples, and the fact that smaller samples are more susceptible to the influence of sampling error than larger, more homogenous groups (Stevens, 2005).

In its original form, NCLB allowed states to determine their own minimum subgroup size, with the understanding that schools would only be held accountable for the performance of those subgroups that met this minimum size. Schools with too few students to constitute a subgroup would not be required to report disaggregated student results, but would instead automatically be counted as having met AYP expectations for that subgroup (NCLB, 2001). In response to growing concerns about the reliability of performance results based on small subgroup sizes, however, the USDE recently passed more flexible regulations allowing states to increase the minimum number of students required to form a subgroup (Erpenbach, Forte-Fast, & Potts, 2003; Forte-Fast & Erpenbach, 2004). In addition, the USDE has also approved the use of confidence intervals in AYP calculations (Forte-Fast & Erpenbach, 2004).

The recent changes in the government's regulations concerning subgroup size and confidence intervals prompted researchers to more fully investigate the effect of these policies on schools' ability to meet AYP. As an example, M. Simpson et al. (2005) evaluated schools' ability to meet AYP for the special education subgroup at various minimum subgroup sizes and confidence intervals. Using a single year of math and reading achievement scores for elementary and middle school students in five states, the authors calculated the percent of special education students and general education students meeting predetermined proficiency targets for the state. These results were then compared to the state annual measureable objectives (AMO) to determine whether each group met AYP. Each school was determined to have met AYP if both groups (special education and general education) met the state's AMO in both reading and math. Alternatively, a school was also determined to have met AYP if the general education group met AMO targets, and the special education group was automatically passed because they did not have enough students to meet the minimum sample size required for accountability purposes. Finally, school passing rates were calculated at varying minimum subgroup sizes (between 10 and 100) and confidence intervals (70%-99%) to evaluate whether passing rates were affected by fluctuations in minimum cell size and interval level.

Results of the M. Simpson et al. (2005) study clearly indicated that, in each of the five states evaluated, an increase in minimum cell size was strongly associated with a school's ability to meet AYP. In one state's analysis, the percent of schools meeting the state AMO using the largest observed cell size (100) was almost 60 percentage points larger than the percent of schools in that same state meeting AMO when the minimum cell size was set to 10. On average, states showed a 32% increase in schools meeting AMO targets when passing rates were calculated using a minimum-n of 100 versus a minimum-n of 10. A similar increase in the percent of schools passing AMO was observed when confidence intervals were used; however the size of this increase was not as great as that observed when varying minimum cell sizes were analyzed. On average, the difference between the percent of schools meeting AMO expectations when no confidence interval was applied, and the corresponding percent passing when a 99% confidence interval was used was approximately 13%. Thus, this study suggests that the practice of increasing minimum cell sizes and using confidence intervals in AYP calculations significantly increases the percentage of schools meeting AYP targets.

Although the results of M. Simpson et al.'s (2005) study appeared to provide support for the use of increased subgroup size and confidence intervals, the authors warned that the increase in the percentage of schools using these methods had negative implications for the inclusion of students with disabilities in state accountability systems. In their analysis, the authors noted that only 20% of the schools meeting the special education subgroup AMO did so while still assessing the performance of this subgroup; the other 80% were only able to meet the subgroup AMO by raising the minimum subgroup size to a level which no longer required schools to account for the performance of the students with disabilities subgroup. In other words, the vast majority of schools were only able to meet state AMOs by exempting themselves from NCLB's accountability provisions (M. Simpson et al., 2005).

Results of M. Simpson et al.'s (2005) study highlight another important concern regarding the practice of increasing minimum subgroup sizes, namely, the mislabeling of "passing" and "failing" schools. The authors noted that 80% of the schools meeting AYP for the students with disabilities subgroup were only able to do so by increasing the minimum sample size. In other words, 80% of schools were said to meet AYP for the disability subgroup when, in fact, they did not. These results have important implications on the accuracy of the information provided to school administrators, policymakers and the general public concerning both school and student performance. The very act of meeting AYP signals to many that a school has been successful in increasing or maintaining high levels of student performance commensurate with state expectations. Unfortunately, in the event that a school passes AYP due only to an inability to meet required subgroup sizes, this information is not accurate.

Thus, one of the drawbacks to the policy of increasing required minimum subgroup sizes is that it may create a dichotomy between a school's official AYP status and the 'truth' regarding student performance. In addition, this dichotomy limits the policy's ability to accurately differentiate between successful and failing schools, and increases the potential for invalid inferences to be made regarding school and student performance. In order for the required minimum subgroup size policy to be useful, therefore, it is essential that policymakers be able to judge the policy's accuracy and precision.

Researchers and policymakers in other fields frequently face similar demands. As an example, epidemiologists have long established the need to be able to accurately judge the usefulness of screening or diagnostic tests. As noted by Jekel et al. (2007) epidemiologists are explicitly interested in determining a screening test's ability to correctly identify those individuals who are known to have a disease (and whose test results, therefore, should be positive), as well as those individuals who are known not to have a disease (and whose test results should be negative). The rate at which a diagnostic screener is able to correctly classify the former is referred to as a test's sensitivity; the latter, on the other hand, is known as test specificity. A test with low sensitivity is one that frequently misclassifies diseased individuals as non-diseased, a type of error known as a false-negative. It is important to note that a test's sensitivity is the inverse of its false-negative error rate; in other words, as one increases, the other decreases. Similarly, a test with low specificity will yield a high false-positive error rate. This type of error relates to the rate at which a test incorrectly labels non-diseased individuals as having the disease. A test's specificity and false-positive error rate always add up to 1.0 (100%); thus, they, too, are inverses of one another.

Table 1
Comparison of Patients' True Disease Status and Screening Test Results

	True Disease Status		
Test Result	Diseased	Non-Diseased	Total
Positive	*A*	*b*	*a+ b*
Negative	*C*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c+d*

Note. Values in cells may be interpreted as follows: *a*= subjects with a true-positive test result; *b*= patients with a false-positive test result; *c*= patients with a false-negative test result; *d*= patients with a true-negative test result; *a+b*= all patients with a positive test result; *c+d*= all patients with a negative test result; *a+c*= all patients with the disease; *b+d*= all patients without the disease.

Once rates of specificity, sensitivity, false-negatives, and false-positives are known, these values may be used to identify the most appropriate cut-off point for a specific screening tool (i.e., the point at which there is a desirable balance achieved between sensitivity and false positives). Receiver operating characteristic (ROC) curves are increasingly used to graph the sensitivity and specificity for a screening tool at different cut-off points. Figure 1 below provides an example of such a ROC. As illustrated, a screening tool is considered most useful if it has a high rate of sensitivity and low false-positive error rate.

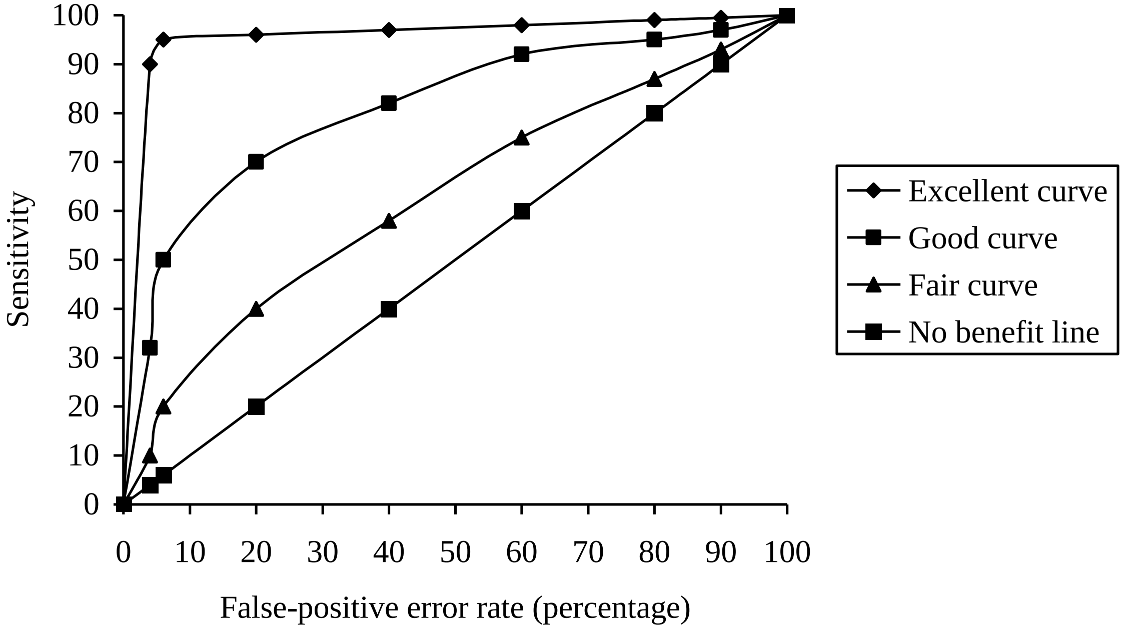

Figure 1. Example of receiver operating characteristic (ROC) curve for four tests. Adapted from "Understanding Errors in Clinical Medicine" by Jeckel et al., 2007.

A similar method may be applied to determine the accuracy of an educational policy, such as the required minimum subgroup size policy. As in epidemiology, the required minimum subgroup policy may be considered accurate if it correctly identifies 'passing' schools (schools that do, in fact, meet AYP targets) while at the same time correctly identifying "failing" schools (see Table 2 below).

Table 2

Comparison of Schools' True AYP Status and Minimum Required Subgroup Policy Results

	True AYP Status		
Policy Result	Failed	Passed	Total
Failed	*a*	*b*	*a+ b*
Passed	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c+d*

Note. Values in cells may be interpreted as follows: *a*= schools with a true-positive test result; *b*= schools with a false-positive test result; *c*= schools with a false-negative test result; *d*= schools with a true-negative test result; *a+b*= all schools with a positive test result; *c+d*= all schools with a negative test result; *a+c*= all schools that did not truly meet AYP; *b+d*= all schools that did truly meet AYP.

Unfortunately, M. Simpson et al.'s (2005) study suggests that the required minimum subgroup size policy may fail such accuracy tests, as the false-negative error rate actually increases as the minimum required subgroup size gets larger. It is due to this unintended consequence of the government's regulations concerning minimum sample size and confidence intervals that some authors, including Stephenson (2006), denounced the use of such policies in school calculations of AYP. In his discussion of recent amendments to NCLB's original accountability mandates, Stephenson (2006) suggested that such strategies present opportunities for states to evade accountability requirements regarding student outcomes, thereby circumventing the law's original intent to 'leave no child behind' (NCLB, 2001). With no ability to distinguish between states requesting to use these accommodations due to genuine concerns regarding the reliability of school performance measures, and those who seek a way of eluding NCLB’s accountability requirements, Stephenson (2006) warned that the new regulations may provide schools with a mechanism for manipulating AYP results to their benefit. Thus, although these regulations were originally developed to address concerns about the reliability of subgroup performance measures, the resulting exclusion of students with disabilities from large-scale assessments creates new concerns about the validity of AYP determinations (Stephenson, 2006). As summarized by M. Simpson et al.,

> If the implicit theory of action guiding NCLB accountability requirements is to improve instruction and thus outcomes for all students, schools and districts must be accountable for all subgroups in order to ensure that these students are appropriately served. Therefore, tinkering with the minimum-n to exclude substantial portions of special education students must be considered a threat to the validity of the Accountability system (2005, p.23).

Both the adoption of increased minimum subgroup sizes and the use of confidence intervals have the potential to reduce the influence of measurement error on measures of school performance, thereby increasing the accuracy of school-level disaggregated results. However, concerns regarding the unintended consequences of these approaches remain. To date, however, disputes regarding the use of both increased minimum subgroup sizes and confidence intervals remain largely speculative, as only one investigation has evaluated the effect of these approaches on schools’ ability to meet AYP (M. Simpson et al., 2005). Thus, the purpose of this study was to evaluate the effect of increased minimum subgroup sizes and confidence intervals on schools’ ability to meet (or be exempted) from AYP thresholds for their disability subgroup. In addition, this study empirically evaluated the extent to which the use of different minimum sample size thresholds results in the exclusion of the students with disabilities subgroup from North Carolina’s accountability system.

The problem of initially lower achievement for the students with disabilities subgroup. As noted previously, NCLB’s accountability mandates require each state to identify annual measurable objectives (AMO) that outline the specific percentage of students in each subgroup and the general education population who must meet designated proficiency goals on large-scale assessments in order for that school to pass AYP targets (NCLB, 2001). Additionally, states must ensure that proficiency targets increase steadily until 2014, at which point 100% proficiency is expected. Initial starting points are calculated on a state-by-state basis using baseline student outcome data for the 2001/02 school year. Each starting point is based on the higher of the two percentages: (a) the percent of students scoring at the proficient or above proficient level on large scale assessments in the least-achieving student demographic subgroup, or (b) the percentage of proficient students in the school at the 20th percentile of the state’s total enrollment among all schools. To calculate the latter, schools are first ranked according to the percentage of proficient students within each school. Starting from the school with the smallest percentage of proficient students, the state administrators were then required to identify the school at the 20th percentile for total student enrollment in the state of North Carolina. The percentage of proficient students in this school was then used as the initial starting point for all schools in the state, if this percentage was larger than the percentage identified in option a above. Separate starting points were calculated for both reading/language arts and math (USDE, 2011).

As such, initial starting points, as well as the difference between these starting points and the 100% proficiency target, differ substantially from state to state, creating great variability across states in the amount of gains in student performance expected. In addition to the variability between states, there is also substantial variation in individual school's starting performance within a state, and the difference between these starting points and percent proficiency targets. Schools with lower starting points will have further to go in ensuring that all students meet expected proficiency targets relative to those schools with higher initial percentages of proficient students. As Linn et al. (2002) noted, NCLB's required method for calculating starting points creates a situation in which states start off on an unlevel playing field, and this same concern extends to individual schools.

Thus, absolute proficiency targets may be more or less attainable, depending on the beginning level of proficiency of the school or population of students trying to reach them. For the students with disabilities subgroup, meeting state AMOs might prove particularly difficult, as these students are repeatedly shown to have lower average performance levels than their general education peers (Eckes & Swando, 2009; M. J. McLaughlin, 2006, USDE, 2011). Ysseldyke and Bielinski (2002) provided evidence for the gap in performance between special and general education students, as has been discussed elsewhere in this manuscript. In addition, Eckes and Swando (2009) presented further evidence for this achievement gap by comparing average proficiency levels for both general and special education students in three states. Results of this study confirmed that not only was the proficiency level of the students with disabilities subgroup lower than the general education group at baseline, the size of this achievement gap remains stable across grades. In regards to meeting state AMOs, students with disabilities therefore need to demonstrate larger gains in proficiency than general education students in order for the school to meet AYP targets.

Given the potential consequences that a low-achieving group may have on a school's ability to meet AYP, it is important to identify those factors that might contribute to a subgroup's lower proficiency. In the case of the special education subgroup, students with disabilities differ from students in the other subgroups in that many possess limitations in their ability to learn (Eckes & Swando, 2009). Furthermore, the very definition of disability posits the condition as something internal to the child, and outside of the control of the school. Although research has shown that access to a high-quality curriculum can result in increases in student learning and improved educational outcomes for students with disabilities (Bray & Kehle, 2011; Marzano, Pickering, & Pollock, 2001), school remediation efforts are unlikely to produce the increase in student performance needed to bridge the achievement gap between this subgroup and their general education peers. Yet, despite evidence supporting the lower proficiency of the students with disabilities subgroup, as well as limitations in schools' ability to correct for lower than expected growth, current NCLB policy mandates that this population of students reach the same level of proficiency at an identical rate as general education students (Eckes & Swando, 2009; M. J. McLaughlin, 2006).

As noted previously, a school's AYP determination is based on the ability to demonstrate that each subgroup of students meets an absolute level of proficiency on large-scale assessments. Thus, the main challenge for schools unable to meet AYP targets due to the lower average performance of the students with disabilities subgroup is not that special education students are not making gains in achievement, but that they are not making enough gains to allow the school to meet the proficiency target required to pass AYP. Within this system, relative growth is not considered when making AYP determinations. Therefore, a school that is able to show gains in academic proficiency for its special education students will only receive credit for this increase if it moves the subgroup's absolute proficiency level up to or above the target determined by the state. As an example, North Carolina schools were only able to meet AYP expectations for the 2009/10 school year if they could evidence that 43% of students within each subgroup met proficiency targets in math and reading. Schools that were able to increase the percent of students meeting proficiency standards, but that still failed to meet this absolute target, were not recognized for these improvements.

Thus, NCLB's current approach to determining AYP status appears to create a dichotomy between schools evidencing high performance (i.e., the ability to meet absolute proficiency targets, or AMOs), and schools demonstrating high growth (i.e., the ability to demonstrate increases in the number of students meeting proficiency targets on large-scale assessments), as illustrated by Figure 2 below. In some instances, these two strands come together to produce valid inferences about a school's performance. For example, group A in Figure 2 represents schools evidencing both high growth (i.e, an increase in the percent of students deemed proficient on large-scale assessments) and high performance (the school or subgroup meets the state AMO). These schools will, appropriately, meet AYP targets. Conversely, a school evidencing low growth and low performance (group D) will fail to meet AYP targets.

In some instances, however, AYP determinations may lead to invalid inferences regarding school quality and performance. For example, a particular subset of schools (group B in Figure 2 below) may succeed in substantially increasing the percentage of students meeting proficiency targets on large-scale assessments (high growth), but still fail to meet AYP (low performance). In this scenario, schools evidencing significant gains in student achievement are not credited for their success. This is particularly problematic for schools with significant populations of students with lower average performance, including students with disabilities. In both these instances, the lower starting performance of students means that a school must demonstrate unrealistically large gains in student proficiency in order to bridge the gap between its starting point and expected levels of achievement, or risk failing AYP. Additionally, the fourth group of schools represented in Figure 2 (group C) may fail to evidence significant student gains in achievement (low growth), but still meet NCLB’s absolute proficiency targets (high performance). Schools within this latter category typically begin with higher starting points, and therefore do not need to evidence large gains in achievement to meet state AMOs. Despite little or no student gains in academic performance, these schools will evade NCLB’s identification as a ‘failed school’ and associated federal sanctions.

	Performance		
Growth		High	Low
	High	(A) School meets AYP	(B) School does not meet AYP
	Low	(C) School meets AYP	(D) School does not meet AYP

Figure 2. A school’s ability to meet AYP targets may best be illustrated by the interplay between gains in student achievement (growth) and the school’s ability to meet absolute proficiency targets, or AMOs (performance).

In summary, NCLB’s method for determining AYP leads to an overly narrow focus on absolute proficiency targets and cut scores, thereby ignoring the ability of many schools to evidence gains in achievement for the lowest-performing students (including students with disabilities). To counteract this problem, some researchers have proposed that index scores be used in calculations of AYP (Linn et al., 2002). Index scores award a school credit for all levels of student proficiency, including that which occurs below the absolute proficiency cut score. Figure 3 provides one example of how an accountability model which incorporates index scores (often referred to as a Performance Index) assigns points to students at varying levels of proficiency, as compared to NCLB’s current percent proficiency model.

		Accountability Model	
		Status/Percent Proficient	**Performance Index**
Proficiency Level	**Advanced**	1.0	1.0
	Proficient	1.0	1.0
	High Basic	0	0.8
	Low Basic	0	0.6
	High Below Basic	0	0.4
	Low Below Basic	0	0.2
	Not Assessed	0	0

Figure 3. Weights assigned to each proficiency level compared across two different accountability models.

As illustrated in Figure 3, the Performance Index awards partial credit to schools for students who fall below the Proficient level. Thus, schools that succeed in moving students from the Low Basic to the High Basic level, for example, will be rewarded by receiving a score of 0.8 points for each of these students. In contrast, the status model only recognizes student improvement if it results in moving students from below proficiency to the proficiency cut-off or above, regardless of how much improvement may have been observed in students below this cut score. The goal of both models remains the same: to obtain an average score of 1.0 (equivalent to 100% proficiency) for all students, including each student subgroup, by 2014. The difference, however, is that a performance index allows the inclusion of all students' achievement, regardless of where on the proficiency spectrum these improvements occur, thereby providing a more valid and representative indication of how well students within a school are performing, on average, in any given year.

Given the potential advantages associated with the use of index scores, the U.S. Department of Education has supported the inclusion of performance indices as part of a state’s accountability plan since NCLB was signed into law in 2001. As of 2007, the USDE had approved the use of index scores in calculations of AYP in 12 states (Erpenbach, 2009). Since then, limited and mixed evidence has been published regarding the impact of performance indices on states’ AYP determinations. In one evaluation of the impact of Pennsylvania’s Performance Index (PPI), Erpenbach (2009) noted that the percentage of schools meeting AYP was less when calculated using only the PPI (30.8%) than when AYP was determined using the Status model (45.3%). However, when PPI was used in combination with the status/percent proficient and the Safe Harbor (a policy that enables schools to meet AYP if they are able to evidence a ten percentage point decrease in the number of students not proficient in any particular area) models, the total number of schools meeting AYP increased to 74.4%. The situation appears only slightly more positive in New Hampshire. Here, Erpenbach (2009) reported that more than half of schools met proficiency index targets in both reading and math (61.3 and 55.8%, respectively); however, no comparison data were provided as to the number of schools meeting AYP under the status model.

Despite initial mixed results, few studies have attempted to empirically evaluate the effects of performance indexing on school-level student outcomes. Thus, this study proposes to evaluate the relationship between two AYP models (percent proficient and a proficiency index) and school growth. One of the advantages of using a performance index is its ability to credit schools for all increases in student performance, regardless of whether these increases occur above or below the proficiency threshold. Therefore, one might find that the use of a proficiency index is a better predictor of growth in schools than is the current percent proficient model. If this holds true, the application of performance indices might yield more valid indications of a school's performance and ability to raise student achievement than is currently feasible within a percent proficient model.

Statement of the Problem and Hypotheses

The No Child Left Behind Act (2001) has placed significant pressure on schools to increase student performance, with the overall expectation that 100% of students will be proficient by 2014. In addition, NCLB's disaggregation policy holds schools accountable for the achievement of individual student subgroups, including the students with disabilities subgroup. By holding schools accountable for the achievement of all children, and not just those in general education, policymakers hope to reduce achievement gaps between groups of students and ensure that all children have the opportunity to benefit from a high-quality education.

One of the main purposes of the disaggregation policy is to provide a mechanism for policymakers, school administrators, and the general public to monitor the performance of all students, including those who, like students with disabilities, have typically been excluded from large-scale assessments and accountability policies (Linn, 2000). However, researchers have raised concerns about the application of the disaggregation policy to the students with disabilities subgroup (Allbritten et al., 2004; Eckes & Swando, 2009; Koretz & Barton, 2003; M. Simpson et al., 2005; Thurlow, 2000; Ysseldyke & Bielinski, 2002). In particular, three issues have proven particularly problematic for producing accurate depictions of the achievement status of SWDs and the performance of schools that serve them: (a) the impact of reclassification on the composition of the special education subgroup, (b) the effect of small subgroup size, and (c) the effect of lower test scores, on average, for students with disabilities. Each of these issues has the potential to distort achievement outcomes for the special education subgroup, thereby jeopardizing the accuracy of inferences made about student and school performance.

The purpose of the proposed study is to empirically evaluate the effects of different variants of school-level disaggregated reporting policies on the students with disabilities subgroup. More specifically, this study proposes to assess school performance under three different policy variants: (a) the two-year post exit inclusion policy, which allows schools to include the achievement results of students who exited special education up to two years prior in calculations of AYP; (b) increased minimum subgroup sizes and the use of confidence intervals; and (c) the use of index scoring in AYP determinations.

Hypotheses

1. Schools will demonstrate an absolute change in the percent of students with disabilities reaching proficiency on large-scale assessments when the two-year post exit inclusion policy is applied.
    a. Schools will demonstrate an increase in the percent of students with disabilities reaching proficiency on assessments of reading.
    b. Schools will demonstrate an increase in the percent of students with disabilities reaching proficiency on assessments of math.

The first hypothesis will evaluate whether the implementation of the two-year post exit inclusion policy results in an absolute increase in the number of schools with students in the disability subgroup meeting proficiency. It is also important to assess whether this absolute change results in a subsequent increase in the number of schools meeting AYP for the special education subgroup. In other words, is the absolute increase observed enough to affect school AYP outcomes for the disability subgroup? To evaluate this question, two different sets of AMO targets will be used to assess whether changes in the percent of students with disabilities who are proficient results in changes in AYP outcomes for this group. The first set of targets corresponds to the AMOs set by the state of North Carolina for the 2009/10 school year in both reading (43.2%) and math (77.2%). These are referred to as "medium" level targets throughout the remainder of this document. During the 2010/11 school year, however, these targets increased sharply to 71.6% proficient in reading, and 88.6% proficient in math (hereafter referred to as "high' targets"). This increase corresponds to NCLB's mandate that all schools increase their proficiency targets until the 100% proficiency goal in both reading and math is attained in 2014. Thus, this study proposes to evaluate schools' ability to meet not only their actual 2009/10 medium level targets, but also the high targets that were implemented shortly thereafter.

2. A net increase in the number of schools meeting AYP targets for the students with disabilities subgroup will be observed when the two-year post exit inclusion policy is

applied to calculations of AYP, as compared to when the percent proficient or status model is used.

   a. An increase in the number of schools meeting the medium AYP target of 43.2% proficient in the area of reading/language arts will be observed.

   b. An increase in the number of schools meeting the medium AYP target of 77.2% proficient in the area of math will be observed.

   c. An increase in the number of schools meeting the high AYP target of 71.6% proficient in the area of reading/language arts will be observed.

   d. An increase in the number of schools meeting the high AYP target of 88.6% proficient in the area of math will be observed.

3. Compared to the policy of increasing minimum subgroup sizes, the use of confidence intervals in calculations of AYP will result in a net decrease in the number of schools meeting AYP for the students with disabilities subgroup when, in fact, they haven’t (i.e., there will be a decrease in false-negative reporting of AYP proficiency for the disabilities subgroup).
4. The type of AYP model used will have an impact on the relationship of school-level AYP outcomes and student achievement growth at the school.

   a. The application of a performance index will significantly predict school growth. Furthermore, it will be a stronger predictor of school growth than the application of the percent proficient/status policy.

Method

Data Source

The present study made use of extant data supplied by the North Carolina Department of Public Instruction (DPI) and housed by the North Carolina Education Research Data Center (NCERDC) at Duke University in Durham, North Carolina. NCERDC houses multiple datasets that can be accessed by researchers with common identifiers across datasets for schools and students. Two of these datasets were used in the present study: (a) individual - level accountability data files (AYP_ABC_PUB2010), including annual test scores for over a million students attending public schools in North Carolina between the mid-1990s and 2011, and (b) the NC school data contained in the Common Core of Data Public School Universe Survey (CCDPSU) database, a national statistical database of all public elementary and secondary schools and school districts. This study drew primarily on student- and school-level data from the 2007/08, 2008/09, and 2009/10 school years.

Participants

Students. This study used extant data from a sub-sample of children in grades 3-5 attending public elementary schools in North Carolina. Students' test results were included if they (a) had taken the general assessment in either reading or mathematics or both in 2009/10, and had been identified as a student with a disability according to the eligibility criteria outlined by the Individuals with Disabilities Education Act (IDEA) in the 2009/10 school year, (b) had reading or mathematics scores available in the NCERDC database for the general assessments in reading and mathematics, and (c) were attending a school with the appropriate grade levels. Approximately 123,000 students met these criteria, and were included as participants in the current study. Furthermore, participants for Hypotheses 1 and 2 also included an additional 21,369 students who had been identified as a student with a disability according to the eligibility criteria outlined by IDEA in either the 2007/08 or 2008/09 school year, but were no longer identified as a student with a disability.

In order for a child to be identified as a student with a disability under IDEA, he or she must (a) meet eligibility criteria for one of a number of specified disabilities, (b) have experienced adverse educational performance as a direct result of this disability, and (c) require special education or related services. In North Carolina, the Department of Public Instruction has outlined thirteen categories of disability: (a) autism, (b) deaf-blindness, (c) emotional disturbance, (d) hearing impairment, (e) mental retardation, (f) multiple disabilities, (g) orthopedic impairment, (h) other health impairment, (i) specific learning disability, (j) speech or language impairment, (k) traumatic brain injury, (l) visual impairment, and (m) hearing impairment. For the purpose of this study, the term “student with disabilities” referred to individuals who fell into any of these thirteen categories.

Schools. Approximately 2400 North Carolina schools are represented within the larger NCERDC database. For the purposes of this study, only a subsample of elementary schools with 3rd, 4th, and 5th graders were included. In addition, for elementary schools that met this criterion, only those with students that met the student participant eligibility requirements outlined above were included.

Tables 3 and 4 provide additional descriptive information about the population of students across schools in the current study. These tables provide information about how students were distributed within schools in the sample, under two different policies. Under the percent proficient policy, the average school had approximately 232 students in grades 3 to 5 in regular education, and 26 students in special education who had participated in the general education assessments in reading and mathematics. However, when the two-year post-exit inclusion policy was applied, students who had exited special education up to two years prior were counted as ‘special education’ students, instead of ‘regular education’ students. Under this policy, then, the average school had approximately 223 students in grades 3 to 5 in regular education, and 34 students in special education.

Table 3
Mean Number of Students in Different Student Groups across All Schools (n= 1111)

Student Populations	*M*	*SD*	Min	Max
Regular Education Students under the Percent Proficient Policy	231.32	95.96	0	726
Students Identified as ‘Student with Disability’ under the Percent Proficient Policy	25.55	12.69	0	73
Regular Education Students under the Two-Year Post-Exit Inclusion Policy	222.80	93.24	0	692
Students Identified as ‘Student with Disability’ under the Two-Year Post-Exit Inclusion Policy	34.06	15.49	1	97

Table 4

Mean Percent of Students in Different Student Groups across All Schools (n= 1111)

Student Populations	*M%*	*SD*	Min	Max
Regular Education Students under the Percent Proficient Policy	89.60	4.92	0	100
Students Identified as 'Student with Disability' under the Percent Proficient Policy	10.76	5.06	0	100
Regular Education Students under the Two-Year Post-Exit Inclusion Policy	86.19	5.32	0	98.94
Students Identified as 'Student with Disability' under the Two-Year Post-Exit Inclusion Policy	13.81	5.32	1.1	100

Overall, 1,111 public elementary schools were included in the final sample. To test Hypothesis 3, all schools in the sample were divided in to four categories, depending on the number of students in the students with disabilities subgroup. Table 5 summarizes the cumulative number of schools in each of the four subgroup sizes.

Table 5

Cumulative Frequency of Schools in Each of Four Minimum Subgroup Size Categories

(n=1111)

Subgroup Size Equal or Exceeding	N	Percent
10	183	16.47
20	762	68.59
40	1041	93.70
60	1111	100.00

The schools in the sample were drawn from 85 counties and 100 Local Educational Agencies (LEA) in the state of North Carolina. In addition, schools in the sample represented a range of locales, with approximately 29% of schools located in cities (i.e., located inside an urbanized area and principal city), 14.5% of schools located in suburbs (i.e., territory outside of a principal city but inside an urbanized area), and 21% of schools located in towns (i.e., territory inside an urban cluster but outside of an urbanized area). A further 46% of schools were located in urban areas (i.e., census-defined rural territory that is outside of an urban area and/or cluster).

Of the 1,111 schools in the sample, the majority of them (75.6%) were Title 1 School-wide Programs, defined as a school in which all of the pupils in the school are designated under appropriate state and federal regulations as being eligible for participation in programs authorized by Title 1 of the Improving America's Schools Act (Public Law 103-382). On average, 47% of students in each school were considered eligible to participate in the Free Lunch Program under the National School Lunch Act; furthermore, 9% of students (on average) met eligibility criteria for the Reduced-Price Lunch Program.

Measures

The primary outcome variable of interest in this study was student proficiency in math and reading, as assessed by student results on the End-of-Grade Test in Reading Comprehension (EOG-R) and the End-of-Grade Test in Mathematics (EOG-M). These assessments were designed to measure student performance against the competencies and skills outlined in the North Carolina Standard Course of Study. In Hypothesis 4, school growth was calculated and used as an outcome variable; in addition, a performance index was also developed and used as an independent variable. The following sections provide further information about these measures.

EOG-R. At the elementary school level, the EOG-R is designed to assess student knowledge and skills in the specific competency areas outlined in the North Carolina English Language Arts Standard Course of Study. In particular, the EOG-R aims to assess a child's ability to read and interpret text and apply strategies to comprehend and evaluate what has been read. The EOG-R requires students to read passages and answer multiple-choice questions about each passage. A total of ten passages are presented. Selected passages represent a range of different content areas to reflect the variety of reading content required of students, including literature passages, informational selections in particular content areas (e.g., science and social studies), and consumer and practical selections (e.g., brochures). The multiple choice questions are intended to tap four key constructs that are conceptualized as part of reading comprehension including (a) cognition, (b) interpretation, (c) critical stance, and (d) connections (PSNC, 2004, 2007).

The number of questions increases with grade level, and ranges from 56 questions in third grade to 68 in eighth grade. Scores on each grade level test are vertically linked. Student results on the EOG-R can be reported as percentiles or scaled scores. Scaled scores are subsequently used to classify a student's achievement according to one of four predetermined achievement levels. To be determined proficient in reading at his or her grade level, a student must demonstrate proficiency equivalent to a Level III or IV. Students at Level III "consistently demonstrate mastery of the grade level subject matter and skills and are well prepared for the next grade" (PSNC, 2004, p. 12).

Internal consistency reliability estimates for the EOG-R are presented as coefficient alphas, and range from .88 (tenth grade) to .94 (sixth grade). The standard error of measurement is two to six points for grades 3-8 (PSNC, 2004).

In the test manual, multiple sources of evidence are used to make the case that the EOG-R is a valid measure of students' reading comprehension skills and the states' language curriculum. Evidence of content relevance is demonstrated by matching each of the items to the strand or construct it is meant to assess (cognition, interpretation, critical stance, or connections). The distribution of items across each of these constructs is provided for each grade level. The percent of questions intended to assess cognition ranges from 25.9% in seventh grade to 39.3% in fourth grade. Similarly, the percent of questions assessing interpretation ranges from 36.7% in third grade to 42% in seventh grade. A smaller percentage of items are included to assess critical stance (18% in fourth grade to 26.8% in seventh grade) and connections (4% in fourth grade to 6.7% in third grade) (PSNC, 2004).

Evidence of criterion-related validity used for the EOG-R is determined by examining the relationship between student scores on the EOG-R and other measures of student achievement, including expected grades, assigned achievement levels, and teacher judgments of student achievement. Pearson correlation coefficients for the EOG-R and these measures of student achievement ranged from 0.49 to 0.65, indicating a moderate to strong correlation between scaled scores on the EOG-R and associated measures of student achievement (PSNC, 2004).

EOG-M. North Carolina's End-of-Grade assessment in mathematics is intended to evaluate students' performance in seven key areas included in the North Carolina Mathematics Standard Course of Study: numeration, geometry, patterns, pre-algebra, measurement, problem-solving, data analysis, statistics, and computation. The EOG-M consists of two parts administered separately. The mathematics computation section consists of 12 questions in grade 3-6 and eight questions for grades 7 and 8. The mathematics applications section of the EOG-M is comprised of 68 questions for grades 3-6, and 72 questions for grades 7 and 8. Students are allowed to use calculators during the mathematics applications section, but not during the mathematics calculations section of the EOG-M (PSNC, 2006, 2007).

As with the EOG-R, student results on the EOG-M can be reported as percentiles or scaled scores. In addition, scaled scores are used to determine a student's proficiency level using the achievement level classification system defined by the North Carolina Testing Program. Grade-level proficiency in math is evidenced by achieving either a Level III or IV on the assessment (PSNC, 2006).

Internal consistency reliability estimates for the EOG-M are presented as coefficient alphas, and range from .94 (tenth grade) to .96 (third grade). Similarly high reliability coefficients were found when analyzed by ethnicity, gender, and disability category. The standard error of measurement is two to six points for grades 3-8 (PSNC, 2006).

Evidence of content validity is provided for the EOG-M, and is demonstrated through teacher ratings. More specifically, content-area teachers were asked to evaluate the appropriateness of items on the EOG-M according to the following criteria: (a) test content reflects the goals and objectives of the grade level curriculum; (b) test content reflects the goals and objectives of the grade level curriculum as it is taught in the teacher's school or school system; (c) items are clearly and concisely written, and the vocabulary is appropriate to the target age level; (d) the content is balanced in relation to ethnicity, race, sex, socioeconomic status, and geographic districts of the state; and (e) each of the items has one and only one answer that is best; however, the distracters appear plausible to someone for someone who has not achieved mastery of the represented objective. Responses were reported on a 5-point scale, with the "5" representing "to a superior degree", and "1" corresponding to "not at all." On average, teacher responses to these questions indicated that the EOG-M met the criteria outlined to a "superior" or "high" degree (PSNC, 2006).

Evidence of criterion-related validity for the EOG-M was determined by examining the relationship between student scores on the EOG-R and other measures of student achievement, including expected grades, and assigned achievement levels. Pearson correlation coefficients for the EOG-M ranged from 0.49 to 0.89, indicating a moderate to strong correlation between scaled scores on the EOG-M and associated measures of student achievement. In addition, evidence of concurrent validity was further demonstrated by comparing trends in student performance between the EOG-M and students' progress on the National Assessment of Education Progress (NAEP). Similar trends in performance for students scoring “basic” or “proficient” on the NAEP and students who scored at Level III or IV on the EOG-M were observed across 4th and 8th grade students (PSNC, 2006).

Test administration and participation. Both the EOG-R and the EOG-M are administered in May of each year to students in grades 3-8 as part of North Carolina’s statewide assessment system. Training on the proper administration of the End-of Grade tests is provided to school test coordinators according to the training specifications outlined in the NCDPI Testing Policy. School test coordinators are responsible for monitoring test administration within a school and responding to any questions or concerns that might arise on the day of testing. Under the supervision of the school test coordinator, school employees are permitted to administer tests to group of students and are responsible for ensuring that correct test security and confidentiality is maintained throughout the administration process.

As per both federal and state accountability policies, students in grades 3-8 are required to participate in the EOG tests, or alternative assessments of reading and mathematics achievement. Some students, including students with disabilities and students with limited English proficiency, may receive testing accommodations; however, the need for these accommodations must be evidenced through appropriate documentation. In addition, a small percentage of students may be excused from participating in EOG tests due to the presence of significant medical emergencies and/or conditions. Only student outcomes on the North Carolina End-of Grade assessments were considered in this study; student outcomes on the two alternate assessments were not included.

School growth. North Carolina uses a variant of a residual gain score as a measure of individual student growth, and the mean of student growth scores at a school as a measure of school growth. Student scores in reading and mathematics are converted to c-scores, which are z-scores based on the grade-level mean and standard deviation obtained by students in the standard setting year. Within this system, academic change is expressed as the difference between a student's c-scale score in the current year and the average of the student's previous two assessments, in c-scale units, with a correction for regression to the mean (PSNC, 2006a). The formula for determining academic change is as follows:

AC = CSc-scale - (0.92 X ATPAc-scale), where

AC = academic change

CS= current score

ATPA= average of the two previous assessment scores

When only one previous score is available for a student, that c-scale score is multiplied by .82 and subtracted from the current score.

To determine academic change at the school level, the mean of all student academic change scores is calculated separately for reading and math. At the school level, a change of "0" indicates that students, on average, made as much growth as expected and maintained their ranking relative to the students in the standard setting year. A negative change score indicates that students made less growth than predicted and fell in ranking relative to students in the standard setting year (PSNC, 2006a).

Performance index. For the purposes of the current study, a performance index was developed to allow for the assignment of partial credit to students at varying levels of proficiency. As illustrated in Figure 4, this performance index consists of four proficiency levels, each with a corresponding weight. To calculate a school's overall performance index score (PI), the percentage of students at each level is multiplied by the corresponding weight, resulting in a score for each performance level. These scores are then summed to get the school's overall PI score. Figure 4 provides an example of how school-level PIs were calculated in this study. Under this policy, the maximum score that a school can receive is 120 (100% of students reaching the advanced level), and the minimum possible score is 0 (all students not assessed). As North Carolina does not currently use a performance index when determining AYP calculations, the index developed for the purposes of this study was modeled after the performance index currently used in the state of Ohio (Ohio Department of Education, 2006).

Performance Index			
Proficiency Level	**% of Students at Level**	**Weight**	**Performance Level Score**
Advanced	10	1.2	12
Proficient	40	1.0	40
Basic	35	0.6	21
Below Basic	12	0.3	3.6
Not Assessed	3	0	0
School-Level Performance Index Score:			76.6

Figure 4. Example of how a performance index can be used to calculate a school-level performance index score. The proficiency levels and weights presented in this example are equivalent to those used in the performance index developed for the current study.

Results

Analyses Specific to Hypotheses

Hypothesis 1. As described in the previous section, Hypothesis 1 asserted that the implementation of the two-year post exit inclusion policy would result in an absolute increase in the percent of students with disabilities at a school meeting grade-level proficiency standards compared to the proficient/status model. To test this hypothesis, the percent of students with disabilities meeting proficiency on the EOG-R and EOG-M (defined as Achievement Standard Level III or above) was calculated for each school in the sample, first, only for students who were in special education during the 2009/10 school year, and then again including students who had exited special education during the 2007/08 and 2008/09 school years.

A binomial test (sign test) was used to evaluate whether the change in who was included in the students with disabilities subgroup increased the percent proficient for this subgroup at the school level. The binomial test analyzed the proportion of schools falling within one of two conditions: (a) an increase in the percent of students with disabilities reaching proficiency on large-scale assessments under the two-year post exit inclusion policy, or (b) a decrease in the percent of students with disabilities reaching proficiency on large-scale assessments under the two-year post exit inclusion policy. Schools that showed no difference in the percent of students reaching proficiency under each of the two policies were excluded from the analysis. The null hypothesis (H0 : p= .50) assumed that there was no difference between conditions; thus, by chance alone, one would expect half of the sample of schools to show an increase in the percent proficient and half to show a decrease when the two-year post exit inclusion policy was applied.

For the EOG-R, 952 (85.7%) of the 1,111 schools saw an increase in the percent of students with disabilities meeting grade-level proficiency standards when the two-year post exit inclusion policy was applied. Similarly, 961 (86.5%) of the 1,111 schools saw an increase in the number of students with disabilities meeting grade-level proficiency standards on the EOG-M under this same policy. In both cases, the number of schools showing increases in the percent proficient was greater than would be expected by chance ($p < .001$), confirming Hypothesis 1.

To determine the extent to which the average school increased its proportion of students with disabilities reaching grade-level proficiency standards when the two-year post-exit inclusion policy was applied, the average percent proficient for the students with disabilities subgroup was calculated under both policies. As indicated in Table 6, a 6-7 percent increase in the number of students with disabilities meeting grade-level proficiency standards was observed when the two-year post-exit inclusion policy was applied.

Table 6

School Mean Percent of Students with Disabilities who Reached Proficiency in Mathematics and Reading Under Two Policies (n = 1111)

	Reading/Language Arts		Mathematics	
Policy	*M*	*SD*	*M*	*SD*
Percent Proficient/Status	40	21	59	20
Two-Year Post-Exit Inclusion	47	19	65	18

Hypothesis 2. The results from Hypothesis 1 indicated an overall significant increase in the percent of students with disabilities meeting proficiency on assessments of math and reading when the two-year post-exit inclusion policy was applied. However, for the purposes of AYP, it is perhaps most important to determine whether this change resulted in an increase in the number of schools meeting annual measurable objective (AMO) targets in both reading and math. Binomial sign tests were again used to test Hypothesis 2, this time examining whether the proportion of schools meeting two sets of AMO targets ('medium' and 'high') increased under the two-year post-exit inclusion policy. With both the 2009/10 'medium' AMOs (i.e., 43.2% proficient in reading/language arts and 77.2% proficient in mathematics) and 'high' 2010/11AMOs (i.e., 71.6% proficient in reading/language arts and 88.6% proficient in mathematics), the number of schools meeting AMO targets increased (all p's <

.001) when the two-year post exit inclusion policy was applied (see Table 7), thereby confirming Hypothesis 2.

Table 7

Number of Schools Meeting Medium and High AMO Targets for the Students with Disabilities Subgroup in Reading/ Language Arts and Mathematics (n = 1111)

	Assessment Domain	
Policy	Reading/Language Arts	Mathematics
"Medium" Targets		
Percent Proficient/Status Only	436	213
Two-Year Post-Exit Inclusion	615*	292*
"High" Targets		
Percent Proficient/Status Only	90	67
Two-Year Post-Exit Inclusion	129*	96*

** Increase in number of schools p< .001*

Note. "Medium Targets" refer to the 2009/10 North Carolina AMOs (43.2% proficient in reading/language arts and 77.2% proficient in mathematics); "High Targets" refer to the 2010/11 North Carolina AMOs (71.6% proficient in reading/language arts and 88.6% proficient in mathematics).

Hypothesis 3. Hypothesis 3 asserted that, compared to the minimum required subgroup size policy, the use of confidence intervals in calculations of AYP would result in a decrease in false-negative reporting of AYP proficiency results for the students with disabilities subgroup. Testing Hypothesis Three required multiple steps to prepare the data and then test the hypothesis. Each of these steps is described here, with results presented in tables in the text or in Appendices A and B.

The first step in testing Hypothesis 3 was to determine the percentage of schools meeting the 2009/10 North Carolina AMOs for the students with disabilities subgroup (43.2% proficiency in reading/language arts, and 77.2% proficiency in math) when different policies regarding how these results should be calculated were applied. Three different policy alternatives were applied: (a) using the percent proficient/status policy; (b) using the percent proficient policy with confidence intervals placed around each school's results; or (3) only requiring schools to compare the percent of students with disabilities reaching proficiency to an AMO criterion if the school's subgroup size exceeded a minimum subgroup size. For the last policy alternative, 4 minimum subgroup sizes were examined: 10, 20, 40 and 60.

For the percent proficient/status policy alternative calculations, the percent of students with disabilities performing at or above the proficient level on assessments of reading and math was computed for each school in the sample and compared against the AMO criterion. The number of schools meeting or surpassing the AMO was then summed. The first row of values in Table 8 represents the percent of the 1,111 schools in the sample that met AYP for reading and mathematics under this policy alternative. For the confidence interval policy calculations, one-sided (upper tail) confidence intervals (α=0.05) were calculated and placed around the percent proficient value obtained at each school. A school was considered to have met the state AYP when the 95% confidence interval around the percent proficient figure included the required AMO value, regardless of whether the school would have met AYP without the confidence interval. The second row of values in Table 8 represents the percent of schools that met AYP under this policy alternative. Finally, to calculate the number of schools meeting AYP under four different minimum required subgroup sizes (10, 20, 40 and 60), the percent of students with disabilities performing at the proficient level was calculated and compared to state AMO's only for those schools that met or surpassed the minimum required subgroup size. All other schools (i.e., any school with too few students in the students with disabilities subgroup to constitute a subgroup) were deemed as having met AYP for the subgroup. The results of these calculations are presented in the last four rows of Table 8.

Table 8

Percent of Schools Meeting 2009/10 AMO Targets for the Students with Disabilities Subgroup when AYP is Calculated Using Three Different Policies (n=1111)

	Assessments	
Policy	Reading/Language Arts	Mathematics
Percent Proficient/Status	39.24	19.17
Confidence Intervals	82.80	52.34
Minimum Subgroup Size		
10	56.08	27.54
20	62.83	37.26
40	88.30	77.22
60	98.74	96.04

The second step in preparing the data to test Hypothesis Three was the construction of 2x2 contingency tables that compared school-level AYP results under both the confidence interval policy and the minimum required subgroup size policy. This step was completed separately for each of the four different subgroup sizes (10, 20, 40, and 60 students) for reading and then math (see Appendix A). In each table, school pass and fail status, as determined after the application of confidence intervals, was treated as a school's "true" AYP status, and the school's results calculated with each of the minimum subgroup sizes was treated as an "obtained" result, where that result contained error as a consequence of applying the minimum subgroup size policy. Thus, the extent to which the minimum subgroup size policy, intended to guard against incorrect portrayals of schools' performance (as well as preserve student confidentiality), actually resulted in incorrect portrayals of a school’s subgroup performance was determined. Two types of incorrect portrayals of a school were possible: false negatives where a school is considered to have met its AYP goal for the subgroup when it did not, and false positives, where a school is considered to have not met AYP when it actually did when confidence intervals were applied. Although false negatives are the focus of Hypothesis 3, the tables in Appendix A provide the number of both false negatives and false positives, with the number of false-negative results in the lower left quadrant of each of the contingency tables, and the number of schools with false positive results in the upper right quadrant.

The last step in testing Hypothesis 3 was to calculate the false negative rate under different minimum subgroup sizes by dividing the number of false negatives in each table by the sum of the true and false negatives. As can be seen from examining Tables 9 and 10, with the smallest minimum subgroup size of 10, the false negative rate approaches 0, but each increment in minimum subgroup size increases the false negative rate. In fact, the false-negative error rate for the largest subgroup (60 students) is approximately 96 percentage points larger than the value for the smallest subgroup size (10 students).   By definition, the false negative rate when confidence intervals are applied is 0, as these results are treated as the “true” results for each school. Although the false-negative error rate for the minimum required subgroup policy varies depending on the specific subgroup size used, it is, in all cases, higher than the false-negative error rate for the confidence interval policy. Thus, Hypothesis 3 is confirmed.

Table 9

False Negative and False Positive Error Rates at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Reading/Language Arts

Subgroup Size	False Negative	False Positive
10	1	32.5
20	14.1	27.1
40	71.7	8
60	96.9	0.9

Table 10

False Negative and False Positive Error Rates at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Mathematics

Subgroup Size	False Negative	False Positive
10	0.9	48.3
20	12.1	39.9
40	68.2	14.6
60	94.9	2.9

To further explore the impact of changes in minimum subgroup size on the accuracy of decisions about the disability subgroup’s performance relative to AYP, sensitivity and specificity values (see Table 1 for formulas) were calculated at each minimum subgroup size and are presented in Appendix B. These values were then used to plot receiver operating characteristic (ROC) curves for reading and math. These curves graphically depict the impact of changes in the minimum subgroup size on the accuracy of portrayals of the disability subgroup’s performance (see Figures 5 and 6).

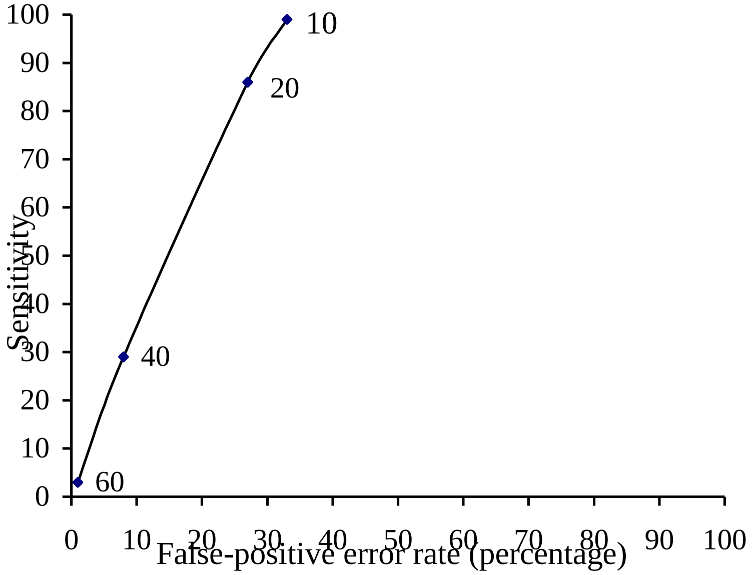

Figure 5. Receiver operating characteristic (ROC) curve illustrating the sensitivity and false-positive error rate of the minimum required subgroup policy at various subgroup sizes. Calculations of sensitivity and false-positive error rate are based on the number of schools in the sample meeting AYP for reading/language arts.

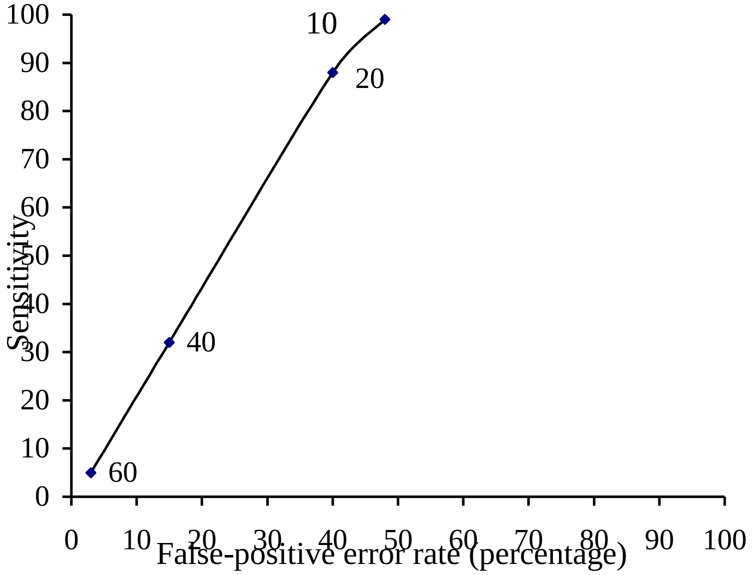

Figure 6. Receiver operating characteristic (ROC) curve illustrating the sensitivity and false-positive error rate of the minimum required subgroup policy at various subgroup sizes. Calculations of sensitivity and false-positive error rate are based on the number of schools in the sample meeting AYP for mathematics.

Sensitivity is the inverse of false negatives, and as indicated in these figures, although the false positive error rate increases somewhat as the minimum subgroup size decreases, it is only at the 10 and 20 minimum subgroup sizes that the trade-off between sensitivity and false positives results in data points on the ROC curves that fall near the “good” or “excellent” curves depicted in Figure 1.

Hypothesis 4. Hypothesis 4 asserted that the variables used as the basis for determining a school’s AYP status under two different policy alternatives (percent proficient and a performance index) would each be related to school growth for students with disabilities in reading and mathematics, although the performance index would be a better predictor. To test this hypothesis, a linear hierarchical regression strategy was used with school growth as the outcome variable and the two AYP variables entered sequentially in separate regression equations, with a test for whether the increase in variance accounted for was significant with the addition of the index score in the second step.

Prior to running the regression analyses necessary to test the hypothesis, descriptive analyses were conducted to examine whether the variables used met the assumptions of hierarchical linear regression (e.g., normality, collinearity). As outlined in Table 11, this analysis included a calculation of each variable's mean, standard deviation, minimum and maximum value. Additionally, Table 12 summarizes each variable's skew and kurtosis. In each instance, the significance of both the skew and kurtosis value was calculated by converting each score to a z-score, and comparing the resulting value to 1.96 (Field, 2009). Based on these calculations, significant kurtosis values were found for each of the dependent variables. Although significant kurtosis values typically violate the assumption of normality, these results were assumed to result from the study's large sample size (n=1,111) rather than extreme skewness; thus, these data were not transformed (Field, 2009). Skewness data were also found to be significant for two of the independent variables (Percent Proficient/Status Policy for Reading/Language Arts and Performance Index for Mathematics) using a 95% significance criterion. However, as regression only assumes normality of data for dependent variables, these independent variables were not transformed.

In addition, a correlation matrix was generated to examine the relationships between all independent variables (Percent Proficient/Status policy and Performance Index) and the dependent variables (school growth in reading and mathematics). Results of these correlations are presented in Tables 13 and 14. Significant, positive correlations were found between the dependent variables, school growth in reading and mathematics, and both the percent proficient of students reaching proficiency and the performance index. In addition, both independent variables were also found to be significantly and positively correlated with each other.

Table 11
Distribution of Variables

Variable	*M*	*SD*	Min	Max
Independent Variables				
Status Policy (Reading/Language Arts)	.40	.21	.00	1.00
Status Policy (Mathematics)	.59	.20	.00	1.00
Performance Index (Reading/Language Arts)	.72	.13	.34	1.20
Performance Index (Mathematics)	.86	.11	.50	1.14
Dependent Variable				
School Growth (Reading/Language Arts)	.04	.19	-1.02	1.02
School Growth (Mathematics)	.11	.21	-.72	.85

Table 12
Skew and Kurtosis of Variables

Variable	Skew	Kurtosis
Independent Variables		
Status Policy (Reading/Language Arts)	.43**	-.21
Status Policy (Mathematics)	-.09	-.44**
Performance Index (Reading/Language Arts)	.13	-.35*
Performance Index (Mathematics)	-.19*	-.35*
Dependent Variable		
School Growth (Reading/Language Arts)	.08	1.89**
School Growth (Mathematics)	.03	.50**

Note. * $p \leq 0.05$, ** $p \leq 0.01$

Table 13
Correlation coefficients of relations between independent and dependent variables based on Reading/Language Arts achievement results (n = 1111)

Variable	School Growth	Percent Proficient	Performance Index
School Growth	1.00	.37**	.42**
Percent Proficient		1.00	.90**
Performance Index			1.00

Note. ** $p \leq 0.01$

Table 14
Correlation coefficients of relations between independent and dependent variables based on Mathematics achievement results (n=1111)

Variable	School Growth	Percent Proficient	Performance Index
School Growth	1.00	.40**	.42**
Percent Proficient		1.00	.92**
Performance Index			1.00

Note. ** $p \leq 0.01$

As indicated in these tables, significant, positive correlations were found between the dependent variable for each planned analysis of school growth (in reading or mathematics), and the independent variables, percent of students reaching proficiency, and the performance index (in reading or mathematics). In addition, the independent variables for the two planned analyses were also found to be significantly and positively correlated with each other, indicating substantial multicollinearity. The high level of collinearity between the independent variables was not surprising, given that both the percent proficient/status and performance index variables were ways of reporting the same test results; nevertheless, the issue of multicollinearity has the potential to pose significant problems for the regression analysis originally proposed in this study (Field, 2009). Thus, a test of the difference between dependent correlations was conducted instead to test Hypothesis 4. This test examines whether the correlation between one predictor variable and an outcome variable is significantly different than the correlation between a second predictor and the same outcome variable. When using this statistical test, Hypothesis 4 would be confirmed if the correlation between the outcome variable, growth, and the performance index would be significantly higher than the correlation between the same outcome variable and the other predictor, percent proficient. The test uses the following formula to calculate the difference between two correlations from the same sample; the value obtained can then be checked against the critical values of the t-distribution to determine whether it is statistically significant (Field, 2009):

$$t_{Difference} = (r_{xy} - r_{xz})\sqrt{\frac{(N-3)(1+r_{xz})}{(1-r_{xy}^2-r_{xz}^2-r_{zy}^2)+(2r_{xy}r_{xz}r_{zy})}}$$

To test Hypothesis 4, the correlation between the status index and school growth was subtracted from the correlation between percent proficient and school growth for language arts, and then the same test conducted for mathematics. The difference between the two correlations was statistically significant for both language arts (*t*(1108)= -5.79, *p* < .01) and math (*t*(1108)= -2.66, *p* <.01). In both cases, the correlation between school growth and the performance index was significantly higher than the correlation between school growth and percent proficient. As such, Hypothesis 4 was confirmed.

Discussion

The No Child Left Behind Act (2001) requires schools to report on the performance of student subgroups, with the aim of focusing attention on subgroups of students who have historically been at risk for low achievement. The application of this disaggregation policy to the students with disabilities subgroup is problematic, as research suggests that it may lead to inaccurate information regarding the achievement and proficiency of the students in this subgroup (Cole, 2006; Mintrop & Sunderman, 2009; M. Simpson, Gong, & Marion, 2005; Ysseldyke & Bielinski, 2002). The purpose of the current study was to empirically evaluate the effects of different variants of school-level disaggregated reporting policies on the students with disabilities subgroup.

The following discussion will begin with a summary of this study's findings as they relate to each of the four hypotheses proposed earlier in this document and the previous literature examining school outcomes for students with disabilities. The next section discusses the implications of these results, especially in regards to the impact of various school-level reporting policies on schools' ability to meet AYP for the students with disabilities subgroup. Finally, this document will conclude with a discussion of the study's limitations and recommendations for future research.

Summary and Discussion of Findings

The effect of reclassification (Hypotheses 1 & 2). The first aim of the current study was to evaluate the effect of North Carolina's two-year post-exit inclusion policy on the percent of students with disabilities obtaining large scale test scores that fell in the proficient range or above. A second, but related, aim was to evaluate the two-year post-exit inclusion policy's effect on schools' ability to meet both medium and high AMOs set by the state of North Carolina. Results confirm findings from past studies illustrating the advantage of expanding special education subgroup membership for AYP purposes to include students who have recently exited special education (Parker, 2001; Ysseldyke & Bielinski, 2002). When the increased performance of these additional students is accounted for, the overall achievement of the special education subgroup significantly increases, thereby bringing schools closer to annual measurable objectives and making it easier for schools to meet AYP targets.

Despite overall positive effects of the two-year post-exit inclusion policy on school-level outcomes, it is noteworthy that the number of schools meeting AYP targets is still highly dependent on the specific AMO set by the state. Although more schools were able to meet both medium and high AMOs under the two-year post-exit inclusion policy than the percent proficient policy, the proportion of schools meeting AMOs (under either policy) relative to the total number of schools in the sample grows smaller as AMOs approach 100%. Thus, although implementation of the two-year post-exit inclusion policy greatly improves schools' ability to meet AYP, its overall effect may be limited as AMOs increase over time.

Comparisons of minimum subgroup size and confidence intervals calculations on the accuracy of AYP results (Hypothesis 3). The second aim of the study was to examine how two policy alternatives, both intended to guard against the impact of small sample sizes on school-level results, functioned in terms of accurately representing the performance of the students with disabilities subgroup. One such strategy was the use of confidence intervals. With this policy alternative, the extent to which sampling error may have distorted results is estimated, and then taken into account in determining whether or not a school met a particular AYP criterion. Any school where the confidence interval around the obtained percent proficient includes the state's yearly AMO is considered to have met the criterion. In this study, the use of confidence intervals in determining AYP was considered the "gold standard" or most accurate approach for dealing with sampling error, as it took into account the school's actual performance and the influence that sampling error was likely to have had on the school's obtained results.

The second policy alternative for adjusting for sampling error is the minimum required subgroup policy. With this policy, if a school does not meet an AYP criterion, and the school's students with disabilities subgroup falls below a minimum subgroup size, the school is still considered to have met that AYP target. The minimum required subgroup size policy assumes that any calculation of AYP based on small sample sizes is statistically unreliable, given the large amount of error inherent within these results. Therefore, the overall effect of this policy is, essentially, to give schools with small subgroup sizes the benefit of the doubt and automatically allow them to meet AYP targets, regardless of what percentage of students in the subgroup actually meet proficiency.

Although the minimum required subgroup policy has the potential to increase the reliability and validity of school-level results, some have argued that its implementation also results in inaccurate information regarding school quality and student achievement, as well as a general disregard of NCLB's accountability requirements (M. Simpson, 2005). Regarding the former, results of the current study confirm that the reporting of false-negative results increases as subgroup sizes get larger. In other words, as the minimum required subgroup size gets larger, so, too, does the percentage of schools reported as having met AYP targets when, in fact, they have not. These results are commensurate with those reported by M. Simpson (2005). In comparison, the false-negative error rate for the confidence interval policy was zero, regardless of subgroup size. Thus, results of the current study suggest that although both policy alternatives are effective in reducing the influence of sampling error on school-level results, only the confidence interval policy is able to do so without simultaneously compromising the accuracy of the information provided on student performance.

To further explore the effect of the minimum required subgroup size policy on the accuracy of the information reported on student performance, this study also used ROC curves to visually analyze the sensitivity and specificity of this policy at different subgroup sizes. As has been noted extensively in epidemiological research, analyzing the accuracy of a medical screening tool or, in this case, an educational policy, in this way assists in the determination of the most appropriate minimum subgroup size, based on the ratio of true versus false results. In the case of the minimum required subgroup size policy, relatively small subgroup sizes of 10 or 20 appear to have the best balance of sensitivity and specificity, as compared to the larger subgroup sizes. More specifically, a subgroup size of 20 will correctly identify a school that meets its reading/language arts AYP target 68% of the time, and will correctly identify a failing school 86% of the time. Although larger subgroup sizes have a higher specificity and are better able to correctly identify passing schools 92-99% of the time, their sensitivity is strikingly low. Although other authors have similarly discussed the effect of the minimum required subgroup size policy on the accuracy of school-level reporting, this study is the first to (a) evaluate the accuracy of an educational policy using methods borrowed from the epidemiological and medical literature, and (b) to illustrate, using ROC curves, the relationship between specificity, sensitivity, false-positive, and false-negative error rates for this policy at various minimum subgroup sizes.

The use of index scoring in AYP calculations. The third aim of the present study was to explore whether the use of a performance index resulted in a measure of school performance that was more closely related to student growth, as compared to the current percent proficient model used for determining AYP. As outlined earlier in this document, the fourth hypothesis stated that not only will the application of a performance index be a significant predictor of school growth, but it will be a stronger predictor of school growth than the percent proficient/status model. Although the analysis for this hypothesis could not be conducted as originally proposed, results of the test of dependent correlations confirmed that the correlation between the performance index and school growth was higher than the correlation between the percent proficient policy and school growth.

Implications for Policy and Practice

Although heavily criticized, the practice of looking to standardized assessments as primary measures of student performance, teacher effectiveness, and even school quality continues to dominate educational policy, especially under NCLB. Given the influence associated with large-scale assessment results, it is even more important that policymakers and school administrators understand how to use these data to draw the most valid and reliable conclusions regarding student performance. Results of the current study indicate that the implementation of current policies, particularly those designed to improve measures of special education subgroup performance, may have varying implications on the reporting of special education subgroup assessment results.

The two-year post-exit inclusion policy. One of the goals of the recently implemented two-year post-exit inclusion policy is to enable schools to account for the transitory nature of the special education subgroup. The challenges associated with frequent transitions are unique to the students with disabilities subgroup; in fact, no other subgroup is characterized by fluctuations in subgroup membership to the same degree as the special education subgroup. As has been illustrated by this study, as well as in previous research, one of the primary ways to account for such transitions is to permit schools to utilize the achievement scores of students who have exited special education in determining the performance of the students with disabilities subgroup. This strategy results in more stable group membership for the students with disabilities subgroup, countering the potential for a downward bias in measuring the performance of the students with disabilities subgroup caused by the exit of higher achieving students from special education.

Thus, one important function of the two-year post-exit inclusion policy is to help ‘level the playing field’ between the special education subgroup, and other subgroups within a school. In addition, allowing schools to include the achievement scores of students who have recently exited special education in the school’s AYP calculations for the special education subgroup acknowledges the important contribution of special education services on those students’ advancement. In other words, if the special education services received by a student directly help to increase his or her achievement, then it is only fair that these increases in achievement be included in subgroup measures of student performance.

Despite positive effects on subgroup performance, the implementation of the two-year post-exit inclusion policy is not without limitations. As illustrated in this study, this policy does increase the number of schools meeting AYP for the students with disabilities subgroup; however, this positive effect is diminished as AMO’s increase. Furthermore, an increase in AMOs is inevitable, as NCLB mandates that all states steadily increase their AMOs until 100% of students meet grade-level proficiency standards by 2014. At the time of writing, the 2014 deadline is only one year away; therefore, most states will be attempting to meet relatively high performance targets. Thus, results of the current study suggest that the adoption of the two-year post-exit policy at this point may pose little benefit to schools, depending on how high states set their AMOs.

Although the current study provides verification of the two-year post-exit inclusion policy's overall effectiveness, other important questions regarding its implementation remain. For example, no research has yet evaluated the impact of including students in the students with disabilities subgroup who have exited special education more than two years prior. In fact, this demarcation (i.e., including students who exited two years prior, as opposed to three or more years) is entirely arbitrary; however, it has a significant bearing on the overall performance of the subgroup. Including students who exited special education up to four years prior, for example, would include more students with increased achievement scores, thereby serving to increase overall subgroup achievement results even more. However, is it fair to operationally define special education group membership in this way, or is it more appropriate to consider students who have not received special education services in four years 'general education students' for AYP purposes? Perhaps more specifically, is it appropriate to attribute this student's achievement to the special education services he or she received four years prior? As noted by Parker (2011) in his review of the effectiveness of special education services, decisions regarding special education subgroup membership (such as how many years prior a student may have exited special education and still be counted as 'special education' for AYP purposes) are closely tied to society's views regarding the overall benefit of special education.

Increasing the required size of the special education subgroup. Like the two-year post-exit inclusion policy, the practice of increasing the required size of the special education subgroup also resulted in a dramatic increase in the number of schools meeting AYP for the students with disabilities subgroup. In fact, the effect of increasing required subgroup sizes is such that when the minimum subgroup size equals 60, nearly all schools in the sample met AYP in both reading/language arts and mathematics for the students with disabilities subgroup. However, results of the current study also indicate that many of the schools meeting AYP when minimum subgroup size is large may not actually be meeting proficiency targets; instead, they are only able to meet AYP because the larger minimum subgroup size excludes them from accountability measures entirely. Thus, results of the current study validate the concerns originally raised by M. Simpson (2005) that implementation of the minimum subgroup size policy may have negative implications on the inclusion of students with disabilities in school-wide accountability policies. In fact, the implementation of this policy actually appears to counteract one of the main goals of the disaggregation policy; namely, to ensure that the performance of all students, including those with disabilities, is accurately reported. It is important to note that results of the current study cannot confirm that schools are, in fact, using the minimum subgroup size policy for the purpose of circumventing accountability requirements. However, it is clear that this policy does provide a “loophole” for schools, and the motivation for schools to use it may increase as pressure to meet all of NCLB’s mandates increases.

Thus, the task of selecting an appropriate subgroup size is complicated by a number of factors, including statistical unreliability and the potential for the exclusion of students with disabilities from accountability policies. In addition, this study confirms that additional concerns regarding the minimum required subgroup size's effect on the accuracy of reported AYP results are also warranted. Based on methods borrowed from the epidemiological literature, however, this document has illustrated an alternative way of determining the accuracy and precision of the required minimum subgroup size policy at various subgroup sizes based on rates of specificity, sensitivity, false-positive and false-negative error rates. Using this information, it is possible to predict the percentage of schools that will be correctly identified as either passing or failing AYP, as well as the percentage of schools that will be incorrectly labeled. For example, based on the information in the ROC curve in Figure 5, it is possible to determine that the implementation of a minimum subgroup size policy with a minimum subgroup size of 20 will result in the correct identification of failing schools 86% of the time; inversely, this also means that the remaining 14% of schools will be incorrectly identified as passing when, in reality, they have actually failed to meet AYP.

This emphasis on the ratio of "truth" to error is similar to the kind of cost/benefit analysis often used by economists and government agencies to analyze the desirability of a given policy based on the expected balance of benefits and costs. It is important to note, however, that this seemingly objective method of policy analysis still necessitates a degree of subjective decision-making; after all, once the respective sensitivity and false-positive error rate of each minimum subgroup size is known, it is up to individual policymakers to determine what particular ratio of truth to error is acceptable. Reverting to the example above, is it socially and ethically admissible for 14% of elementary schools to be incorrectly labeled as meeting AYP targets? In practice, this means that parents, teachers, and administrative staff in 14% of North Carolina's schools will be given false information regarding the overall performance of their students. Would adopting a minimum subgroup size with a lower false negative rate be more desirable, even if it also means decreasing the sensitivity as well? Decisions such as these have no simple, clear-cut answers; however, it's of utmost importance that policymakers fully understand the statistical and practical implications of adopting various subgroup sizes as part of the minimum required subgroup size policy.

Thus far, this document has solely evaluated the minimum required subgroup size policy's main goal of reducing the susceptibility of small subgroups (such as is often typical of the special education subgroup) to various sources of volatility (e.g., sampling error) and encouraging more statistically reliable results. However, this policy serves a second purpose: namely, to help ensure the confidentiality of individual students in a particular subgroup when that subgroup's total membership is low. In instances in which the membership of a subgroup is too low, the analysis or reporting of student scores may jeopardize the confidentiality of the individuals in that subgroup; thus, one goal of the required minimum subgroup size policy is to allow schools to forgo separate reporting of these subgroups to prevent confidentiality breaches. For this purpose alone, the minimum subgroup policy appears to achieve its goal; the identity of individual students is certainly better protected when the membership of the subgroup is required to be of a certain number. However, problems arise when policymakers attempt to address both of these goals with one policy. As results of the current study illustrate, the success of this policy in ensuring confidentiality may be greatly overshadowed by the many other problems it produces, especially in regards to the inclusion of students with disabilities in accountability systems.

The use of confidence intervals in calculations of AYP. Compared to the minimum subgroup size policy, the practice of using confidence intervals in AYP calculations provides a more accurate and representative indication of the special education subgroup's actual performance, while still enabling an increased number of schools to meet AYP targets compared to the current percent proficiency model. As one example, using confidence intervals when calculating the percent of students in a subgroup meeting performance targets accounts for possible sources of volatility (such as the problem of small subgroup sizes) that other policies, such as the percent proficient/status policy, do not. In addition, the calculation of a confidence interval is based on the amount of error likely to be found in a group of students of a certain size; thus, compared to policies that rely on arbitrary cut-offs (like the percent proficient policy), the policy of using confidence intervals in calculations of AYP will yield more accurate results. However, although the use of confidence intervals may potentially be a more 'truthful' representation of subgroup performance, it is not a perfect policy. By definition, a 95% confidence interval around the percent of students meeting proficiency indicates the limits within which we are certain to find the true percent proficient 95% of the time. If the state AMO also falls within this range, the school is considered to have met AYP. However, it is also acknowledged that there is a 5% chance that this AYP determination is wrong, and that a school is considered to have met AYP when, in fact, it hasn't. Like the schools that automatically meet AYP as the minimum subgroup size increases, these schools are also considered "false negatives." Thus, the policy of applying confidence intervals when calculating AYP does not eradicate the problem of false negatives; however, it does greatly reduce it. Furthermore, it is possible to control the level of false negatives by altering the confidence interval used (e.g., a 99% confidence interval will only yield false negatives 1% of the time, compared to a 95% confidence interval in which the rate of false negatives is 5%).

The use of performance indices for AYP purposes. Finally, the current study also highlighted the positive effects of using a performance index for AYP purposes, including its ability to account for both the lower initial starting points of the students with disabilities subgroup, as well any gains in achievement made below the absolute proficiency targets. Results clearly indicated that the relationship between the performance index and school growth was stronger than the relationship between the percent proficient policy and school growth.

The benefit of the performance index lies in its ability to account for small changes in student achievement not currently captured by the percent proficient/status policy, thereby acknowledging all student achievement, not just that which occurs around the percent proficient target. However, it is important to acknowledge that, although an obvious improvement over the percent proficient model, the performance index does not completely eliminate the problems associated with targets and cut-off scores. The development of a performance index merely increases the number of targets or levels that students may meet to receive ‘credit’ for AYP purposes. For example, the performance index utilized in the current study had four such performance levels; other indices may have more or less. However, schools may still only receive credit for certain predetermined increments of growth. In other words, not all increases in achievement are acknowledged; as with the percent proficient policy, only increases that meet the performance level cut-offs are acknowledged. Thus, although the current study clearly outlines the potential advantage to using a performance index, this policy does not completely address one of NCLB’s largest criticisms- its reliance on cut-off scores for AYP determination purposes.

Limitations

Although the results of the current study have significant implications for policy and practice, a number of limitations must be taken in to consideration. First, it is important to note that the data used in the present study represent demographic and achievement information from a sample of students in only one state. Characteristics of the students in the students with disabilities subgroup, including such key variables as socio-economic status and type of disability, will naturally vary from state to state, as will school-level characteristics as well. In addition, states also differ in the specific assessments administered, their content, and the cut-off scores selected to differentiate between ‘passing’ and ‘failing’ students. Furthermore, state-level differences in the adoption of varying policies governing AYP determinations (e.g., the varying use of performance indices across states) have also been noted. Given these variations, it is unclear how well the results of the current study will generalize to schools in other states.

A second, but related, limitation to the current study pertains to the population of students chosen to be included in the sample. Results of the current study only use achievement results from students in a limited number of grades (3rd-5th); it follows, therefore, that these students are only representative of a sub-sample of all the elementary schools in North Carolina (only those with a lowest grade of at least 3 and a highest grade of 5). In reality, the composition of elementary schools across the state varies significantly, with some schools limiting the number of grades taught to K-3rd grade, and others extending the highest grade included in the school to 6th grade students. Furthermore, although AYP is also calculated at the middle and high-school level, these schools were not included in the analysis.

An additional limitation relates to the use of confidence intervals in Hypothesis Three. In particular, it is important to note that confidence intervals were not applied to all school-level proficiency scores when determining AYP. Instead, confidence intervals were only used with schools whose proficiency levels fell below the AMO (and therefore, were at risk for failing AYP), and then the confidence intervals were only used to examine whether a school's true proficiency level might have fallen above the AYP cut-off score . As has been noted elsewhere, the use of confidence intervals in calculations of AYP acknowledges the fact that school-level results are subject to numerous sources of volatility and attempts to compensate for this by defining an interval of scores within which one is 95% certain the true proficiency score lies; if this interval includes the AMO, a school is given the benefit of the doubt and identified as passing. Thus, the use of confidence intervals allows school administrators to guard against false positive results, or the likelihood of a school being identified as failing AYP when its true proficiency level might have been above the AYP cut score. However, a second type of error is also possible: false negative results, or the identification of schools as passing when they've really failed. In order to guard against this second type of error, confidence intervals would have to be applied to the proficiency results of *all* schools and the possibility that a school's true percent proficient fell below the cut-off for AYP considered even though their obtained percent proficient was above the AYP cut-off score. It is important to note that using confidence intervals to account for both types of error would result in substantially different results than what is reported by this study.

Finally, it is important to note that the policies chosen for analysis in this study are not the only policies relevant to AYP determinations. In many instances, other policies have been proposed, and, in some cases, accepted by the state as possible methods for determining AYP at the school level. However, in North Carolina, these policies must be applied in a very specific, pre-determined order. For example, public schools must first attempt to meet AYP using the percent proficient policy. If a school fails to meet AYP under this policy, then the Safe Harbor provision may be applied. If AYP is still not met, confidence intervals may be used to determine AYP status. Additional policy options are also available for Title 1 targeted assistance (TAS) schools who continue to fail AYP even after the above policy options have been invoked. After attempting the to meet AYP using the Safe Harbor, confidence interval, and TAS policy options (if applicable), a growth trajectory may be calculated and used to determine AYP status.

The order in which policies may be applied is particularly relevant to this study, as it quickly becomes clear that not every school will have a chance to apply the policies evaluated here, as they may meet AYP before they get to that point in the sequence (e.g., a school that meets AYP under the percent proficient policy will not have the opportunity to calculate AYP using confidence intervals, as this option is only invoked at a later stage in the AYP determination process). Alternatively, other schools who do not meet AYP under the policies evaluated in this document may still have a chance to meet AYP using other policies (e.g., growth trajectories) not evaluated here. Thus, the application of this sequence of policies in a predetermined order will have an effect on how many schools are actually identified as 'passing' and 'failing' for any given year, above and beyond what has been reported in the results of this study.

Directions for Future Research

The current study has important implications for policy makers and school administrators involved in the evaluation and implementation of educational policies, especially those pertinent to the students with disabilities subgroup. For this reason, research in this area should be continued and expanded to ensure that the performance of all students, including student with disabilities, is accurately represented by school-level achievement results. The following suggestions for directions for future research are based, at least in part, on the limitations identified above.

As noted previously, the limitations of the current study pertain in large part to the external validity of the results. More specifically, current results are based on a very specific sample of students; thus, it will be important that future research replicate the current study using more diverse populations of students located in other parts of the country. This is particularly important given the variability in disability identification policies, proficiency assessments, and cut-off scores utilized across states. A replication of the present study using diverse populations of students would help to support the external validity of the current results.

In addition to broadening the sample of students and schools evaluated, further research examining the effectiveness of the policies described in this document within the larger context of AYP determination would be beneficial. As mentioned earlier, policies and methods of determining AYP status are invoked in a certain order when determining each school's AYP status. Unlike the current study, which examines the effectiveness of each policy in isolation, further research should seek to analyze the effect that these policies have when applied in combination, so as to provide a more realistic demonstration of their overall effectiveness.

Along these same lines, many of the policies and methods for determining AYP, particularly those that utilize cut-scores to indicate levels of proficiency within a school- are quickly becoming discarded due to copious amounts of research invalidating their use as accurate indicators of academic proficiency and performance. Thus, researchers would be well advised to focus their efforts on identifying and evaluating the validity of more complex indicators of performance, such as growth models. Such models are already being authorized and implemented within some states (including North Carolina) as part of the larger “package” of policies used to make school-level AYP determinations (Chester, 2005, Dunn & Allen, 2009). However, further research is needed to firmly establish the effectiveness of growth models for accountability purposes, especially for the students with disabilities subgroup (Buzick & Laitusis, 2010, Dunn & Allen, 2009). Ideally, researchers will continue to contribute to the literature on policy effectiveness, thereby helping to bridge the gap between educational research, policy, and practice.

References

Allbritten, D., Mainzer, R., & Ziegler, D. (2004). Will students with disabilities be scapegoats for school failures? *Educational Horizons, 82*, 153-160.

AYP in Pennsylvania: The 41 Possible Targets. (2006). Retrieved February 2, 2008, from http://www.paayp.com/possible_41_targets.html

Bray, M.A. & Kehle, T.J. (2011). *The oxford handbook of school psychology.* New York, NY: Oxford University Press.

Buzick, H.M., & Laitusis, C.C. (2010). Using growth for accountability: Measurement challenges for students with disabilities and recommendations for research. *Educational Researcher, 39,* 537-544.

Chester, M.D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice, 24,* 1-13.

Cole, C. (2006). *Closing the achievement gap series: Part III: What is the impact of NCLB on the inclusion of students with disablities?* Retrieved from http://www.ceep.indiana.edu/pub.shtml#ed.

Dunn, J.L., & Allen, J. (2009). Holding schools accountable for the growth of nonproficient students: Coordinating measurement and accountability. *Educational Measurement: Issues and Practice, 28,* 27-41.

Eckes, S. E., & Swando, J. (2009). Special education subgroups under NCLB: Issues to consider. *Teachers College Record, 111*, 2479-2504.

Erpenbach, W. J. (2009). *Determining adequate yearly progress in a state performance or proficiency index model*. Retrieved from the Council of Chief State School Officers website: http://www.ccsso.org/Resources/Publications/Determining_Adequate_Yearly_Progress_in_a_State_Performance_or_Proficiency_Index_Model.html

Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under NCLB: Central issues arising from an examination of state accountability workboos and U.S. Department of Education Reviews under the No Child Left Behind Act of 2001*. Retrieved from http://eric.ed.gov/PDFS/ED481838.pdf.

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA: SAGE Publications.

Forte-Fast, E., & Erpenbach, W. J. (2004). *Revising statewide educational accountability under NCLB: A summary of state requests in 2003-2004 for amendments to state accountability plans*. Washington D.C.: Retrieved from http://eric.ed.gov/PDFS/ED484706.pdf.

Four Pillars of NCLB. (2004). Retrieved January 17, 2012, from http://ed.gov/nclb/overview/intro/4pillars.html

Gordon, S. (2006). Making sense of the inclusion debate under IDEA. *Brigham Young University Law Review*, 189-225.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). Does special education raise academic achievement for students with disabilities? *NBER Working Paper Series* (pp. 1-53).

Hardman, M. L., & Dawson, S. (2008). The impact of federal public policy on curriculum and instruction for students with disabilities in the general education classroom. *Preventing School Failure, 52*(2), 5-11.

Huefner, D. S. (2006). *Getting comfortable with special education law: A framework for working with children with disabilities* (2 ed.). Norwood, MA: Christopher-Gordon Publishers.

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1412 (2004).

Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1414 (2004).

Jacob, S., Decker, D. M., & Hartshorne, T. S. (2011). *Ethics and Law for School Psychologists* (6 ed.). New Jersey: Wiley.

Jeckel, J.F., Katz, D.L., Elmore, J.E., & Wild, D. (2007). *Epidemiology, Biostatistics, and Preventative Medicine.* Philadelphia, PA: Saunders Elsevier.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal Of Economic Perspectives, 16*(4), 91-114.

Keele, C. E. (2004). Is the No Child Left Behind Act the right answer for children with disabilities? *UKMC Law Review, 1111*, 1-27.

Koretz, D. M., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence*. (Technical Report 587). Retrieved from http://www.cse.ucla.edu/products/reports.php?action=search&query=587.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16. doi: 10.3102/0013189x031006003

Marzano, R.J., Pickering, D.J., & Pollock, J.E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement.* Alexandria, VA: Association for Supervision and Curriculum Development.

McLaughlin, M. J. (2006). *Closing the achievement gap and students wtih disabilities: The new meaning of a "free and appropriate public education"*. Unpublished manuscript. University of Maryland. College Park, Maryland.

McLaughlin, M. J., & Thurlow, M. (2003). Educational accountability and students with disabilities: Issues and challenges. *Educational Policy, 17*, 431-451.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement--and why we may retain it anyway. *Educational Researcher, 38*, 353-364.

Nagle, K., Yunker, C., & Malmgren, K. W. (2006). Students with disabilities and accountability reform: Challenges identified at the state and local levels. *Journal of Disability Policy Studies, 17*, 28-39.

No Child Left Behind Act, 20 U.S.C. § 6301 et seq. (2001).

Ohio Department of Education. (2006). *Performance index.* Retrieved from Ohio Department of Education website: http://education.ohio.gov

Parker, J. M. (2011). *The use of large-scale assessment data to investigate special education effectiveness.* North Carolina State University, Raleigh, NC.

Public Schools of North Carolina. (2004). *North Carolina reading comprehension tests: Technical report (Citable draft).* Raleigh, NC: Public Schools of North Carolina, North Carolina Department of Public Instruction, Office of Curriculum and School Reform Services. Retrieved from http://www.ncpublicschools.org/accountability/testing/technicalnotes

Public Schools of North Carolina (2006). *The ABCs model for 2005-06.* Raleigh, NC: Public Schools of North Carolina, North Carolina Department of Public Instruction, Office of Curriculum and School Reform Services. Retrieved from http://www.dpi.state.nc.us/accountability/reporting/growthformulas

Public Schools of North Carolina (2006). *The North Carolina mathematics tests: Technical report.* Raleigh, NC: Public Schools of North Carolina, North Carolina Department of Public Instruction, Office of Curriculum and and School Reform Services. Retrieved from http://www.ncpublicschools.org/accountability/testing/technicalnotes

Public Schools of North Carolina. (2007). *Assessment brief: North Carolina end-of-grade tests.* Raleigh, NC: Public Schools of North Carolina, North Carolina Department of Public Instruction, Office of Curriculum and School Reform Services. Retrieved from http://www.dpi.state.nc.us/accountability/policies/briefs/

Schulte, A. C., & Villwock, D. N. (2004). Using high-stakes tests to derive school-level measures of special education efficacy. *Exceptionality, 12*, 107-126.

Simpson, M., Gong, B., & Marion, S. (2005). *Effect of minimum cell sizes and confidence interval sizes for special education subgroups on school-level AYP determinations* (NCEO Synthesis Report 61). Retrieved from the National Center on Educational Outcomes website: http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis61.html

Simpson, R., LaCava, P. G., & Graner, P. S. (2004). The No Child Left Behind Act: Challenges and implications for educators. *Intervention in School and Clinic, 40*, 67-75.

Stephenson, E. (2006). Evading the No Child Left Behind Act: State strategies and federal complicity. *Brigham Young University Education and Law Journal*, 157-188.

Stevens, J. (2005). The study of school effectiveness as a problem in research design. In R.W.Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 1-38). Maple Grove, MN: JAM Press.

Thurlow, M. (2000). Standards-based reform and students with disabilities: Reflections on a decade of change. *Focus on Exceptional Children, 31*(3), 1-16.

U.S. Department of Education, Institute of Education Sciences. (2011). *The Nation's Report Card: Reading 2011: National Assessment of Educational Progress at Grades 4 and 8.* Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf

U.S. Department of Education, Office of Elementary and Secondar Education. (2002). *No Child Left Behind: A desktop reference*. Retrieved from http://www2.ed.gov/admins/lead/account/nclbreference/reference.pdf

U.S. Department of Education, Office of Elementary and Secondary Education. (2010). *State of North Carolina Consolidated State Application Accountability Workbook.* Retrieved from http://www2.ed.gov/admins/lead/account/stateplans03/nccsa.pdf

U.S. Department of Education, Office of Elementary and Secondary Education. (2011). *State of North Carolina Consolidated State Application Accountability Workbook.* Retrieved from http://www.dpi.state.nc.us/docs/stateboard/meetings/2012/01/gcs/01gcs.pdf

Ysseldyke, J., & Bielinski, J. (2002). Effect of different methods of reporting and reclassification on trends in test scores for students with disabilities. *Exceptional Children, 68*, 189-200.

Appendices

Appendix A

Table A1

Crosstabulation Results of Schools Meeting AYP in Reading/Language Arts when Minimum Subgroup Size = 10 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	189	299
Met AYP	2	621

Table A2

Crosstabulation Results of Schools Meeting AYP in Reading/Language Arts when Minimum Subgroup Size =20 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	164	249
Met AYP	27	671

Table A3

Crosstabulation Results of Schools Meeting AYP in Reading/Language Arts when Minimum Subgroup Size =40 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	56	74
Met AYP	135	846

Table A4

Crosstabulation Results of Schools Meeting AYP in Reading/Language Arts when Minimum Subgroup Size = 60 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	6	8
Met AYP	185	912

Table A5

Crosstabulation Results of Schools Meeting AYP in Mathematics when Minimum Subgroup Size =10 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	524	281
Met AYP	5	301

Table A6

Crosstabulation Results of Schools Meeting AYP in Mathematics when Minimum Subgroup Size =20 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	465	232
Met AYP	64	350

Table A7

Crosstabulation Results of Schools Meeting AYP in Mathematics when Minimum Subgroup Size =40 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	168	85
Met AYP	361	497

Table A8

Crosstabulation Results of Schools Meeting AYP in Mathematics when Minimum Subgroup Size =60 (n=1111)

	95% Confidence Intervals	
Minimum Subgroup Size	Did Not Meet AYP	Met AYP
Did Not Meet AYP	27	17
Met AYP	502	565

Appendix B

Table B1

Rate of Sensitivity and Specificity at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Reading/Language Arts

Subgroup Size	Sensitivity	Specificity
10	99	67.5
20	85.9	72.9
40	29.3	92
60	3.1	99.1

Table B2

False Negative and False Positive Error Rates at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Reading/Language Arts

Subgroup Size	False Negative	False Positive
10	1	32.5
20	14.1	27.1
40	71.7	8
60	96.9	0.9

Table B3

Rate of Sensitivity and Specificity at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Mathematics

Subgroup Size	Sensitivity	Specificity
10	99.1	51.7
20	87.9	60.1
40	31.8	85.4
60	5.1	97.1

Table B4

False Negative and False Positive Error Rates at Various Subgroup Sizes Based on Number of Schools Meeting AYP for Reading/Language Arts

Subgroup Size	False Negative	False Positive
10	0.9	48.3
20	12.1	39.9
40	68.2	14.6
60	94.9	2.9