National Research and Development Center on Assessment and Accountability for Special Education

Significance of the Focused Program of Research

This is a re-application for a National Research and Development Center on Assessment and Accountability for Special Education. With this R&D Center, our ***first goal*** is to conduct research that provides evidence about the natural developmental progress in achievement of students with disabilities. Our ***second goal*** is to examine the reliability and validity of alternative accountability models where student academic growth is used to describe and evaluate school effectiveness in serving students with disabilities.

Nature and Scope of the Problem to be Investigated

Students' summative test scores in reading and mathematics in our current accountability system are key educational outputs used to evaluate the effectiveness of schools and teachers, and, in some cases, individual student's learning. Many educators and policy-makers believe that test-based accountability can be a useful strategy for raising student achievement, especially for low-performing students. The theory of action for this approach is based on the supposition that by setting standards and taking a snapshot of current student performance relative to these standards, teachers and students will both work harder and more learning will result (Marion, White, Carlson, Erpenbach, Rabinowitz, & Sheinker, 2002).

The centerpiece of federal accountability legislation, the No Child Left Behind Act (NCLB, 2001), requires reporting of school-level outcomes as well as the disaggregation of achievement test scores for subgroups who have historically performed poorly relative to other students. Individual states, however, determine many of the specifics of the accountability and reporting mechanisms used to comply with this legislation. As a result, states vary considerably in their assessment instruments, testing and reporting mechanisms, definition of grade level proficiency, and means for schools to demonstrate progress toward universal proficiency by 2014 (Heck, 2006; Linn, 2008). We know relatively little about the implications of states' assessment choices in terms of the reliability of building-level scores and the validity of the inferences and decisions about schools made on the basis of these scores (Heck, 2006; Linn & Haug, 2002; Zvoch & Stevens, 2005). Unfortunately, we know even less about the reliability and validity of scores for targeted subgroups reported at the school level (Kiplinger, 2008).

The lack of knowledge about impact of particular assessment and accountability choices on the reliability and validity of disaggregated test scores at the school level, or higher levels (e.g., district, state), is particularly acute for students with disabilities (SWD), one of the targeted subgroups in NCLB. First, schools' poor performance with this group of students has been a concern for decades (Carlberg & Kavale, 1980; Schulte, Osborne, & Erchul, 1998) and remains so today. With many states reporting that over 70% of students with disabilities are below expectations in reading and mathematics on annual statewide achievement tests, there is a critical need to provide accurate information to schools about whether their practices with this subgroup are effective. Second, in a recent 3-state study of schools that failed to make their adequate yearly progress (AYP) targets, Eckes and Swando (2009) found that the most frequent reason for schools' failure was the performance of this subgroup. Thus, unreliable measurement of schools' progress with this subgroup not only threatens the validity of inferences about schools' performance with SWDs, but inferences about schools' performance as a whole. Finally, several characteristics of students with disabilities present particular challenges for assessment and accountability programs. Student mobility in and out of special education is high, and exits and entrances into special education are correlated with achievement. This results in biased cross sectional reports of performance gains and losses (Schulte, 2010; Ysseldyke & Bielinski, 2002). SWDs are more likely to be retained in grade and/or perform at the bottom of grade-level score distributions. Both retention (because of a lowered standard relative to promoted peers) and low scores (because of measurement error and regression toward the mean) increase the chances of invalid inferences about the improvement of this subgroup of students when scores are compared across years with current status-based accountability models in NCLB.

In this proposal, we address the accuracy of current status-based student and school-reporting mechanisms in representing the performance of students with disabilities. We also investigate whether there are better ways to examine student outcomes. A number of educational scientists have argued for shifting our accountability metric away from achievement status to students' achievement growth (e.g., Betebenner, 2008; Hanushek & Raymond, 2005; Schulte & Villwock, 2004; Teddlie & Reynolds, 2000; Willms, 1992). The argument for using achievement growth rather than status as the basis for accountability is based on the dual premise that (a) schools should be held accountable for achievement outcomes they can control, such as how much students learn during the school year, rather than their prior achievement, and (b) status models incentivize schools to focus on students near the threshold of proficiency rather than focusing on the achievement growth of all students, including those functioning well below these thresholds (Ladd & Lauen, 2009). Thus, many educators want to learn more about ways to measure and characterize students' academic progress, especially for the heterogeneous group of students identified with disabilities, who as a group are disproportionately functioning below proficient status in virtually all states.

Finally, with the realization of high levels of student participation and the systematic collection and management of individual student annual test data by states, it is now possible to conduct longitudinal growth analyses of the academic achievement of students receiving special education services. Despite the promise of growth models for assessing students with disabilities and evaluating schools' performance, many issues remain largely unexplored in employing growth models with this population, as well as for their use with all children regardless of disability status. Thus, the need for this R&D Center is evidenced by the paucity of longitudinal research on achievement growth for individual students with disabilities who take general achievement tests. In addition, we simply don't yet know how well current (status-based) and proposed (growth-based) school-level accountability models represent student achievement outcomes, particularly for students with disabilities. Interestingly, the academic performance status of students with the most severe disabilities is very different according to the recent report from the National Study on Alternate Assessments (Cameto, et al. 2009) where nearly 75% of these students, on average across states, performed at or above the proficient level on alternate assessments based on alternate achievement standards. Although the achievement status for this subgroup of students with disabilities is relatively high compared to the performance standards states set for them, little is known about their achievement growth. The proposed re-application for a National Research and Development Center on Assessment and Accountability for Special Education fully addresses all previous IES reviewers' concerns. We summarize each concern and our response in Figure A.1 in Appendix A. The revisions respond to areas in which more clarity and specificity was needed as well as a more directed program of research linked to the Request for Proposals.

Research Questions and Overview of Datasets and Linked Studies with Research Partners

To accomplish the Center's goals and address the problem of measuring achievement growth for educational accountability purposes, we propose conducting a tightly linked series of growth modeling studies that address 6 key questions:

1. What is the natural developmental progress in achievement for students with disabilities?
2. What models best characterize achievement growth for students with disabilities who are participating in general achievement tests, as well as those taking alternate assessments?
3. How do various growth models represent school effects for students with and without disabilities, and how do results compare to those derived from status models now in use?
4. What are the reliability and validity of the estimates of school effectiveness for students with disabilities produced by alternative growth models and how are these estimates influenced by contextual differences among schools and students?
5. How do results from different types of interim assessments of students' achievement meaningfully contribute to a model of academic growth for students with disabilities?
6. How can information about opportunity to learn and achievement growth be used to enhance academic outcomes for students with disabilities?

To answer these questions with scientifically sound evidence requires the use of longitudinal designs, an understanding of measurement limitations, and a command of an array of statistical analyses and comparison techniques (Barton, 2005; Gong, Perie, & Dunn, 2006; Linn & Haug, 2002; Raudenbush, 2004; Singer & Willett, 2003; Stevens, 2005). It also requires access to large and representative datasets of both summative and interim assessments. Further, we maintain that these large datasets should not only include students with disabilities, but students without disabilities in order to understand how students with disabilities differ from their non-disabled peers, and how well different school-level growth models represent school effects for both populations. This research strategy is in keeping with the principles of "inclusion" and "least restrictive alternative" that underlie many policy decisions regarding students with disabilities, including their participation in current standards-based reforms (McDonnell, McLaughlin & Morison, 1997). A preference will be given to school-level growth models that are valid for both populations (inclusion), unless there is clear evidence that a different standard or growth model for SWDs is merited by the data (least restrictive alternative).

Thus, the proposed R&D Center's Focused Program of Research on achievement growth is based on 10 existing sets of longitudinal achievement data for students, and 1 new dataset. Four datasets are the annual (summative) achievement test results for students in grades 3-8, from North Carolina (NC), Arizona (AZ), Oregon (OR), and Pennsylvania (PA). Four additional extant datasets are results from partner states' annual alternate assessments for students with significant cognitive disabilities. The remaining 2 extant datasets are from the national, longitudinal databases of widely used interim assessments (with SWD and nonSWD), and the to-be-created dataset includes summative, interim, and additional measures. With the exception of the 4 alternate assessment datasets, all datasets include students with and without disabilities.

We use the term *interim assessments* throughout this proposal to refer to technically sound, brief measurements of achievement in reading and mathematics that can be used multiple times within a school year, as well as across school years for cohorts of students in grades 1-8. These measures (a) are linked to a developmental scale, (b) have wide use in schools, particularly with special education students, and can include multiple administrations per year, and (c) offer a window on early achievement growth that is not possible with large-scale summative assessments because interim assessments are frequently used prior to grade 3. As noted by Andrade and Cizek (2009), these measures are distinguished from formative assessments by greater standardization of administration, technical adequacy in terms of reliability and validity, and the use of carefully constructed score scales including equated alternate forms and vertical scaling of forms. These assessments provide teachers with important instructional, evaluative, and predictive information (see Perie, Marion, & Gong, 2009). As a consequence, their inclusion in our research allows modeling within-year growth, as well as between-year growth, and allows examination of important policy questions such as whether summer losses and gains affect student or school results (Downey et al., 2008; Zvoch & Stevens, in press). The 2 interim assessments included in this research, the Northwest Evaluation Association's Measures of Academic Progress (NWEA MAP) and easyCBM (from the UO) both allow examination of growth starting with Grade 1 and have been designed so scores across years are vertically scaled. They both contain students with and without disabilities.

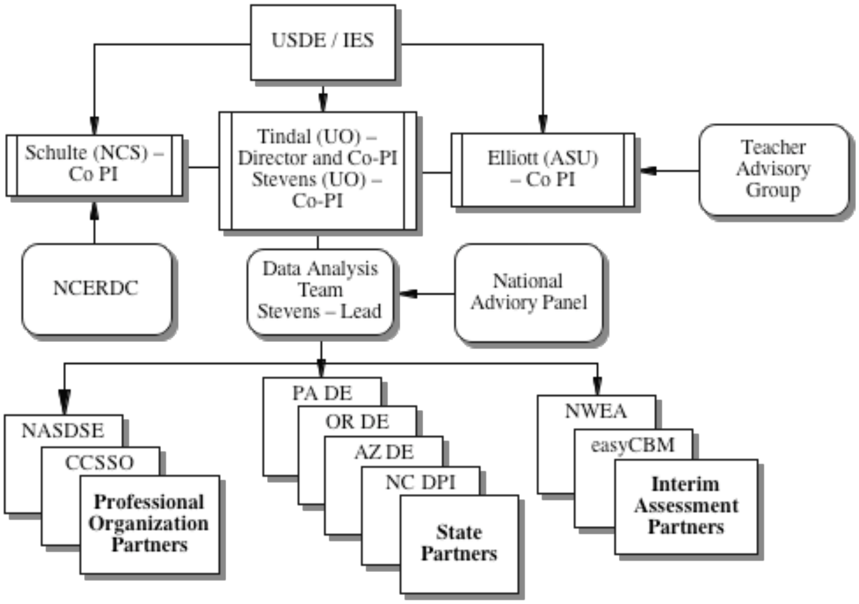
The R&D Center's growth modeling research with these existing and annually expanding extant datasets will be organized into a coordinated series of studies including: a ***Cornerstone Study*** featuring the application of multiple growth models with the NC summative dataset and the examination of which growth models or variants within a single model work best for different assessment and accountability purposes; a ***Multi-State Extension Study***, where the replicability of the Cornerstone Study findings about the growth of students with and without disabilities and validated growth model(s) in NC are examined with the AZ, OR, and PA statewide summative datasets; and an ***Interim Assessments Study*** where the multiple growth modeling methods used with the state datasets are applied to national longitudinal datasets from 2 interim assessments commonly used to provide teachers instructional feedback about students' knowledge and skills.

Two supplemental studies will contribute an additional dataset and analyses that also address our key research questions. Specifically, we propose a ***Multiple Measures Validation Study*** that utilizes a common dataset collected in 200 classrooms in OR and PA and comprised of concurrent measures of opportunity to learn, interim assessment, and summative achievement. Finally, an ***Alternate Assessments Study*** is proposed to focus on the achievement growth of students with significant cognitive delays using each of our 4 partner states' alternate assessment datasets and appropriate growth modeling methods.

In summary, we propose to analyze and compare the longitudinal achievement of students with disabilities in reading and mathematics using 10 extant datasets and 1 new dataset that contain test scores from summative state tests or interim assessments, or both. The selected datasets and featured longitudinal designs allow students to serve as their own controls and are less susceptible to the influences (e.g., student background, intake characteristics, and other confounding factors) that undermine the interpretability of status-based test scores from cross sectional samples for assessing student growth and school effectiveness (Stevens, 2005; Willms & Raudenbush, 1989). Longitudinal statewide assessment datasets, especially those featuring students with disabilities, can be complex due to an array of measurement, design, and quality issues (Schulte & Villwock, 2004; Stevens & Zvoch, 2006), some of which are unique to a population of students with disabilities (e.g., changing special education status across years, higher grade retention rates, more missing data, and small cohort size within schools). However, when analyzed carefully by a research team with expertise in growth modeling and special education, these data have the potential to answer the research questions proposed for this R&D Center.

To enact such a program of research on achievement growth and objectively answer the 6 questions that guide our work, the R&D Center's research team will consult with a panel of national experts and work cooperatively with the IES, 4 states' departments of education, 2 vendors of interim assessments (Northwest Evaluation Association for access to Measures of Academic Progress, and University of Oregon [UO] for access to easyCBM), and 2 national organizations (NASDSE and CCSSO) to complete a program of exploration and validation research on (a) the expected growth patterns in reading and mathematics for students in grades 3 through 8, and (b) the use of growth measures for school accountability. Figure 1 highlights the relationships among the R&D Center's Principal Investigators (PIs) and partners, all of whom are committed to answering our research questions to advance assessment and instructional practices for students with disabilities (See ***Appendix A*** for Letters of Commitment and documentation of data access).

Figure 1. Partners in the R&D Center on Assessment & Accountability for Special Education



What We Know Now

From our experience and with guidance from rigorous growth modeling research by McCaffrey et al. (2004), Stevens (2005), and Tekwe et al. (2004), as well as the practical CCSSO (2008) *Implementer's Guide to Growth Models*, we focus on 3 common approaches to modeling growth on achievement tests. We contrast these 3 approaches with the more commonly employed status approach, and identify design and data quality issues that are critical to the effective use of growth approaches. An integrated summary of this discussion and key questions that each growth model can effectively answer is provided as Figure 2 (on page 8).

Common Approaches to Modeling Student Achievement and School Effectiveness

Many methods are in current use to measure student progress. Most common, of course, is the federally mandated percent proficient method for determining adequate yearly progress incorporated in the No Child Left Behind Act (NCLB) of 2001. Since 2005, additional federal flexibility has been available for developing and applying growth models. While NCLB's Adequate Yearly Progress (AYP) methods measure student performance in a single year in relation to a performance standard, growth models characterize the amount of students' academic progress over two or more points in time. Although a variety of growth models are being applied by states and districts to augment or inform NCLB proficiency methods, limited research examines the relative efficacy of these alternative models, their reliability and validity, and their ability to generalize to students receiving special education services. The purpose of our posed R & D center is to empirically examine these alternative models of estimating student learning and school effectiveness.

NCLB Status and Improvement. Status models provide a picture of academic performance at a single point in time and have the same advantages and disadvantages of a census. Status models provide a summary of student performance and provide a snapshot of school performance. School status performance (i.e., percent proficient) is interpreted through comparison to a performance standard or benchmark. Another variant in NCLB is the different groups improvement model, embodied in the NCLB "safe harbor" provision. In this model, different cohorts of students are compared from one year to another to determine change in percent proficient. Although status models are sometimes considered growth models, they do not track individual student progress over time and their accuracy in reflecting school change depends on the questionable assumption that the student population has remained stable from one year to the next. As we compare school-level growth approaches in terms of their validity, the NCLB Status and Improvement model will serve as a starting point for evaluating the advantages and disadvantages conferred by various growth modeling approaches.

Transition Matrix. Students' growth is tracked at the performance standard level. A transition matrix is set up with previous performance levels (e.g., Does Not Meet, Meets, Exceeds) as rows and current performance levels as columns. Each cell indicates the percent of students that moved from year to year. The diagonal cells indicate students staying at the same level, cells below the diagonal show the students moving down one or more levels, and the cells above the diagonal show the students moving to higher performance levels. Transition matrices can be combined to show the progress of students across all tested grades to show total performance for the school. This approach is included here because it allows scores from tests on different scales to be aggregated and a substantial number of students with disabilities take alternate assessments that, in many states, are on a different scale than the assessments used in general education for reporting school-level results.

Residual Gain and Value Added Models (ResVAM). A variety of residual models have been used in accountability systems. The simplest predicts current performance from past performance. Each student has a predicted score based on achievement from a previous occasion. The difference between predicted and actual scores in the current year is the residual score. Residual gains near zero indicate growth consistent with prediction, positive scores indicate greater than predicted growth and negative scores indicate performance lower than predicted growth. Residual gain scores can be averaged to obtain a group growth measure, but they are not easily integrated with performance standards because they focus on relative gain. The growth models used by NC and AZ are based on this approach.

Value-added models (VAM) are a variation of residual models. The best known are variations like Sanders' Tennessee Value-Added Assessment System (TVAAS; Sanders & Horn, 1998), the Dallas accountability system (Webster, 2005), the Chicago School Productivity model (Bryk, Thum, Easton, & Luppescu, 1998; Ponisciak & Bryk, 2005), and the RAND model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). All of these models use prior achievement as a predictor of performance, some use multiple years of prior achievement, and some also include other conditioning predictors including measures of student background characteristics. VAMs are now being used and applied in a number of states to estimate and link teacher effects to student achievement results. The growth model used in PA is a variant of Sanders' TVAAS.

Multilevel Growth Models (MGM). Longitudinal designs entail tracking an individual over time and measuring performance at two or more time points (see Rogosa, 1995; Willett & Sayer, 1994; Willett, Singer, & Martin, 1998). There are at least two major advantages to these designs: (a) each student acts as his or her own control (see Stevens, 2005), and (b) the focus is on the outcome of interest, student learning. MGMs fit growth trajectories to each student's data using a two or three level structure. The first level of each model is used to represent measurement occasions and estimate a growth trajectory for each student. The second level in the structure represents student characteristics and the third level represents school context, characteristics, and programs.

An important feature of MGMs is that variation in performance can be separated out into the student and school levels. These models allow the estimation of an intercept and a slope for each individual and each school. Time in school (an indication of opportunity to learn and of the amount of exposure to school practice and policy) and other variables can be explicitly included as predictors of individual student's growth. Application of the MGM will produce individual student growth trajectories, as well as school average growth trajectories (see Appendix A, Figure A.2)

Evidence indicates that multilevel models are more accurate than difference or residual gain score models (Koretz & Hamilton, 2006). As part of our research we plan to study variations of MGMs including the addition or exclusion of covariates, the number of measurement occasions used in the model, the use of piecewise growth models to study seasonal effects of learning, the functional form of growth (i.e., linear, curvilinear, cubic), the timing of assessments (year to year versus multiple occasions within year) and the kind of statistical estimator used (i.e., Empirical Bayes vs. Ordinary Least Squares). Although none of our states employs a MGM as part of their accountability system, all 4 states' large-scale assessments and our 2 interim assessments yield student-level scores that permit their application.

Implementation of Growth Models

Growth models require multiple years of high quality test data for the same students connected to the schools they have attended. We highlight 7 key characteristics of growth models that are critical to ensure integrity of any subsequent analyses. This list of design and measurement issues is informed by the empirical work of Kolen (2008), Stevens and Zvoch (2006), and Willet et al. (1998).

Database of matched student records over time. Measurement of growth requires analysis of individual student's results from two or more occasions. Until recently, most state systems lacked a student ID system that assigned a unique identification number recorded with each assessment. Without an ID number, record matching must be based on some combination of name, birth date and other demographic information and is prone to error. A student ID system with a unique identifier is in place in all our partner states.

Common scale for reporting scores. Most growth methods require student scores to be reported on a common scale. Ideally, this would mean that all the tests were designed for use on a developmental scale intended to support the measurement of growth and based on content standards aligned across grades. It is possible, however, to create a common scale for types of tests using equating methods. A number of technical issues and challenges, however, exist in creating vertically linked assessments and even when developmental scales have been created for an assessment (e.g., NC), construct equivalence over large time spans may be difficult to establish. That is, the construct measured ("math achievement") may change if the time span encompasses a fundamental evolution in what the construct references (e.g., math facts in 3rd grade versus geometry reasoning in 10th grade). For this reason, our research plan only uses assessments that have vertically linked developmental scales and, in any one analysis, we only include time spans in which the construct of interest can be expected to be relatively invariant.

Precision and accuracy evaluated. Proper use and interpretation of growth models requires that uncertainty be taken into account in using analytic results. This can be done in several ways including evaluating parameter reliability, examining the stability of results across cohorts, examining the size of standard errors, and computing and applying confidence intervals. For each of the growth models examined we will evaluate parameter reliability using the following formula: $\lambda_j = (n_j \rho_I) / [1 + (n_j - 1) \rho_I]$, where n_j represents the school sample size and ρ_I is the intraclass correlation coefficient. For either status or growth parameters, λ_j represents the portion of the variation in status or slope that is "true" versus residual variation. We will specifically examine these issues in the application of our various growth models to determine the impact on inferences drawn about the growth of individual students or the effectiveness of schools. We will also explore the way in which particular policy decisions (e.g., minimum n-size or size of confidence interval used) may impact conclusions drawn about growth model results.

Students with missing scores included. The occurrence of missing data can severely undermine the accuracy of estimates of student or school performance and may occur for many reasons (e.g., from a student absence for illness to a change of school, drop out, or a family move to another jurisdiction). While longitudinal designs provide some control over the nonrandom assignment, differential attrition across schools can result in systematic bias, undermining the validity of the conclusions drawn about student, school, or district performance.

If a large percentage of students do not stay in the same school long enough to be tested, then the model results may not be representative of all students. In previous research we have found that missing data may lead to mis-estimation of school performance (Zvoch & Stevens, 2005). For all growth models we analyze in each partner state, missing data rates will be reported and effects of missing scores on estimates of school effectiveness will be examined empirically. More information about our missing data solutions is provided below.

Affected by cohort stability. Another concern is whether the accountability method is sensitive to the particular cohort characteristics. The use of measures that index the achievement status of a single cohort (relative to a proficiency target) or the change in status between two successive cohorts may result in school evaluations biased by factors (e.g., student demographics) that are outside of the school's control (Linn & Haug, 2002; Kane & Staiger, 2002; Raudenbush, 2004). Research has also shown that the cross-cohort performance of schools differs depending on whether the mean achievement status or growth of students is considered and cross-cohort estimates of student achievement are differentially impacted by school characteristics. School composition factors like free lunch percentage are strongly related to the status or average performance level of schools (Raudenbush, 2004; Zvoch & Stevens, 2006) while both status and growth estimates of school performance are related to cohort size. Schools with smaller student cohorts had greater changes in student outcomes than schools with larger cohorts. Estimates for successive cohorts in smaller schools are more volatile with more potential for differences to occur in the make-up of student cohorts (Kane & Staiger, 2002; Linn & Haug, 2002). As part of the center's planned research activities, we examine the stability of estimates of school effectiveness across cohorts for each of the alternative analytic models studied.

Non-linear growth examined. Some growth models assume that each student's growth in achievement is linear. Growth over multiple occasions, however, is often non-linear with greater growth rates early in the school year and over years, with elementary grade growth rates greater than high school growth rates (e.g., Raudenbush & Bryk, 2002; Stevens, 2005). If growth is nonlinear, it may be necessary to apply a growth model that can correctly represent the shape of the growth trajectories (Stevens & Zvoch, 2006). For each relevant model we examine, we plan to study the functional form of the relationship between measurement occasion and student achievement performance, using a variety of empirical methods including empirical plots, parametric statistical procedures (polynomial regression with linear quadratic, and cubic terms) and nonparametric statistical procedures (kernel regression).

Includes results from alternate tests. If alternate assessments (i.e., for students with significant disabilities) do not produce scores on a common scale with the general test, it may not be possible to include those students in the growth calculations. The Transition Matrix model then can be based on student progress as indicated by changes in the performance levels attained by students. However, common performance levels must have been set across different tests for results to be combined. Meaningful results depend on the assumption that the performance standards were set such that it is reasonable to assume that the performance levels on both tests indicate that students have the same knowledge and skills. The analysis of achievement growth on alternate assessments is possible by our partner states; we plan to test the inclusion of these assessments in the same growth models used with the general test.

In summary, the 4 types of growth indicators can be crossed with the 9 issues on data management and interpretation to form a matrix highlighting specific issues and eventually answering different questions about improvement of students with disabilities (see Figure 2).

Figure 2. Data and Psychometric Characteristics of Common Growth Models

Data Requirements	***NCLB***	***TM***	***ResVAM***	***MGM***
Database of matched student records over time (Stdnt ID)	No	Yes	Yes	Yes
Common scale	No	No	Yes	Yes
Precision and accuracy evaluated	Yes	Yes	Yes	Yes
Confidence interval	Ind. Grps.	Std. Errors	Error Var.	Error Var.
Includes students with missing scores	Yes	No	No	Yes
Affected by cohort stability	Yes	Yes	Yes	Yes
Handles non-linear growth	No	No	No	Yes
Includes results from alternate tests (different scales)	No	Yes	No	No
Student performance standards in definition of growth	Yes	Yes	No	No
The primary growth question answered:				
NCLB + Status Improvement	Did this year's students meet AYP?			
Transition Matrix	Are students in a group making adequate progress across performance levels?			
Residual Gain Scores and Value Added Models	How much residual change was produced by a group?			
Multilevel Growth Models	What is the school growth rate?			

Key: NCLB = NCLB +Status Improvement, TM =Transition Matrix, ResVAM = Residual Gain Scores and Value Added Models, MGM = Multilevel Growth Models

What We Need to Know Next

As noted, student and school level achievement status and growth can be represented in a number of ways. States are not only faced with choices between the 4 approaches depicted in Figure 2, but also with choices and options within each approach. Varying specific features, such as how many measurement occasions are included in MGMs, or how many students are required before outcomes for a subgroup are reported separately, often have important implications for how student and school outcomes are represented. Each of our state partner's growth models is described along a set of common dimensions as Figure A.3 in Appendix A. As indicated by the Center’s research questions, important unknowns include (a) the natural developmental progress for students with all types of disabilities, (b) similarities and differences in various growth models in terms of their efficacy and representation of student and school effects, (c) the possible additive value of interim assessments of achievement to those of statewide summative achievement tests, and (d) the best method(s) for accurately revealing the development of students whose knowledge and skills are measured by alternate assessments.

The Performance of Students with Disabilities on Statewide Achievement Tests, Related Academic Measures and Alternate Assessments

Students with disabilities are a diverse group and exhibit a wide array of skills and behaviors. The majority of these students learns at rates slower than their same age peers and often experience difficulty in acquiring reading skills (DiPerna & Elliott, 2000; McDonnell, McLaughlin, & Morison, 1997). Individualized special education services are provided to support learning and provide access to general education curriculum. The development of knowledge and skills in language arts and mathematics has been a priority goal. The expected development of all students' knowledge and skills in these academic domains has been articulated by each state in content standards aligned with annual tests to evaluate achievement. Test results, both general achievement and alternate assessments, have most often been evaluated against achievement standards that yield dichotomous status reports of "proficient" or "not proficient." Although a substantial amount of research exists about the characteristics of students with disabilities and about assessment of their abilities and skills for purposes of classification and intervention, far less is known about their growth rates, general curriculum access, and the effectiveness of the services they receive.

What We Know Now

Recent reports of state test score trends have indicated that students at risk educationally, which includes students with disabilities, have largely participated in their state assessments; however, the majority of their performances have not met the achievement standard of "proficient" (Center on Educational Policy, June 2009; Thurlow, et al., 2008). And yet educators sense that students with disabilities are progressing, but just not as fast as their peers because it takes more time for them to learn. More research is needed to understand the natural developmental growth rates of students with disabilities, and the relation between their instruction and achievement in reading and mathematics. A comprehensive understanding of large-scale testing of SWDs and alternative measures of academic progress is needed to guide the design of such research. Therefore, we summarize the key findings from several research bases and seminal studies from the peer-reviewed and published literature. We also briefly review the developing research on opportunity to learn because it has important implications for the validity of findings about achievement growth, especially when growth is poor.

Participation rates in large-scale assessments. Nationally in 2005-2006, the percentage of students with disabilities tested on reading assessments (regular and alternate) at the elementary and middle school levels has exceeded 95% in 45 states. In our partner states of AZ, NC, OR, and PA, the participation rates in reading for students with IEPs has averaged 99%, 100%, 99%, and 99%, respectively over the past 3 years. The participation rates for mathematics tests for students with IEPs has averaged nearly the same at 98% or higher in all four states.

Accommodated assessments. Many students with disabilities in elementary and middle school grades receive accommodations to facilitate their access and responses to assessments. On statewide reading and mathematics during 2005-2006, nearly 75% of students with IEPs received accommodations in three states, while more typically 50% to 74% of students with IEPs received accommodations in the majority of states (Lazarus et al., 2009). In our partner states of AZ, NC, and PA accommodations were reported for 52%, 33%, and 53% for reading tests and 51%, 36%, and 53% for mathematics tests, respectively. In OR, accommodations are allowed for all students, whether or not they are receiving specialized instruction on an IEP; therefore, data are not collected on accommodations.

Achievement level performance on large-scale assessments. For states in which rates of student proficiency could be calculated for reading and mathematics assessments (both general test and alternate assessment) completed in 2005-2006, generally more than 30% of students with IEPs performed at a level considered proficient. According to the NECO Annual Performance Report (April, 2008), there was a slight improvement in both reading and mathematics achievement for students with IEPs from 2003-2004. For the same period in our partner states, the percentage of students with IEPs who achieved at or above the proficiency level was 24% (AZ), 55% (NC), 55% (OR), and 29% (PA) on the reading assessments, and 27% (AZ), 35% (NC), 57% (OR), and 36% (PA) on the mathematics assessments.

Achievement growth using large-scale assessments. Only a handful of empirical investigations document the achievement growth of students with disabilities using large-scale assessments. In one of the first published studies of the achievement growth of students with disabilities using large-scale assessment results, Schulte et al. (2001) examined whether students with learning disabilities in a single school district met the North Carolina proficiency and growth standards across two school years. Although the students with learning disabilities were far less likely to meet the state's proficiency standards, their mean level of growth met or exceeded the state's growth standards in grades 4 and 5 in both years. The authors noted the complexities of looking at growth across grades due to changes in the special education population across because of entrances and exits of students from special education services.

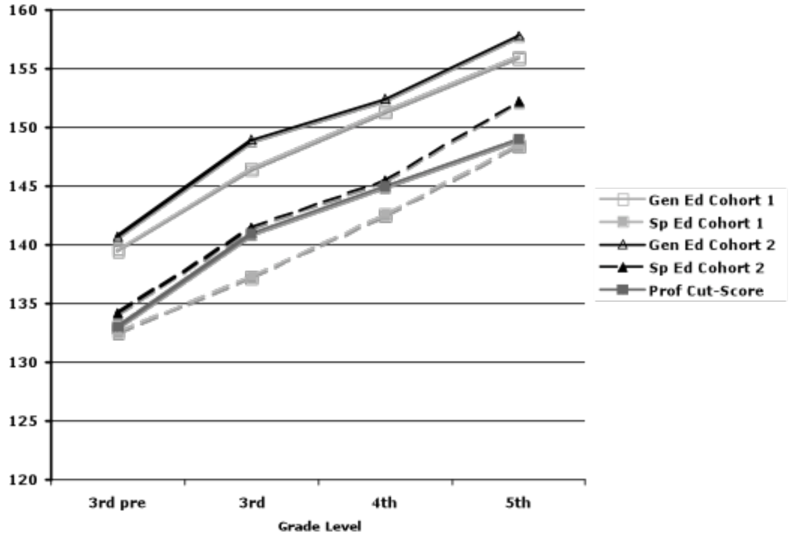
Ysseldyke and Bielinski (2002) expressed similar concerns about tracking progress of students with disabilities without considering changes in student special education status across years. They examined changes in the achievement gap between SWDs and general education students across grades 4 to 8 using statewide test data from Texas. They found that 13% of students receiving special education in 4th grade had exited services by 5th grade, and that 17% of the students receiving special education in 5th grade were not in special education in 4th grade. The cumulative impact of the yearly exodus of high performing students from special education and the entrance of low performing students into special education markedly affected how longitudinal trends in the achievement gap appeared for SWDs. When special education entrances and exits were uncontrolled the achievement gap increased by 93% from 4th to 8th grade. When a stable sample of students in special education was followed from 4th to 8th grade, the achievement gap decreased by 12%.

Schulte and Villwock (2004) examined three years' of school-level special education and general education growth and proficiency outcomes in reading for a sample of six NC elementary schools in one district. They also followed three longitudinal cohorts of students from the same schools across grades 3 to 5 and contrasted annual school-level results for SWDs with longitudinal outcomes tracked by student rather than school. They found that schools varied more on measures of the percent of SWDs reaching proficiency than on the percent of SWDs who had exceeded the state's expected growth standards. They also found that when the predictors of student achievement level at grade 5 were considered for the three longitudinal cohorts, students' third grade achievement level accounted for over 40% of the variance, and school accounted for less than 2%.

In another study of achievement growth for students with disabilities, Zvoch and Stevens (2005) found that students in special populations demonstrated achievement growth at a rate that was indistinguishable from their counterparts in the data submitted for accountability reporting. However, data from a more inclusive student cohort revealed that ethnic minority, impoverished, and special education students grew at a slower rate than their peers. Thus, whereas in one student sample it appeared as though schools were under-serving special student populations (i.e., achievement gaps widened over time), in the other it appeared as though traditionally disadvantaged students kept pace with their more advantaged peers. This study highlights the importance of carefully attending to missing data in assessing growth of students with disabilities.

More recently, Schulte (2010) extended her initial investigations of achievement outcomes and growth for students in special education in a single district by using multiple years of extant state-level longitudinal data from NC's testing program. These analyses confirmed that SWDs differ markedly from non-identified students in terms of achievement level, but less so in terms of growth (see Figure 3), and that the high rate of exits and entrances from special education during the elementary school years (see Figure A.4 in Appendix A) distorts the size of the achievement gap when data are reported cross-sectionally rather than longitudinally.

Figure 3. Growth in Reading Scale Score Across Grades for Two NC 5th Grade Cohorts



Other measures of achievement growth. A number of investigators have studied growth of students with disabilities using measures other than large-scale achievement tests. Two key studies focusing on achievement growth or its determinants with well-defined samples of students with disabilities are briefly reviewed next to highlight the utility of using alternative measures to supplement growth analyses based on large-scale assessments.

Deno et al. (2001) contrasted typical growth for general and special education students across elementary grade levels for reading aloud with a technically sound Curriculum-Based Measurement (CBM) instrument. Their results indicated that for both groups of students a growth rate of 2 words per week in reading aloud from grade level text was typical. In addition, a growth rate of 1 in grades 2-6 may represent a reasonable standard, with a slightly higher rate in grades 2-4 than in grades 5-6.

Finally, in an examination of opportunity to learn and its relationship to achievement, Kurz, Elliott, Wehby, & Simthson (2009) examined the content of the planned and enacted 8th-grade mathematics curriculum for 18 general and special education teachers and the curricula's alignment to state standards via the *Surveys of the Enacted Curriculum* measure. The relation between alignment and student achievement also was analyzed for three formative assessments and a corresponding state test within a school year. Results indicated that alignment to state standards was low with no significant differences between general and special education teachers. Significant correlations between student achievement averages for 238 students and teacher alignment indices were ≥ .48. When teacher groups were examined separately, the relation between alignment and achievement remained significant only for special education with correlations equal or greater than .75. This finding highlights the importance of measuring opportunity to learn content that is present in the test. This research on alignment and opportunity to learn, along with several of the analytical tactics discussed by Muthen et al. (1995) offer promising methods for advancing a more comprehensive understanding of achievement growth for students with disabilities.

What We Need to Know Next

Much more information about the achievement growth of students with disabilities is needed and how various approaches to modeling school effects fare when applied to outcomes for students with disabilities. Five recent growth studies that did not centrally consider SWDs provide promising methods and comparison data to guide our Focused Program of Research.

Bloom, Hill, Black, and Lipsey (2008) used norm group data from seven nationally standardized achievement tests to characterize typical growth in reading, math, science, and social studies as grade level effect sizes (ES). In a pattern that was consistent across tests and content areas, students showed large annual gains in the early grades (K-2) that decreased gradually in the later grades. They also examined performance gaps between policy relevant target groups (e.g., Black vs. White, male vs. female students) at each grade and subject area. They found marked differences in achievement gap ES across different combinations of grade level and target groups. Bloom et al. did not examine growth, or variations in the achievement gaps across grades for SWDs. However, their findings point to the importance of understanding grade-by-grade growth trajectories when setting growth benchmarks or evaluating the impact of interventions in closing the achievement gap between students with and without disabilities.

In addition to the strategy used by Bloom et al. (2008), we will also explore growth percentiles as a means of examining the natural developmental progress in achievement of students with disabilities (Betebenner, 2009). Growth percentiles, a concept adapted from pediatrician's growth charts, provide a means of putting a growth from a given starting point in a normative context (Betebenner & Linn, 2010). They can be used to both characterize observed growth and state probabilities concerning how much growth can be expected for students at a similar starting point. In addition to being useful for characterizing growth for parents and school staff at the individual and school level, growth norms have also proved useful in evaluating the feasibility of the growth expectations contained in different states' growth to standard models, particularly for non-proficient students (Dunn & Allen, 2009).

Williamson, Thompson, and Baker (2006) used multi-level modeling with large-scale assessment data from five cohorts of students to characterize reading and mathematics growth across grades 3-8. The average growth curve showed substantial initial growth with decelerating growth over time. The pattern was similar to the pattern of growth reported by Bloom et al (2008), although their methodologies were different. Williamson et al. found considerable between-student variation in intercept (set at the initial 3rd grade assessment) with less variability in velocity and curvature. The authors did not look at whether status variables such as disability classification accounted for some of the variability in intercept or growth curve shape, but speculated demographic variables might explain the variability observed between students.

Heck (2006) used data from 123 schools to simulate results from different ways of evaluating school effectiveness with large-scale achievement data. He compared school performance measures based on student proficiency levels with those based on growth. He then looked at schools where results based on student proficiency were discrepant from the results based on growth. He found a school’s percent of special education students was one of the factors that helped account for discrepant results between growth and proficiency measures. Heck’s study showed how various accountability models affect portrayals of school performance and his results suggested that exemplary outcomes for special education students may be missed when school performance is assessed only in terms of proficiency levels.

Zvoch and Stevens (2008) conducted an investigation that featured multilevel models to compare the validity of status and growth-based achievement indicators for measuring school performance within the NCLB accountability framework. They analyzed data from a mathematics achievement test to evaluate school performance for a longitudinal sample of several New Mexico middle school student cohorts. Specifically, they compared measures designed to index the within-cohort achievement status (school mean achievement) and growth of students (school mean growth), as well as the status (school mean achievement change) and growth-based changes (school mean growth change) that occurred over years between cohorts. Both the simple and complex statistical models applied to the longitudinal time-series achievement data produced similar descriptive outcome estimates. The results indicated that within-cohort status-based estimates of mathematics performance were closely related to student demographics (i.e., ethnicity, ELL, Special Education status) and between-cohort estimates of performance were associated with cohort enrollment size and initial performance level. The results also indicated the within-cohort measures of performance were estimated with greater reliability than the between-cohort measures. However, of the within-cohort measures examined, status-based indices were more confounded with student demographics than those based on growth. These results suggested that schools using only status-based performance methods may often be rewarded and penalized on the basis of factors over which they have limited or no control.

Validation of Achievement Growth and Use of Growth Data for Accountability Purposes

The overarching research strategy guiding work in this proposal is one of using multiple sources of data and multiple methods to evaluate the validity of inferences made about student growth (Goal 1) or about school impact and effectiveness through the use of alternative statistical models and accountability purposes (Goal 2). Shadish, Cook, and Campbell (2002) describe a process of pattern matching as a method of determining validity that involves the logical, theoretical consideration of the attributes and characteristics of a construct that should be present followed by a process of observation and matching of actual attributes and characteristics. This process is particularly useful in the context of research designs in non-experimental settings. Examination of plausible rival hypotheses (Riechardt, 2000; Rindskopf, 2000) provides another mechanism for studying and validating alternative methods.

For instance, with reference to our Center's Goal 1, replication of findings about the growth of students with disabilities across states, cohorts, and/or summative and formative datasets will serve as an indication of inference validity (e.g., that growth for special education students is more similar to general education growth in mathematics than reading), and that rival hypotheses are implausible (e.g., the observed differences in growth are not due to artificial variations in a particular state's measurement or scaling of reading and mathematics).

With reference to our Center's Goal 2, surprisingly little literature examines the reliability and validity of school effectiveness measures and growth models using actual state and school accountability data. In the few studies that have reported on relationships between confounding factors such as student socioeconomic status and school outcome measures, strong relationships between certain confounding factors have been found for some measures (e.g., status scores, percent proficient: Hill & DePascale, 2003; Linn & Haug, 2002; Sammons, Mortimore, & Thomas, 1996; Teddlie, Reynolds, & Sammons, 2000) and little association found for other measures (e.g., growth: Ballou, Sanders, & Wright 2004; Bryk et al., 1998; Ponisciak & Bryk, 2005; Stevens, Estrada, & Parkes, 2000; Stone & Lane, 2003; von Hippel, 2009).

Given that any accountability method has strengths and weaknesses and the high stakes application of school effectiveness models, it is critical to evaluate the reliability, validity, and utility of alternative accountability designs and methods. Such evaluation should include assessment of consequential validity (Messick, 1989) and critical comparison and evaluation of alternative methods of measuring school effectiveness. One of our goals in this research is to examine the plausibility of rival hypotheses for the performance of certain measures of school effectiveness for students with and without disabilities. Specifically, we gather evidence on different classes of covariates as related to alternative measures of school performance. We hypothesize that measures of school practice and policy presumed to have an impact on student learning, like outcomes from interim assessments and OTL, should show positive relationships with outcome measures in proportion to the validity of the measure. Conversely, covariates that reflect confounding factors (e.g., school poverty, ethnic composition of the school) should show little or no relationship with valid measures of school effectiveness. As Figure A.5 in Appendix A illustrates, we can compare and document differences in how the approaches for representing school effects (in this case, status versus MGM) relate to sources of construct irrelevant variance. Evidence is needed that demonstrates that accountability methods validly capture the effects of school policy and practice, and are relatively immune to construct irrelevant sources of variation.

Rather than trying to find a single, optimal model, we plan to provide evidence for the degree of reliability, validity, and utility of each model we evaluate so educators and state educational agencies can make informed decisions about the methods they adopt for their populations and purposes. Our research program will document the efficacy of each model for school accountability, for representing individual growth trajectories, and for understanding the growth of various populations subsumed under the broad category of special education.

Research Plan for the Focused Program of Research

As indicated from the outset of this proposal, our research focuses squarely on achievement growth measures and relationships among key demographic and instructional variables related to growth in reading and mathematics. Specifically, we have access to the extant longitudinal (1997-2010) and future (2011-2015) achievement data for a diverse population of students from NC, AZ, OR, and PA. We plan to start the primary phase of our research (Years 1-2) with the NC dataset. Given the history of NC's end of grade testing system and the high-quality data management provided by the North Carolina Education Research Data Center (NCERDC), the growth modeling research with the NC dataset comprises our ***Cornerstone Study***. With multiple years of achievement data for roughly 13,000 students with disabilities per grade, the NC dataset provides a rich and robust dataset for growth analyses. We have already begun working with the NC dataset to gain information about the sample of students with disabilities, ease of using the database, and feasibility of using various growth models with the data. The design and analytic methods for the NC studies is then extended, and in many ways replicated, with extant longitudinal (2006-2011) and future (2012-2015) achievement data from the general education test for cohorts of students with and without disabilities from AZ, PA, and OR resulting in our ***Multi-State Extension Study.*** Our primary phase of research also features groundbreaking growth analyses with over 3 million students who have two or more years of test results on nationally used interim assessments (i.e., easyCBM and NWEA MAP) administered during the period 2006-2010. The growth modeling research with these measures comprise our ***Interim Assessments Study.***

In a secondary and overlapping phase of supplemental research (Years 2-5), we plan to (a) examine multiple measures and potential predictors of achievement that provide actionable information for teachers, such as opportunity to learn and repeated interim measures of reading and mathematics, and (b) focus on growth model analyses with extant data sets from statewide alternate assessments for students with significant cognitive disabilities. Information from these measures will expand understanding of students' growth rates, patterns of achievement on statewide summative assessments, and variables influencing growth. This secondary phase of research comprises our ***Multiple Measures Validation Study*** and ***Alternate Assessments Study***, respectively, adding important information for understanding growth and building a more inclusive validity argument for accountability purposes.

Figure 4 summarizes our overall research plan for our 5 study strands and the explicit connections among our 6 research questions and 11 key datasets. Each state's general achievement test and alternate assessment data set will be comprised of extant test results based on at least four years (2006-2011) of testing prior to the start of our research and these datasets will continue to expand, yielding a second wave of growth analyses based on longitudinal cohorts who completed their state's test during the period 2011-2015. Thus, both within state and across state replications of our growth modeling analyses with both the general achievement and alternate assessments will be conducted for large samples of students with and without disabilities in grades 3-8. In the reminder of this section, we address key data and sample issues.

Access to Data and Overview of Samples of Students with Disabilities

We have access and use of extant data from statewide general and alternate assessments (see letters). NC Educational Research Data Center (NCERDC) supports our NC Cornerstone dataset for growth modeling studies. See ***Appendix A*** for letters. Each of these states either has a growth model (AZ, NC, PA), or vertically aligned tests compatible with the application of multiple types of growth models (OR) and all have the NCLB alignment elements and foundational elements specified by the USDE (November 2005). NCERDC houses annual NC test data from 1992-2010 and its store of data will continue to grow and be accessible throughout the existence of the NR&D Center. On average, 62,000 SWDs in grades 3-8 have been assessed annually in NC. In other partner states where our Extension Study and Multiple Measures studies will be conducted, we have access to extant data for 4 years (2007-2010) that consist of nearly 66,000 (AZ), 36,000 (OR), and 132,000 (PA) SWDs assessed annually in grades 3-8.

Figure 4. Research Plan

Study	Primary Research Question(s) Addressed*	Datasets	Year
Cornerstone Study	1. What is the natural developmental progress in achievement for students with disabilities? 2. What models best characterize achievement growth for students with disabilities who are participating in general achievement tests? 3. How do various growth models represent school effects for students with and without disabilities, and how do results compare to those derived from status models now in use? 4. What are the reliability and validity of estimates of school effectiveness for students with disabilities produced by alternative growth models and how are these estimates influenced by contextual differences among schools and students?	NC Statewide Achievement Test (Grades 3-8: Reading & Math)	Years 1-2, 5
Multi-State Extension Study	Across 3 partner states: 1. What is the natural developmental progress in achievement for students with disabilities? 2. What models best characterize achievement growth for students with disabilities who are participating in general achievement tests? 3. How do various growth models represent school effects for students with and without disabilities, and how do results compare to those derived from status models now in use? 4. What are the reliability and validity of estimates of school effectiveness for students with disabilities produced by alternative growth models and how are these estimates influenced by contextual differences among schools and students?	AZ Statewide Achievement Test (Grades 3-8: Reading & Math)	Years 2 & 5
		OR Statewide Achievement Test (Grades 3-8: Reading & Math)	
		PA Statewide Achievement Test (Grades 3-8: Reading & Math)	
Interim Assessments Study	For 2 interim assessments: 1. What is the natural developmental progress in achievement for students with disabilities? 5. How do results from different types of interim assessments of students' achievement meaningfully contribute to a model of academic growth for students with disabilities?	NWEA MAP in reading and mathematics Grades 1-10 easyCBM measures in reading and mathematics Grades 1-8	Years 1 & 2
Multiple Measures Validation Study	5. How do results from different types of interim assessments of students' achievement meaningfully contribute to a model of academic growth for students with disabilities? 6. How can information about opportunity to learn and achievement growth be used to enhance academic outcomes for students with disabilities?	Multiple Measures OTL+ Interim Assessments [CBM + Multiple Choice Tests] + OR & PA Summative Assessments (Grades 3-5: Reading & Math)	Years 2-4
Alternate Assessments Study	1. What is the natural developmental progress in achievement for students with disabilities? 2. What models best characterize achievement growth for students who are participating in alternate assessments?	Alternate Assessment (Grades 3-8 for Reading & Math): AZ • NC• OR • PA	Years 3–5

Question numbers refer to primary research questions listed on p. 2-3.

In addition to the statewide achievement test datasets, we also have data sharing commitments for two interim assessments: NWEA's Measures of Academic Progress (MAP) and the University of Oregon's easyCBM. NWEA MAP is a computer adaptive test in reading and mathematics that uses Item Response Theory (IRT) to locate both students and items on a common scale. The test also provides normative growth, both within and across years. NWEA'S MAP also includes diagnostic information on item difficulty and will provide us access to nearly 10 million students in the NWEA Growth Research Database. easyCBM is a progress monitoring assessment system, launched in 2006 as part of an OSEP funded *Model Demonstration Center on Progress Monitoring*. Alternate forms are available for both benchmarking at the beginning of the year and monitoring growth within the year, using items that have been assembled into alternate forms using IRT. Measures and graphic outcomes are available in grades K-8 with the system currently being used by 50,000 teachers and nearly 700,000 students in schools and districts participating from all 50 states in the country.

In summary, for our initial 3 featured studies (i.e., ***Cornerstone Study, Multi-State Extension Study, and Interim Assessments Study)*** we have approved access to substantial extant and expanding databases of achievement test results used for NCLB accountability purposes in our 4 partner states and an extant nationwide database from schools who voluntarily administered interim assessments. Access also has been secured to the extant and expanding databases of our partner states' alternate assessments for use in the ***Alternate Assessments Study***. The only database that needs to be developed is for our ***Multiple Measures Validation Study*** and we have letters of support for conducting this study from a sample of school districts in OR and PA (See ***Appendix A*** for school district letters). The sample for this latter study provides common cohorts of students with and without disabilities, with known opportunities to learn aligned test content, who complete a common set of interim assessments and one of two state general achievement tests.

Variables and Description of Key Measures

The primary descriptive variables of interest in the research to be conducted by the Center are state, school district, grade, sex, racial group, free and reduced price lunch status, and disability type. It is expected that sample size limitations will require collapsing the 13 federally recognized disability types into fewer categories, such as learning disabilities, mental retardation, emotional/behavior disorders, and speech/hearing impairments. The primary outcome variables of interest are reading and mathematics achievement, measured with the tests used in each state that yield total scores for reading/language arts and mathematics. Each statewide general achievement test is briefly described next. The descriptions of these same states' alternate assessment are provided as Figure A.6. in Appendix A.

North Carolina End of Grade Tests in Reading and Mathematics (NC EOG-R and NC EOG-M). All students in grades 3-8 take annual reading and mathematics tests aligned with the grade-level curriculum. The reading test contains approximately 50 reading comprehension items based on 10 reading passages at each grade. The Mathematics test consists of 82 word problems with calculator active and inactive portions. Shorter reading and mathematics pretests are given at the beginning of third grade and are used to assess 3rd grade growth. At each grade, vertically scaled developmental scores can be converted to four grade level proficiency levels. The developmental scale scores also can be converted to norm-referenced scores using grade-level results from a standard setting year or the current year. Students whose actual scores meet or exceed their predicted scores based on earlier scores have made expected growth. Schools where mean actual growth meets or exceeds expected growth have met the school growth standard.

Arizona's Instrument to Measure Standards (AIMS). The AIMS is a dual-purpose assessment consisting of criterion-referenced and norm-referenced items across grades 3-8 and high school. For grades 3-8, the AIMS combines reading, writing, mathematics, and science items with items from the Stanford 10th edition. The Stanford 10 is given exclusively (i.e., without AIMS items) in grades 2 and 9 and yields national percentile rankings in reading, language arts, and mathematics. The AIMS Reading test contains approximately 64 reading comprehension items based on 11 reading passages at each grade. The Mathematics test contains approximately 90 multiple-choice items for 3-8 and 100 for high school. AIMS scores can be converted to four proficiency categories. To determine growth, AZ uses an index called the measure of academic progress which is based on the difference between a student's individual growth expectation based on a regression prediction and the student's actual growth. School performance profiles are based on a weighted formula that includes: (a) student AIMS performance in reading, writing, and mathematics for the current year; (b) change in the percent of students proficient on the AIMS from the baseline year; (c) student performance on the state English language proficiency assessment; (d) the school's AYP status for the current year; and (e) the average academic progress index across students and subjects for the school. The formula yields a scale score that is transformed to a rating on a 6-point scale from Failing to Excelling.

Oregon Assessment of Knowledge and Skills (OAKS). The OAKS Online test is a computer adaptive test of reading and mathematics in which items are selected according to each student's demonstrated ability: The number of items taken by a student vary as the test is terminated as soon as a reliable estimate of performance is attained. Typically, students take from 35 to 50 items. Students may take online assessments up to three times per year between October and May. This feature allows our research to address growth within the academic year, from October, when the test window opens to May, when the test window closes. Approximately 85% of the students in Oregon take their allotted 3 attempts. On average, students finish the OAKS Online Assessment in 60 – 75 minutes. A paper-pencil version is allowed for students whose IEP or 504 Plan indicates this need. The score from the multiple-choice test is a Rasch scaled score that is vertically articulated across grades 3-10 with the lowest score being approximately 195 and the highest score being 260. Cut scores for each grade level (used to place students in four proficiency categories) begin at grade 3 (201) and extend to grade 10 (239) in approximately 7-point increments per grade.

Pennsylvania Value-Added Assessment System (PVAAS). PVAAS is available for reading and math in grades 4-8. PA Department of Education's approved growth model, PVAAS is a system that began pilot implementation in 2002 and evolved to statewide implementation during the 2005-2006 school year. The PVAAS model uses all available student data as part of its analysis and provides two types of information, *value-added (or growth) data* on cohorts of students and student level *projection data*. The value-added analysis uses available data from previous years to help schools evaluate how much cohorts of students have over- or under-performed in relation to the prediction based on previous years data. Projection data can be used for intervention planning and resource reallocation. PVAAS does not provide measures of growth on individual students.

Other measures of achievement and opportunities to learn. In addition to the state measures of achievement, we will analyze extant data and collect new longitudinal data concerning students' reading and mathematics skills and opportunity to learn such skills. Specifically, interim assessment measures from NWEA's MAP and easyCBM will be used to measure students' skill development and achievement multiple times within each year. To document students' opportunity to learn standards-aligned content, teachers will complete MyiLOGS at the beginning and end of academic years to yield an alignment index for the intended, enacted, and assessed curricula. A follow-up individual measure of opportunity to learn this same content can be conducted with a subset of students in each classroom at the same time they complete interim assessments. For example content from each of these measures see ***Appendix B.***

Focused Research Studies: Systematic Approach and Anticipated Refinements

In the following sections, we provide details for each of the planned and supplemental studies introduced previously. It is important to emphasize that although our systematic series of studies are grounded in extant research and guided by research questions and data analysis plans, they will also be informed by feedback from experts on our National Advisory Panel (NAP) about related work, by concerns from our Teacher Advisory Group (TAG), and by what we learn as we proceed through the analyses. As an example, we anticipate that our initial analyses in NC will yield lessons that will help focus our subsequent work and also sensitize us to issues we had not fully anticipated, but cannot ignore as our research evolves. For example, although our plans for the Cornerstone Study include longitudinal growth modeling across grades 3 to 8, we may find NC's developmental scales do not have the psychometric qualities that permit growth modeling across such a wide age span. In that case, we would limit or exclude those analyses in the Cornerstone Study and also where appropriate in our Multi-State Extension Study. Similarly, we may find that there are so few students in a particular disability category (e.g., students with severe cognitive impairments) participating in the general education assessments that we drop this group from our analyses with large scale assessments and examine growth for this group only in the Alternate Assessments Study or Interim Assessments Study. Both of these examples describe situations where we would narrow our analyses based on early findings; however, we may also expand our analyses based on early findings. For example, in 2008 in NC 37% of students were retested in reading largely because they did not meet proficiency standards. Allowing these retest results to be counted markedly changed schools reaching their achievement proficiency targets for NCLB (Bonner, Hui, & Latifi, 2009). If student or school growth looks quite different when retests are included in the Cornerstone Study analyses, this finding will prompt additional analyses that examine the impact of retesting on student growth in the Multi-State Extension Study. In summary, we have a clear vision for a series of related studies and yet anticipate that refinements which improve our work will emerge from our experienced team members' involvement, the developing research literature, and lessons learned from ongoing data analyses.

Cornerstone Study

The first set of question-driven studies will make use of the extensive NC statewide achievement data available from the NCERDC. The important issues of statistical power and missing data are discussed within the context of the NC dataset, but the methods used are relevant to all our datasets and studies. Later in Year 2, many of these same analyses will be replicated in our Multi-State Extension Study with extant AZ, OR, and PA data to assess the generalizability of the results. The power calculations refer specifically to the two MGM analyses in the Cornerstone Study where growth is modeled at the student and school levels.

Statistical Power and the NC Longitudinal Dataset

The initial growth analyses for the Cornerstone Study are based on a dataset consisting of five cohorts of NC students with longitudinal data from grades 3 to 8. The availability of large samples for the overall (SWD and non-SWD) analyses allows for very good statistical power in detecting both student- and school-level predictor effects as well as accurate estimation of individual school effects. At the elementary school level, we have approximately 90,000 non-SWD students and 13,000 SWD per cohort across 1,100 schools. In middle school we have approximately the same numbers of students across 550 schools. We approximated power analyses for our MGM analyses with the software Optimal Design for Longitudinal and Multilevel Research (Spybrook, Raudenbush, Congdon, & Matinez, 2009) using the multisite trials module. Even in the presence of sizable school effects related to SWD status (an intraclass correlation of .20), and assuming random effects at the school level related to SWD status, we are able to detect even small effects (i.e., an effect size of .05) with power >.99.

Detecting and Managing Missing Data: Insights from the NC Dataset

Missing data and student attrition have the potential to seriously compromise the validity of longitudinal research designs (Goodman & Blum, 1996; Shadish, et al, 2002). While longitudinal designs provide some control over the nonrandom assignment of students into schools, if differential attrition across schools results in missing accountability data, systematic bias can occur. It is unclear at this time what the effects of bias actually are in the context of school effectiveness studies (Zvoch & Stevens, 2005). If some students are not included, then accountability is evaluated for some but not all students. On the other hand, attendance and participation in schooling and assessment serve as markers of opportunity to learn. Students who are less frequently engaged in schooling, who leave one school system for another, or who drop out have received unequal instructional opportunities and intervention.

Understanding the pattern and amount of missing data for all our datasets is therefore important and will provide insights on how attrition impacts estimates of school effectiveness. The few studies published on attrition effects in school accountability show differential patterns of transiency across students and schools and suggest that transiency is non-random and associated with certain characteristics including student and school economic disadvantage and student special education status (Rumberger, 2003; Rumberger & Thomas, 2000; Zvoch & Stevens, 2005).

Williamson et al. (2006) examined missingness patterns in the NC dataset (i.e., panels spanning 1995-2000, 1996-2001, 1997-2002, 1998-2003 and 1999-2004). They examined the temporal patterns of missingness, correlations between missingness and demographic variables, and the composition of samples for students with missing data and for students with complete data. There were statistically significant differences in demographic composition between the students with missing data and those with complete data with higher percentages of missing data for non-White students, students with disabilities (8.8% vs. 5.5%), and male students (52.6% vs. 47.6%). The largest difference occurred for students eligible for free or reduced price lunches (missing data 40.3% FRL, complete data 33.0% FRL).

Missingness and attrition are more problematic for the estimation of school effects. Student absences, transfer to other schools, mobility, and drop-out reduce the complete sample of students in a longitudinal data set for a school, especially when assessment occurs only once per year as in most summative accountability systems. Given the results of the studies cited, it is unlikely that data will be *missing completely at random* (MCAR). If data are *missing at random* (MAR) then the probability of missingness depends on the observed data and likelihood-based methods (e.g., FIML) produce unbiased estimation when joint response distributions are correctly specified (Fitzmaurice et al., 2004; Singer & Willett, 2003). When data are related to the missing values they are *not missing at random* (NMAR) and almost all standard methods of longitudinal analysis will produce biased results (Fitzmaurice et al., 2004).

In our studies and analyses, we will first carefully describe patterns of missing data and student attrition and examine whether missingness is related to student characteristics or to certain types of schools. Following this, for our studies that focus on individual student growth trajectories we will estimate parameters using all available data for each student and conduct analyses based only on observed data. Another means for dealing with missing data is to replace missing values with statistical estimates. Several options are available for this purpose including data replacement methods and model based approaches (Little & Rubin, 1989; Schafer & Graham, 2002). We will apply both of these approaches using imputation procedures to replace missing data before analysis and then in an alternative approach we will use empirical Bayes estimation procedures that apply Full Information Maximum Likelihood (FIML) to estimate intercept and slope parameters for each individual or school (Raudenbush & Bryk, 2002). A strength of the latter approach is that students do not need to be assessed on each measurement occasion and assessments do not need to be administered at the same time for each student. We will compare the results of all three approaches (observed data, imputed replacement, FIML) and evaluate the degree to which inferences and conclusions differ depending on the approach implemented.

It should be noted that we expect these issues to be somewhat more complex when estimating school effects. As previous studies have shown, there are likely some significant differences between the composition of student cohorts at one point in time in comparison to those present at multiple measurement occasions. As a result, we will expend particular care and effort to document and model the effects of missingness and attrition on the estimation and characterization of school effects. We believe it will also be important to carefully attend to these effects for particular disaggregated student groups, especially SWD and economically disadvantaged students. A goal of these analyses will be to determine the degree to which the reliability and validity of inferences about school effectiveness are affected by missingness or attrition.

Cornerstone Study: Preliminary Analyses

A critical first step in building knowledge about the achievement growth of SWDs is a thorough exploration of the characteristics of the data. A number of unique issues this group poses must be addressed when seeking to describe SWD achievement growth or validly estimate SWD growth for accountability purposes. At the individual student level, instability in who is served in special education (Ysseldyke & Bielinski, 2002) raises the issue of who we define as a SWD when tracking longitudinal outcomes. An important issue is whether grade level tests are sensitive measures for students who are achieving below what is expected for their grade. At the aggregate level, research has suggested that districts with higher special education prevalence rates tend to identify less impaired students as disabled (Singer, Palfrey, Butler, & Walker, 1989). Another issue is small sample sizes of SWDs per school and grade. For example, our preliminary analyses using two years of NC achievement data indicate that over 70% of elementary schools have less than 30 SWDs across grades 3-5. These small sample sizes can result in more extreme scores in some growth models (Linn & Haug, 2002) and pull scores toward the mean in others (Amrein-Beardsley, 2008). Understanding the impact of small sample sizes on the validity of various value-added measures of schools' effectiveness with SWDs has implications for the school subgroup reporting in NCLB. Preliminary analyses with the NC Cornerstone dataset will answer the following nine questions about SWDs and the schools that serve them:

1.1 Defined as at least one year where an individual received special education services, what is the "lifetime" prevalence of SWDs in grades 3-8?

1.2 How do the odds of being in special education, and in particular exceptional children's categories (e.g., speech/language impaired) vary by grade?

1.3 What are the modal entrance and exit patterns for special education?

1.4 How stable are classifications to the various exceptionality categories? Do most students stay in one category or change categories? Does the stability of classification vary by category?

1.5 How do students with disabilities differ from students without disabilities on basic demographic variables that may affect interpretation of outcomes, including changes in schools, grade retentions, retests, and absences?

1.6 What are the conditional and unconditional distributions of current grade level developmental scale scores based upon students' test results in the previous grade?

1.7 What is the typical cohort size by school and grade for SWDs at the elementary and middle school level?

1.8 Do school-level variations in who receives special education affect school outcomes for SWD? For example, are schools with a high propensity to label more likely to identify higher achievers, and as a result show higher percentages of SWDs functioning on grade level?

Cornerstone Study: The Natural Developmental Progress of Students with Disabilities

After obtaining the basic demographic information and considering its implications for our planned analyses, we will characterize students' achievement growth using two strategies. First, we will use hierarchical linear modeling to model student growth and proficiency. Second, we will calculate grade-level growth and SWD/nonSWD achievement gaps using Bloom et al.'s (2008) effect size approach and characterize SWD growth with a growth percentiles approach.

Two-level MGM. To effectively characterize growth as well as identify important student- and school-level correlates of growth, we will use Multilevel Growth Modeling techniques. One challenge in the implementation of MGM with our longitudinal data is that most students transition from an elementary school to a different middle school. Therefore we will implement two different sets of MGM models: (a) a two-level model (time, student) that ignores school effects and is used primarily to describe the natural developmental progress through the estimation of individual student growth trajectories across all grades (up to 7 measurement occasions) and (b) a school effects, three-level model (time, student, school) with separate models for the elementary and middle school data. In the three-level model, there will be four measurement occasions per pupil, including a third grade pretest for elementary school students or the fifth grade exit score used as a pretest for middle school students.

In both MGM approaches, reading or math test scale scores are the outcome measure. Models will be used that include all students (both SWD and non-SWD) as well as models for SWD students only. Model development will proceed through a model building strategy (Hox, 2002), in which predictors are added incrementally, through consideration of statistical significance, effect magnitudes, and influence on the overall model deviance. Important considerations for the level 1 model (repeated measures level) concern the appropriateness of a linear growth trajectory (or alternatively quadratic or higher order) as well the possible addition of time-varying covariates. A challenge in modeling the growth of SWD's concerns how to define SWD as a student's status can change over time. We will test two models, constant SWD status over time and SWD status treated as a time-varying covariate that allows SWD status to change at each measurement occasion reflecting the more complex entrances and exits occurring in the schools.

In the first MGM approach, time will be coded such that the pretest given at the beginning of 3rd grade will correspond to the intercept. Of particular interest are not only the average trajectory and variability of trajectories across students, but also the relationship between student-level variables (e.g., SWD category) and growth. While school effects are ignored in this model, appropriate corrections to standard errors will be applied (Huber, 1967; White 1982) so as to account for spurious deflation due to school-related clustering. An important consideration relates to whether SWDs and non-SWDs grow at similar rates (differ only in intercepts, not slopes), and in the three-level models, whether there are school differences in SWD growth rates or in the SWD-nonSWD growth rate difference.

Below is an illustration of one type of two-level, conditional MGM we anticipate using where level 1 represents measurement occasions and level 2 represents students:

Level-1 (Measurement Occasion):

$$Y_{ti} = \pi_{0i} + \pi_{1i}(time) + e_{ti} \qquad (1)$$

Level-2 (Students):

$$\pi_{0i} = \beta_{00} + \beta_{pi}(a_{Pi}) + r_{0i} \qquad (2a)$$

$$\pi_{1i} = \beta_{p1} + \beta_{pi}(a_{Pi}) + r_{1i} \qquad (2b)$$

Where: Y_{ti} is the outcome (i.e., score) at time t for student i

π_{0i} is the status of student i at the first measurement occasion

π_{1i} is the linear growth rate over time for student i

e_{ti} is a residual term representing unexplained variation from the latent growth trajectory

r_i are residual terms representing unexplained variation in student growth parameters

a_{Pi} are student level predictors

As discussed earlier, in some models we will also include squared and cubed time variables to test the functional form of the relationship between time of measurement and outcome performance. Equations 2a and 2b show the modeling of each student's intercept and growth rate. Predictors at this level will include student characteristics and covariates such as SWD status, cohort membership, gender, race/ethnicity, disability classification (dummy coding to represent different disabilities), free/reduced lunch status, Limited English Proficiency status, etc. In this example, we assume SWD to be a level 2, student-level variable (constant over time), although as noted above, we will also explore models treating SWD as time-varying covariate included at level 1. Our questions of interest generally relate to the effects of other predictors that can be similarly entered at the student level. For example, the question of whether differential effects are observed for different subcategories of SWDs and how these subgroups should be constituted can be studied by entering dummy-coded variables related to subgroup designation as predictors of student intercepts and slopes.

Three-Level MGM. The next group of MGM models will incorporate school effects into a three-level structure (time, student, school). Because of the shifting enrollment of students in schools, separate models will be used for the elementary and middle school data. Elementary models will analyze data from grades 3, 4, and 5 while middle school models will analyze data from grades 6, 7, and 8. Scores from the fall, grade 3-pretest and from the end-of-grade test for 5th grade will be defined as the intercept in each model for elementary and middle school grades, respectively, resulting in four measurement occasions for each set of models. In these three-level models we will explore the relevance of school characteristics to student status, growth, and cross-level interactions with student level predictors.

The MGM analyses incorporating school effects will be built from a similar type of growth modeling structure to that described earlier, but incorporate a third level representing school. Because separate analyses will be performed for elementary and middle schools, growth will be examined over a more limited grade range, and thus the functional form of growth may possibly differ from the models above. The illustration below is for a three-level MGM model assuming linear growth.

Level-1(Measurement Occasion):

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(time) + e_{tij} \tag{3}$$

Level-2(Students):

$$\pi_{0ij} = \beta_{00j} + \beta_{pij}(a_{Pij}) + r_{0ij} \tag{4a}$$
$$\pi_{1ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij} \tag{4b}$$

Level-3(Schools):

$$\beta_{p0j} = \gamma_{000} + \gamma_{pqs}(W_{sj}) + u_{00j} \tag{5a}$$
$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j} \tag{5b}$$

where: Y_{tij} is the outcome (i.e., score) at time *t* for student *i* in school *j*
π_{0ij} is the status of student *ij* at the first measurement occasion
π_{1ij} is the linear growth rate over time for student *ij*
a_{Pi} are student level predictors
W_{sj} are school level predictors
e_{tij} is a residual term representing unexplained variation from the latent growth trajectory
r_{ij} are residual terms representing unexplained variation in student growth parameters
u_j are residual terms representing unexplained variation in school growth parameters

One issue to be explored further at the student level is the determination of which effects should be treated as random across schools. We will use variance components analysis to evaluate whether there is statistically significant variation between schools (Raudenbush & Bryk, 2002). For example, we may find that SWD status slopes vary from one school to another. If so, we will then attempt exploratory analyses to attempt to understand school level differences. Deviance testing (Hox, 2002; Sniders & Bosker, 1999) will be used to evaluate model improvement as we add predictors from unconditional models to more complex models. School-level effects on student intercepts, slopes, and predictors will also be studied using school-level predictors including but not limited to community type, school size, pupil-teacher ratio, percent free/reduced school lunch, percent special education students, school calendar (traditional or year-round), and staff mean education and experience level. An important consideration in evaluating this set of three-level MGM models concerns the reliability of student- and school-level effects. We will examine reliability of model parameters as well as calculating the intraclass correlation (ICC) of intercept and growth parameters. An important outcome of the three level MGM analyses will be estimation of individual school effects. From our final MGM models we will obtain estimates of school intercepts and slopes, in the form of Empirical Bayes' (EB) and ordinary least squares (OLS) estimates. Such estimates, based on three longitudinal cohorts from each school, will provide an additional standard against which annual estimates of school effects can be compared in our pattern matching strategy for examining the validity of growth approaches and specific models within the three broad growth model approaches to be examined. Key questions to be answered include:

2.1. How is growth for SWDs across grades 3-5 and 6-8 best characterized?

2.2. What are student-level and school-level predictors of growth and does disability status interact with these predictors?

2.3. How much do proficiency and growth outcomes for SWDs vary between schools, in proportion to the variation observed between students?

Effect size benchmarking. A second way of describing typical change in achievement across grades for groups of children is ES benchmarking (Bloom et al., 2008). Achievement growth at each grade in mathematics and reading will be expressed as an ES based on the change in mean score from grade to grade, divided by the pooled standard deviation. In this analysis, the two approaches to achievement size benchmarking described in Bloom et al. will be applied to the NC EOG developmental scale scores in reading and mathematics. Results from these analyses will be helpful in providing another way of describing the growth of SWDs and for putting ESs for interventions for SWDs at different grades in context. Questions to be examined in this analysis include:

3.1. What are the normative expectations for academic growth in reading and math by grade for all students, and SWDs?

3.2. Do specific disability subgroups deviate from these normative expectations?

3.3. Stated as an ES, what is the achievement gap between SWDs and those without disabilities, by grade? How does this gap compare to other achievement gaps of national interest (e.g., Black-White, Eligible-Ineligible Free/Reduced Price Lunch) across grades?

3.4. Do entrances and exits from special education distort ES's and grade level trajectories examined cross-sectionally compared to ES's from a longitudinal sample?

Cornerstone Study: Comparisons Among Models

The NCLB status model and representative models from each of the 3 growth approaches described in the Significance section can all be applied to NC EOG test data (see Figure 5). Our examination of the suitability of different models for particular accountability purposes begins with the application of exemplars of the 3 growth approaches (and the NCLB status model) to a common dataset. Specifically, for our initial examination of the validity of different growth models, we will contrast annual student and school-level results obtained from NC's NCLB status/proficiency model with (a) NC's current growth model (an example of a residual gain model), (b) a transition matrix that shows transitions between the four levels of NC's reading and mathematics tests, (c) AZ's VAM model, (d) PA's multilevel VAM models, and (e) the three-level MGM developed in the analyses described in the previous section. This research strategy for investigating how different school effects models perform has a long history (Dunn & Allen, 2009; Hanushek & Rivkin, 1996; Linn & Haug, 2002; Richards, 1975; Tekwe, et al., 2004). In these analyses, we will compare student and school-level results for students with and without disabilities. Among the characteristics to be examined are sensitivity of the models to individual growth, parsimony (are the same results from complex models obtained with simpler or more inclusive models?), and "existence proof" (Linn, 2003). For example, if it rarely occurs that the actual growth for a subset of SWD's (e.g., those with cognitive impairments) is similar to that of general education students, it suggests this growth model (or the standard articulated within it) is not appropriate for this population.

Figure 5. NC Data Used for Common Growth Model Analyses

	NCLB	***ResVAM***	***TM***	***MGM***
Student-level Growth	None	Gain score adjusted for regression to mean and expected growth	Proficiency level last year vs. proficiency level this year (4 proficiency levels)	Individual Slope
School-level Growth	% Students AYP in current year; change in % proficient Y1 to Y2	Ave. residual gain score across students	Ave. change in level	School Ave. Slope

When applied to schools, the question of how to best measure growth is a question of the validity of the inferences made about the school based on the results from the particular growth model. As stated earlier, we will use a pattern matching strategy for evaluating validity of school effects measures (Shadish, Cook, & Campbell, 2002) where the pattern of results from each of the alternative models for representing growth is examined relative to expectations for a valid measure of school effects for a particular purpose (Stevens, 2005). Among the overall criteria to be used for evaluating measures of school effects are whether the model: (a) produces results that are relatively stable across time (large annual fluctuations in school outcomes are unlikely to reflect true change) for students with and without disabilities, (b) evidences low correlations with predictors unrelated to school effects, such as the kinds of students served by the school, and (c) evidences higher correlations with predictors related to school policy and practices (e.g., staff educational level), particularly in comparison to status measures of student achievement. When different accountability uses require different inferences, we can investigate whether the use of particular growth models supports these inferences. For example, NC's accountability program provides annual salary increases to staff when their school meets a predetermined growth criterion. This particular use requires school growth results reliable at the school, but not at the subgroup level. As such, we would be most concerned with the reliability of school-level scores for this use rather than individual student growth scores.

Comparison of multi-cohort growth results with annual results. To allow the comparison of annual growth results with multi-year, multi-cohort results, we will use annual student data from the NCERDC database that completely overlaps with the achievement data used in the MGM longitudinal analyses. These contrasting school-level results (annual and cumulative across 3 cohorts) will be examined in a variety of ways to provide information on the stability and validity of annual growth results yielded from the different growth models. For example, the stability of results (Heck, 2006) will be examined by looking at the variability in school level results across the three years within each growth model. Models that show large fluctuations across the three years are more likely to have results influenced by sampling error or other artifacts.

These initial Cornerstone Study analyses take advantage of extant data in NC. For these initial analyses, we will use extant data from 1997 (year of the first longitudinal cohort's 3rd grade pretest and end-of-grade achievement results) to 2006 (year of fifth longitudinal cohorts 8th grade results). The particular years of test results selected: (a) reduces the number of students affected by the introduction of a new edition of either the mathematics or reading test, minimizing the use of scores equated across editions; (b) takes advantage of the 3rd grade pretest introduced in 1997; and (c) allows us to produce results in the first year by employing Schulte’s already created longitudinal datasets which cover two of the five years targeted. In Years 2-5, we will also continue to assemble longitudinal data on additional cohorts from NC using extant NC test data, as well as to-be-collected results from 2011-2015. In the last year of the Center, as part of our overarching strategy of pattern matching as one way of examining validity, we will test the generalizability of our findings across time, cohorts, and editions of the NC test by replicating key analyses from the first year where extant data were featured. The new data should also permit additional questions to be asked about SWDs. For instance, in 2006 information about special education students’ placement setting was added to the NC database. This information will allow setting to be examined as a predictor of growth in the Year 5 dataset.

Multi-State Extension Study

In Year 2, we will use the same overall analysis strategy described in the Cornerstone Study to examine student growth, and the validity and applicability of different options for measuring student growth and school effects, in AZ, OR, and PA. States vary in many ways, including demographic characteristics, educational systems, content standards, assessments, and procedures related to identifying and serving SWDs. For example, identification criteria for particular disabilities vary by state, affecting state prevalence rates within and across categories (Shattuck, 2006). Repeating the analyses done in NC with longitudinal samples from three additional states is an important examination of the generalizability of conclusions about testing practices, achievement growth, and school accountability across a wide range of contextual variables. The same, careful initial description of the sample in the Cornerstone Study, including SWD prevalence, patterns of entry and exit from special education, and distributions of scores, will be repeated to help understand the degree of state variation and what factors may be contributing to differences in findings for SWDs across states.

We also expect that large-scale test data from the additional states will allow us to address questions that could not be engaged in the Cornerstone Study because of limitations in the NC data. For example, within NC's testing program student scores cannot be linked to teachers, only schools. However, examining student growth by teacher is an important policy issue and there are critical research questions about the validity of valued-added and growth models for examining teacher effects, particularly in cases where responsibility for instruction is shared (McCaffrey et al., 2004; Tekwe et al., 2004), as is often the case with SWDs. Although NC's extant data does not allow this question to be addressed, Pennsylvania's growth model does. As another example, because Oregon's large scale testing program includes multiple tests per year, we will be able to examine the impact of using within and between year summative assessments to estimate school effects. We expect to use the datasets generated for the Multi-State Extension Study to address these and similar research questions relevant to key policy issues in education. Collection of statewide achievement test scores for elementary and middle school students in all states will continue during the period 2011-2015. The new data collected will be used to extend the databases of existing cohorts so that at least three cohorts cover grades 3-8 in each state and it will also provide for additional new elementary and middle school cohorts. These data will then be analyzed during Years 4 and 5 using each state's growth model and promising models identified in the earlier analyses.

Interim Assessments Study

Our 6 questions address summative statewide tests that are administered one time per year (with the exception of OR as noted elsewhere) and interim assessments (including opportunity to learn) that are administered multiple times (3 to 10+ times per year). This distinction is important for examining growth both within and across years. The use of interim assessments allows us to (a) model within year growth trajectories, (b) compare the trajectories over years obtained from the summative and the interim assessments, (c) examine various predictors of student and school characteristics to model growth on the interim assessments and opportunity-to-learn (e.g., minutes of instruction, tiers in response-to-intervention), and (d) use status and slope on these interim assessments as predictors (in conjunction with other important student variables) of both status and growth on the statewide tests. Tracking the developmental trajectory of individual students enables more precise estimation of student and program performance (Linn & Haug, 2002; Raudenbush, 2001), enables the separation of effects due to individual differences from effects due to schools and programs (Stevens, 2005; Zvoch & Stevens, 2003) and can serve as a robust means for assessing the impact of interventions on changes in student performance (Boyle & Willms, 2001). Furthermore, achievement, the outcome of interest for the study and evaluation of special education students' progress, is fundamentally a problem in the analysis of change that can best be addressed with longitudinal designs (Stevens & Zvoch, 2006).

In the present proposal, we have an opportunity to measure growth within the year on multiple occasions using the interim assessments and we can model growth across years as well. Thus, during the end of Year 1 and throughout Year 2, growth model analyses will be extended to interim assessments of achievement. Specifically, we will conduct growth analyses with extant datasets for nationally representative samples of elementary students with and without disabilities from our four partner states that completed interim assessments on two or more occasions (MAP or easyCBM) across one or more years. The Status, Residual Gain Score, and Multilevel growth models will be used to examine achievement growth within and across school using the interim assessments. The growth results from each of the interim assessments will cover grades 1-8, be based on three or four measures of achievement per year, and compared to the growth results based on statewide (summative) assessments. In addition, by having multiple assessments that span more than one school year we will be able to estimate the out-of-school changes in student achievement that occur during the summer. Research has documented the "summer slide" that often occurs during the summer break (relative to school year learning) and shows little or negative growth for students from disadvantaged families (Alexander et al., 2001; Burkham et al, 2004; Cooper et al., 1996; Downey et al., 2008; Downey et al., 2004). Using piecewise MGMs (see below), we will examine growth rates during each school year as well as the growth or decline that occurs during the summer break and also examine whether the summer slide is significantly different for SWD than non-SWD students. It is also important to recognize that these "out of school" changes often masquerade as part of the school effect estimated in many growth models. We will estimate and describe the size and pattern of these seasonal effects (Downey et al., 2008).

Both the NWEA MAP assessments and the easyCBM tests have established vertical scales and item banks based on IRT methodology. Characteristics of these scales and item banks are (a) the item difficulty calibrations are sample free, (b) achievement level estimations are sample free, and (c) the item difficulty values define the test characteristics. These IRT properties have facilitated the development of item banks with content that extend beyond a single grade level or school district and enables measurement scales in reading and mathematics that extend from 1st grade to high school. Thus, these interim assessments are ideal for examining questions of achievement progress.

Data collection. In states with interim assessments and where our supplemental study involves collection of data on opportunity-to-learn, we plan to sample performance 3 or 4 times on these measures within the academic year (fall, winter, and spring) and apply HLM growth model within each year (3-4 occasions) as well as across multiple years (6-12 or more occasions). We will examine students' initial level of performance (intercept) and the rate of change (slope) and the relationship of student characteristics and covariates to these parameters. Careful examination of both intercept and slope will be important (Seltzer, Choi, & Thum, 2003) because interventions with special education students may not be able to exert limitless influence over the absolute level of the child’s functioning but can influence the learning rate of the child. In addition, taking the child’s initial status into account may be important both pedagogically and in terms of evaluating growth or progress for instructional or accountability purposes.

Analytic procedures. Multilevel modeling techniques will be used to analyze student growth trajectories (Raudenbush, 2001). Three-level longitudinal models will be used to estimate a growth trajectory for each student, to partition the observed parameter variance into its ‘within’ and ‘between’ school components, and to estimate each student’s initial achievement level and growth rate. Conditional three-level models will also be run to regress the achievement outcomes on student and school characteristics. These models are comparable to those described above for the summative growth models in equations 1- 5b with level 1 composed of a longitudinal growth model that fits a linear, curvilinear, or cubic regression function to each individual student’s scores over the multiple measurement occasions, level 2 composed of student level means, predictors, and residuals, and level 3 composed of school level means, predictors, and residuals. Level 2 predictors will include covariates that we wish to control (e.g., gender) as well as student level predictors that address research hypotheses (e.g., SWD status, OTL). Level 3 predictors will include relevant school characteristics or covariates (e.g., SES, school size). We also will include interaction terms that allow us to test the cumulative effect of interventions over time. For example, do outcomes increase more rapidly for schools with well-developed RTI models?

In addition to analyses that parallel those for the summative assessments, we will use piecewise models in order to specifically identify learning rates during different periods in the multiple years of schooling (Singer & Willett, 2003). For example, to model the effects of schooling across two years of elementary school with three interim assessments delivered in each year we will estimate multilevel piecewise growth models with three segments: (a) year 1 fall, winter and spring assessments, (b) the change in performance across the summer break between year 1 and year 2, and (c) the year 2 fall, winter, and spring assessments. Estimation of these growth segments during instruction and the summer break will allow us to address a number of important research questions including (a) what is the magnitude of summer slide? (b) how does the summer slide vary for different students (e.g., SWD vs. non-SWD, economically disadvantaged students, etc.)? and (c) how can an estimate of within-year learning purged of summer effects refine school effect estimates? The equations below specify this example of the type of piecewise growth models we apply on student (a_{Pij}) and school level (W_{sj}) predictors:

Level-1 (measurement occasions):

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(First\ Year) + \pi_{2ij}(Summer\ Break) + \pi_{3ik}(Second\ Year) + e_{tij} \qquad (6)$$

Level-2 (students):

$$\pi_{0ij} = \beta_{p0j} + \beta_{pij}(a_{Pij}) + r_{0ij} \qquad (7a)$$

$$\pi_{1ij} = \beta_{p1j} + \beta_{pij}(a_{Pij}) + r_{1ij} \qquad (7b)$$

$$\pi_{2ij} = \beta_{p2j} + \beta_{pij}(a_{Pij}) + r_{2ij} \qquad (7c)$$

$$\pi_{3ij} = \beta_{p3j} + \beta_{pij}(a_{Pij}) + r_{3ij} \qquad (7d)$$

Level-3 (schools):

$$\beta_{p0j} = \gamma_{000} + \gamma_{pqs}(W_{sj}) + u_{00j} \qquad (8a)$$

$$\beta_{p1j} = \gamma_{pq1} + \gamma_{pqs}(W_{sj}) + u_{10j} \qquad (8b)$$

$$\beta_{p2j} = \gamma_{pq2} + \gamma_{pqs}(W_{sj}) + u_{20j} \qquad (8c)$$

$$\beta_{p3j} = \gamma_{pq3} + \gamma_{pqs}(W_{sj}) + u_{30j} \qquad (8d)$$

We also will conduct an additional analysis that examines how well growth on the state summative assessments can be predicted using interim assessments and other measures. As described for our Cornerstone Study, we will use MGMs to estimate the growth over three years in student achievement as measured by the summative assessments. The school growth estimates provided by this analysis will be used as the outcomes in a sequent regression analyses that include initial status and growth estimates for each school based on the interim assessments as well as school characteristics predictors that describe and characterize each school (e.g., school SES, school size, percentage of special education students, etc.). Using these analyses we will be able to evaluate how well interim growth models and school context and characteristics can be used to predict and model school growth on the high stakes, summative assessments.

Other Center Activities: Supplemental Studies & National Leadership

The Center's PIs have a bold and broad vision of what it means to conduct research on the achievement growth of students with disabilities. In addition, they have a dynamic leadership team to ensure the research results and lessons learned guide future assessment and accountability practices by others across the country. In cooperation with IES project leaders, we plan to refine plans for "quick response supplemental studies" and a number of leadership activities. To illustrate our research and leadership vision and document our capacity for accomplishing this vision, we propose two supplemental studies designed to expand understanding of achievement growth for all students with disabilities beyond what is learned from tests administered once per year. Drawing conclusions about the academic achievement of students with disabilities from a single state test alone is questionable, if not unreliable and invalid (AERA, 1999; Hershberg, 2005). Thus, the need for supplemental studies that expand on our focused growth modeling studies. In the remainder of this section, we outline two proposed supplemental studies, provide evidence that we have the capacity to conduct this research, and share a vision of leadership activities and our capacity to accomplish the activities.

Supplemental Studies

Two supplemental research studies address the comprehensive nature of accountability systems and (a) the need for multiple measures that can be focused on interim classroom measures as well as summative state wide tests and (b) the degree to which the models of growth studied with the general education tests can also be applied to alternate assessments.

Multiple Measures Validation with Opportunity to Learn

It is commonly observed that annual statewide achievement measures do not provide sufficient detail about student performances in a timely fashion for educators to make informed instructional decisions for either individuals or groups of students. Therefore, to accomplish the goal of having meaningful data accessible for teachers and administrators, we consider it imperative to include other measures to supplement and complement statewide tests. The two primary types of measures include (a) opportunity to learn (OTL) as measured classwide with My instructional Learning Opportunity Guidance System or MyiLOGS, and (b) interim assessments, using a curriculum-based measure (easyCBM) and a brief multiple-choice test (NWEA MAP tests). The study is designed for three years (2013-2015) and collects OTL and achievement data 3 to 4 times within a year. This study will be conducted in approximately 200 classrooms. To increase the power of our analyses, we will conduct this study in two of our partner states (OR & PA). This will allow us to better understand growth that appears generalized (but confounded by teacher) versus un-confounded with specific instructional strategies (either opportunity to learn or instructional evaluation from progress monitoring). The unit of analysis for this component of our research will be the student. We will select students so that all disability types are represented over a range of classrooms and geographic areas within a state following the 4 steps below:

1. We will assemble a list of buildings and districts within the state; stratify district sampling into National Center for Educational Statistics (NCES) categories: city (large and mid-size), suburb (large and mid-size), town (distant, fringe, remote), and rural (fringe and remote). These categories will be collapsed into large, medium, and small based on student population to ensure a sufficient sample size.
2. We then will randomly sample 2-3 buildings within each district and contact the special education coordinator to solicit participation. Each teacher will select 2 SWDs from an assigned category (considering only high incidence categories of learning disabilities, emotional disturbance, and speech-language) and 2 non-SWDs. We will sample approximately 600 students (300 SWD and 300 non-SWD) in total.
3. We will provide a webinar to train teachers in the data collection procedures.
4. We will analyze each data set to document changes within the year (change in the opportunity-to-learn and growth in the interim assessment measures). For this *within instrument* analysis, we will use three levels: time (fall, winter, and spring), students, and schools. For level 2, we propose the following *student characteristics* to predict achievement (using two separate analyses with fluency and comprehension as the criterion variables): (a) English language learner, (b) special education, (c) gender, (d) race-ethnicity, and (e) grade. We also will use either outcome as a predictor for analyzing state test performance at level 3 where we focus on schools and programs: school size (number of students and teachers) and percent of students receiving Free or Reduced Price Meals (as a proxy for SES).

This study's focus on the integration of OTL and multiple measures of achievement provides an innovative approach to understanding academic growth and is consistent with some current education policy initiatives designed to provide more frequent and instructionally relevant feedback to educators and students about their learning.

Alternate Assessments Study

A significant supplemental research strand will be the documentation of status and growth for students with the most significant disabilities who are participating in alternate assessments for alternate achievement standards (AA-AAS). The extant (2006-2010) and prospective (2011-2015) datasets for all four states will be used, allowing us to investigate different types of assessment approaches – a multi-method approach where indirect observations, multiple choice, and performance tasks are used in AZ compared to performance assessments in NC, OR, and PA. Note that none of our participating states use a portfolio assessment approach; such an approach is impossible to scale adequately. For more details on each of these assessments, see Figure A.6 in Appendix A**.** In each of these states, three growth models will be considered: (a) NCLB Status, (b) residual gain scores (adjusting the change regressed on prior performance), and (c) transition matrix (tracking students' growth at various performance standard levels). In OR and PA, given their administration format includes well structured performance tasks with sufficient technical scales also allows us to examine a multi-level growth model with time at level 1 and student disability at level 2. Using these models, we will be able to evaluate models for documenting growth of students with the most significant disabilities. With the transition matrix, we will also be able to compare outcomes on the alternate assessment with those on the general education assessment.

Additional Supplemental Studies

The large amounts of data compiled for the primary analyses offer the promise to address many other questions of interest. Many potential studies can be envisioned. For example, by Year 4 we anticipate being able to do a cross-state comparison in reading by converting scores from Oregon to either MAP or easyCBM, given that both of these interim assessments are scaled using IRT, individual items are available, and a group of students can be identified who have participated in all three types of tests. This supplemental study allows us to examine natural developmental progress in reading, and examine reading and mathematics outcomes for SWDs across type of test (measure) on a common equated score scale. In later years of the Center, we will have high school outcome data (e.g., graduation, 10th grade reading level) for some students in the four states' longitudinal samples, permitting additional analyses of developmental progress.

Capacity for Completing Studies

The PIs and their team of fellow measurement and assessment experts collectively have more than 150 years experience conducting collaborative research. In addition, they have authored nearly 1,000 empirical publications, many of which focus on the assessment of students with disabilities. Capacity for completing large-scale, school-based studies, however, requires more than knowledge and experience. It requires management and teamwork skills, sound measures of the targeted behaviors or abilities, knowledge of research ethics, connections with communities of practice in educational agencies, and resources. Key resources essential to the completion of the proposed work include (a) expert data management and software handling and analyzing large, complex databases, (b) assistance of state department of education leaders in the recruitment of schools, (c) assistance of state assessment data experts in accessing test results for all participating students, (d) technical support for and use of online formative assessments of reading and mathematics skills in grades 1-8, and (e) expert consultation in the use of online surveys for documenting instructional content coverage. As indicated by staff vitae, the R&D Center PIs and their team of measurement and assessment experts have proven capacity to manage and complete large and complex research projects.

National Leadership Activities

We propose to undertake and support a number of leadership activities with the potential for national impact throughout the duration of the R&D Center. These activities will involve several technical papers, annual research briefings, webinars, presentations at national conferences, and a "growth modeling toolkit" for state assessment and special education leaders. We also plan to host summer institutes on growth modeling. The opportunities for enacting these leadership activities will be maximized by our partnerships with the Chief Council of State School Officers (CCSSO) and the National Association of State Directors of Special Education (NASDSE). These two professional organizations have long established channels for working with state leaders. The collaborative workgroups in CCSSO's State Collaborative on Assessment and Student Standards (i.e., Accountability Systems and Reporting, Assessing Special Education Students, and Technical Issues in Large-Scale Assessment) include the vast majority of the state leaders responsible for implementing inclusive assessments and reporting student achievement. Thus, we plan to direct a portion of our leadership efforts and materials toward these CCSSO workgroups and state leaders at NASDSE conferences via its web-based Project Forum.

Our initial (Years 1 & 2) leadership efforts will focus on the development of two white papers on growth expectations and modeling of achievement growth for students with disabilities; one of the papers will focus on technical issues and multiple measures of students who take the general achievement test and the second paper will examine similar topics relevant to alternate assessments for students with the most significant disabilities. A second featured leadership activity will be bi-annual research briefings via national webinars (Years 2-5). Both of these activities will allow us to use the vast talents of our research team and national advisory panel members to address technical and practice issues while also providing actual results from our empirical studies. A culminating leadership activity will be the collaboration with CCSSO to produce an "Implementer's Guide for Measuring the Achievement Growth of Students with Disabilities." CCSSO's current *Implementer's Guide to Growth Models* (2008) serves as a model to be updated and expanded so that the unique technical and practical considerations of including students with disabilities are understood and used to advance reports on achievement growth.

Capacity for Completing Leadership Activities

The PI team of Tindal, Schulte, Elliott, and Stevens individually and collectively has established nationally recognized records of research and grant management, scholarship, professional development training for educators and assessment leaders, and professional service to the fields of education and psychology. Over 20+ years, these PIs have had many opportunities to share knowledge and provide leadership to colleagues and teachers who work with students of all ages and abilities. As a result, the PIs' capacity and desire to provide leadership on issues of assessment and accountability for students in special education is substantial. The vita of each PI provides evidence of his/her advanced knowledge and skills in the areas of testing and measurement, inclusive assessment, and special education services. Leadership, however, requires more than strong research records. Leadership requires an understanding of the technical, social, political, and fiscal issues that influence an area of inquiry. It also requires excellent communication skills, high levels of organization and discipline, and the ability to collaborate to accomplish goals. Evidence of the leadership skills of the PI team has been amassing since the 1980s. Examples of leadership activities for each PI follow.

Tindal is the Department Head of Educational Methodology, Policy, and Leadership as well as the Director of Behavioral Research and Teaching (BRT), a research center housing federal and state grants and contracts (http://brtprojects.com). He has received $30 million of grant funds over 25 years, including multi-state projects as part of a research consortium with the CCSSO and has published over 80 peer reviewed papers three books, and scores of chapters and technical reports; this work focuses on curriculum-based measurement and validity issues and SWDs in large-scale testing programs. With four grants from the Institute of Educational Sciences (IES), he is currently expanding easyCBM to include more measures across more grades with more technical data supporting the use of these interim assessments for evaluating instruction and predicting state test proficiency outcomes. He has developed state alternate assessments and currently and operates them for OR and AK; in addition, he has helped developed and operated an on-line training for state alternate assessments in WY and PA. He also serves on state and national technical and advisory committees, including the National Alternate Assessment Center (NAAC), National Center on Educational Outcomes (NCEO), the National Study of Alternate Assessments, and technical committees in NY, OR, and AK.

Schulte has a longstanding interest in accountability and students with disabilities, particularly the use of value-added and growth models with SWD. She published one of the first studies looking at the use of a value-added accountability model and its impact on students with disabilities (Schulte et al., 2001), and also a study that contrasted student and school-level cross sectional and longitudinal achievement results for SWDs (Schulte & Villwock, 2004). She has worked with the NCERDC datasets since 2007, and has constructed a longitudinal dataset that allows the examination of important questions about the achievement of SWDs and their growth. She has also chaired a dissertation that used the NCERDC achievement data across 5 cohorts to replicate Hanushek, Kain, and Rivkin's (1998, 2002) examination of growth of students after entering and exiting special education (Ewing, 2009), and has completed state-level analyses examining the stability of school-level performance and growth outcomes for SWDs using longitudinal data from 500,000 students and 1100 schools.

Elliott has directed or co-directed two of the largest educational research institutions in the country; first the Wisconsin Center for Education Research at the University of Wisconsin and the Learning Sciences Institute at Vanderbilt. Both of these research enterprises serve more than 125 externally funded investigators. In moving to Arizona State University, Elliott will be directing a new Learning Sciences Institute, based on the successful research support models from Wisconsin and Vanderbilt. He also has served as Editor of *School Psychology Review*

(1984-1990), has guest co-edited other journal issues on testing accommodations (2005 issue of *Assessment for Intervention*) and alternate assessments (2009 issue of *Peabody Journal of Education*), and created CEC's first online assessment course *Assessing One and All*. He also has been selected to serve on a number national committees concerning testing and assessment of SWDs (e.g., National Academy of Sciences' Committee on Education Goals 2000 and Services to Student with Disabilities (1995-97), National Alternate Assessment Study Panel (2005-08), NAEP Technical Work Group (2005-07), and most recently ETS's Visiting Research Panel. Finally, Elliott has directed or co-directed 20 USDE grants/cooperative agreements, 6 involving the validity of testing accommodations and/or alternate assessments for SWDs. Two projects, in particular, required substantial leadership and coordination with 6 other research enterprises (i.e., *Consortium for Alternate Assessment Validity and Experimental Studies* (2006-09) and *Coordination, Consultation, and Evaluation Center for Implementing K-3 Behavior & Reading Intervention Models* (2002-06).

Stevens is the Associate Dean for Academic Affairs of the College of Education and the Director of the Center for Assessment, Statistics, and Evaluation (CASE) at the University of Oregon. He was formerly a Measurement Statistician at the Educational Testing Service (ETS) and a Project Director at the Psychological Corporation. He has extensive experience in assessment, measurement, and instrument validity. His expertise also lies in the application of statistical models like HLM and SEM to problems of educational policy and practice. He has worked for over 15 years in the design and development of state assessment and accountability models, has consulted with several states on growth models, and was the primary architect of the Oregon Growth Model Pilot.

Management and Institutional Resources

The proposed Center will be directed by Tindal with assistance from Co-PIs Ann Schulte, Stephen Elliott, and Joseph Stevens. Tindal has experience leading major multi-site research projects and has chaired a department for over a decade at the University of Oregon. Schulte has experience leading research projects and university programs/departments. Elliott has lead major university-wide research operations at the University of Wisconsin, at Vanderbilt University, and currently Arizona State University, and has been Project Director of six multi-site, multi-year grant projects, where a similar team structure has worked successfully. Finally, Stevens is an Associate Dean and Director of the Center for Assessment, Statistics, and Evaluation at the University of Oregon. All four PIs embrace a management by objectives approach that features frequent communication, progress monitoring, and objective accountability. Collectively, this 4-person leadership team has the scientific and organizational expertise to ensure the Center accomplishes its goals efficiently and effectively. More information is provided in the Personnel section about these and other people and how they will be functionally organized to accomplish the Center's goals. The goals are to: (1) conduct research that provides valid evidence about the natural developmental progress in achievement of students with disabilities and (2) develop and test various approaches for measuring the academic growth of students with disabilities for purposes of valid accountability decisions. To accomplish the Center's goals on a scale that yields meaningful and reliable results, we have assembled a team of researchers, state partners, and professional organizations committed to advance research-based practices for all students. Many of the resources that enable our team of behavioral scientists and educational leaders to accomplish their work will be provided by three institutions: the Learning Sciences Institute (LSI) at Arizona State University, the North Carolina Education Research Data Center (NCERDC), and the Behavioral Research and Teaching Center at the University of Oregon. The coordinating institution and fiscal agent responsible for the Center will be Oregon's BRT Center.

Management of Data, Analyses, and Reporting

The primary work of the Center revolves around the management, analysis, and interpretation of datasets concerning the achievement of SWDs. The scope and amount of data that we propose processing is substantial and requires the management of a team of measurement and statistics experts. As indicated in Figure 6, work assignments will be structured around the 5 planned studies, 11 datasets, and the expertise of our personnel.

Figure 6. Lead Researchers for 5 Planned Studies

Study	Data Sets	Lead Researchers	Years
Cornerstone Study	NC	Schulte • Stevens • Zvoch • McCaffrey	1, 2, 5
Multi-State Extension	AZ • OR • PA	Schulte • Stevens • Zvoch • McCaffrey	2 & 5
Interim Assessments	MAP • easyCBM	Elliott • Tindal • Thum • Levy	1 & 2
Multiple Measures	OTL • Interim • States	Elliott • Kurz • Gorin • Levy	2 – 4
Alternate Assessments	NC • AZ • OR • PA	Tindal • Zigmond • Gorin	3 – 5

The R&D Center database will include over 1,500,000 student cases from across each of 4 states collected over a period of at least 10-years (2005-2016). This comprehensive dataset will be stored in a Microsoft Access database on a secure server at the University of Oregon. It is capable of storing large amounts of data while making the data quickly accessible. Data can efficiently be imported or exported within Access between databases and spreadsheets. It will allow multiple people to work in the database, and will provide multiple options for updating, sorting, querying, and reporting. Access security features can be set to allow users limited or widespread usage rights and can be encrypted. The records are identifiable only by identification number, which is used to connect students across years, as well as to identify students within classrooms, schools, districts, and states. Members of the Oregon data management team have experience using this type of longitudinal data set. The Center Data team will work closely with the University of Oregon's and other institutions' IRBs to ensure participant protection and the integrity of the database during and 5 years after completion of the project.

Data Access and Publication Rights

A data access and publication policy has been drafted by the co-PIs. This document provides guidelines for all project Principal and co-investigators regarding (a) access to and use of data and (b) authorship rights and credits for public reports of research funded by the Center. These guidelines are intended to facilitate decision-making about data use and authorship credit. This policy is founded on the ethical principles and consensus professional practices set forth in the *Publication Manual of the American Psychological Association* (2010) and influenced by policies used to guide the publications of the multi-center, NICHD Study of Early Child Care (1998-2004) and the OSEP K-3 Reading and Behavior Intervention Project (2000-2006).

Major Institutional Resources

University of Oregon Special Education and Behavioral Research & Teaching (BRT). The UO College of Education had external funding expenditures of $33 million in 2008-2009. The COE ranks among the *top five* public and private graduate institutions of education in the nation and is *first* among all publics. The UO special education program is ranked *third* in the nation. The BRT supports educational research and professional development with a mixed computing environment of high-end workstations and servers. The computers have high processing speeds and access to large data storage systems. Network services are accessible via the LAN providing multiple file sharing, backup and structure data storage options. Secure servers are deployed in a production data center with multiple (and diverse) networking and power connections, onsite power backup including diesel generation, and redundant cooling.

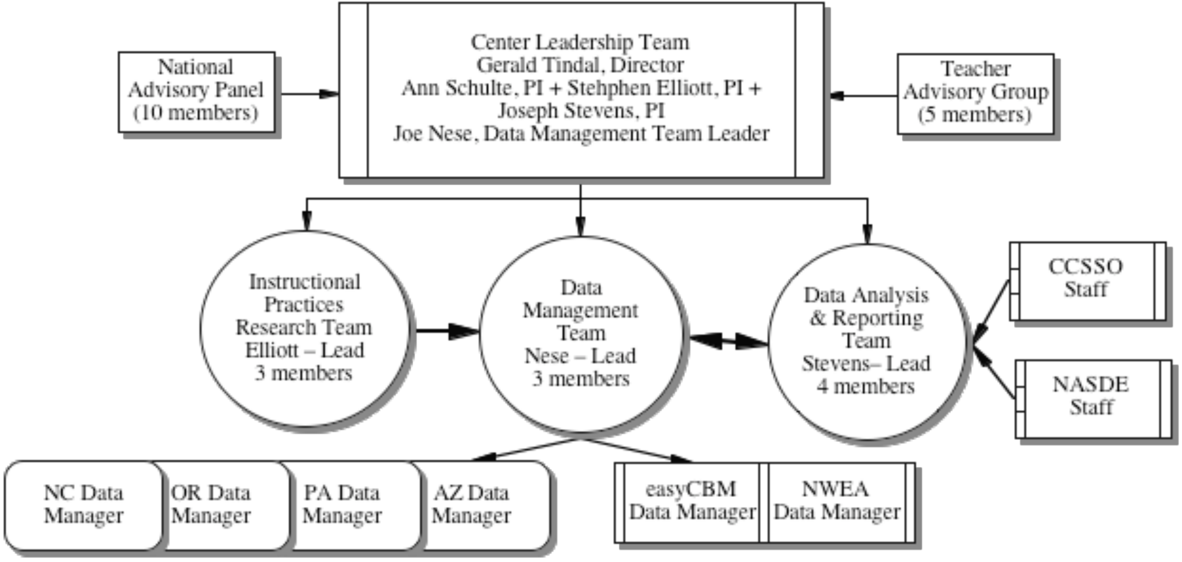
The North Carolina Education Research Data Center (NCERDC). This Data Center was established in 2000 at Duke University houses data on every district, school, teacher, and student in North Carolina public schools from 1995 to the present. One of the most valuable resources at NCERDC is its archive of large-scale assessment results for students attending NC public schools. Since the 1992-93, 3rd through 8th grade students enrolled in NC public schools (including charter schools) have been tested annually in reading and mathematics with the NC End of Grade Tests in Reading (EOG-R) and Mathematics (EOG-M). Within the NCERDC, student test results are available in a variety of datasets, such as those that include all children tested at each grade each year and those that provide expected and actual growth achieved for each student by grade and year using the NC growth model. Demographic data are included in datasets and information on student participation in exceptional children's programs is available by exceptionality in all years, as are the accommodations the student received while taking the EOG tests. Starting in 2001, the setting in which each student's exceptional children's services were delivered also is available. Starting in 2006, information about the special education service plan, exit date, and specific services provided is available. The NCERDC also has corresponding datasets for teachers, classrooms, and schools (although teacher and classroom are not linked).

Learning Sciences Institute (LSI) at Arizona State University. The LSI, directed by Stephen Elliott, at Arizona State University brings together interdisciplinary teams of faculty and students from across the university to address basic and applied learning sciences in ways that create new knowledge. Funded projects are supported through budget monitoring, meeting management and hosting, information dissemination, and technical services such as multimedia, computer support, and data management. LSI staff assists faculty with the development/maintenance of project web sites, and the dissemination of research and development findings via the LSI website, the university news service, and the national media. The Learning Sciences Institute has large meeting rooms and multimedia capabilities that will allow the leadership team to hold both on-site conferences and teleconferences with assessment professionals nationwide.

Personnel

A highly organized and productive network of research scientists and state assessment leaders together will conduct the proposed research and leadership activities. The Center Director will be Gerald Tindal of the University of Oregon. He will be joined by his three co- principal investigators: Ann Schulte, Stephen Elliott, and Joseph Stevens. These experienced large-scale assessment researchers will collaborate with measurement and statistics experts – Dan McCaffrey, Keith Zvoch, Joanna Gorin, Roy Levy, and special education co-investigators Naomi Zigmond and Alexander Kurz, all of whom they have collaborated with on previous externally funded research projects. This team of researchers will work with assessment and special education leaders in state departments of education and lead psychometricians at NWEA. Key personnel will be organized into four functional teams: (a) Center Leadership Team, (b) Data Management Team, (c) Data Analysis & Reporting Team, and (d) Instructional Practices Team. The Leadership Team is comprised of the Project Director, the Co-PIs, and the leaders of three task-specific teams (See Figure 7). The Center's Leadership Team will benefit from the intellectual guidance and sage wisdom of a National Advisory Panel comprised of leading measurement scientists (Bennett, Cizek, Linn, & Plake), special populations researchers (Abedi, Cook, Kame'enui, Vaughn, & Ysseldyke), and an accountability and policy analyst (Erpenbach).

Figure 7. R&D Center's Personnel and Management Structure


National Advisory Panel (10 members)
Center Leadership Team
Gerald Tindal, Director
Ann Schulte, PI + Stehphen Elliott, PI +
Joseph Stevens, PI
Joe Nese, Data Management Team Leader
Teacher Advisory Group (5 members)
Instructional Practices Research Team Elliott – Lead 3 members
Data Management Team Nese – Lead 3 members
Data Analysis & Reporting Team Stevens– Lead 4 members
CCSSO Staff
NASDE Staff
NC Data Manager
OR Data Manager
PA Data Manager
AZ Data Manager
easyCBM Data Manager
NWEA Data Manager

The Center's efforts to have influence and input beyond its partner states will be enhanced by the support of the Council of Chief State School Officers (CCSSO) and the National Association of State Directors of Special Education (NASDSE). A brief description of the Center's Leadership team and key Co-Investigators follows. The brief vita and effort documentation of all key personnel is provided in this application.

Leadership Team

Gerald Tindal, PhD, Project Director & Co-PI, is a Castle-McIntosh-Knight Professor and the Department Head of Educational Methodology, Policy, and Leadership as well as the Director of Behavioral Research and Teaching (BRT). His focus will be replicating the growth modeling research in OR and working with both supplemental strands (alternate and interim assessments) along with extension of this information to professional development. Tindal will contribute 40% each academic year and 75% for summer months in Years 1-5.

Ann Schulte, PhD, Co-PI, is a Professor of Psychology at North Carolina State University with expertise in large-scale assessment and measurement. She has conducted research on achievement growth for special education students in North Carolina over the past decade and worked closely with investigators from the NCDE and NCERDC. During Years 1 and 2, Schulte will contribute 75% effort during the academic years and 2.5 months each summer. In Years 3-5, she will contribute 50% effort during the academic year and 1.5 months each summer.

Stephen N. Elliott, PhD, Co-PI, is a Professor of Education and the Mickelson Foundation Professor in Learning Sciences at Arizona State University. He also is the Director of the Learning Sciences Institute at ASU. He has collaborated with five states in the design and validation of alternate assessments for students with significant disabilities and currently is the Senior PI on a USDE project that is helping both AZ and IN develop an AA-MAS. Elliott will contribute 40% effort during each AY and 2.25 summer months in Year 1 and 5 and 2.0 summer months in Years 2-4.

Joseph Stevens, PhD, Lead Statistician and Co-PI is a Professor in Educational Methodology, Policy, and Leadership with expertise in multi-level modeling and advanced statistical analyses of large data sets. He has conducted research on growth modeling for over 15 years. He consults regularly with state departments of education. He will contribute 25% effort during the academic year and 50% during the summer for each of the five years of the grant.

Data Analysis Team Members

Joseph Stevens, PhD, University of Oregon will serve as the lead statistician and will work with two other members from the University of Oregon: *Keith Zvoch, PhD* (funded at .10 FTE during the academic year) and *Gina Biancarosa, EdD* (funded .25 during the academic year and.20 during the summer). They will work with *Daniel McCaffrey, PhD,* RAND (Pittsburgh), *Joanna Gorin, PhD*, Arizona State University*, Roy Levy, PhD,* Arizona State University*,* and *Y.M. Thum, PhD,* NWEA.

Data Management Team Members

Joe Nese, PhD., Research Associate, University of Oregon will serve as the lead (1.0 FTE for 12 months) and work wit*h Peter Beddow, PhD*, Research Assistant Professor, Learning Sciences Institute at ASU (.60 FTE for 12 months). In addition, a *Post-Doctoral Data Specialist* North Carolina State University *(funded at 1.0 FTE for 12 months)*, will participate as a member of the Data Management Team, assisting with, recruiting schools, collecting, structuring, and analyzing data during Years 2-4. Finally, a *Data Specialist*, NCERDC at Duke University will be funded to facilitate access and efficient use of the NC dataset. *Denise Swanson,* University of Oregon (.25 FTE for 12 months) will assist all members of the Data Management team in data quality assurance.

Instructional Practices Research Team Members

Stephen Elliott, PhD, at ASU, Alexander Kurz, MS, Department of Special Education, Vanderbilt University, has experience in conducting research with MyiLOGS, *Naomi Zigmond*, Distinguished Professor of Special Education at the University of Pittsburgh, and *Andrew Roach*, Assistant Professor at Georgia State University.

State Department of Education Data Analysis and Reporting Members

Charles Bruen, *PhD*, is the Arizona Director of Data Analysis, Budget and Technology for the Assessment Section and performs the psychometric analyses of the regular assessment along with the alternate assessments. *Louis M. Fabrizio, PhD,* is the Director of Accountability Policy & Communications for the North Carolina Department of Public Instruction. Fabrizio currently serves on the National Assessment Governing Board. *Doug Kosty,* is the Assistant Superintendent of the Office of Assessment and Information Technology at the Oregon Department of Education. Kosty was a key developer of the Oregon Database Initiative Project (DBI). *Kristen Lewald, EdD*, is statewide Project Director for the Pennsylvania Value-Added Assessment System (PVAAS) under the leadership of the PA Department of Education (PDE). She is responsible for communication with school districts across PA and supports growth model implementation for AYP, and supervision of PVAAS staff. Lewald also coordinates PD in PA.

National Advisory Panel & Teacher Advisory Group Members

A total of 10 individuals have agreed to serve on our National Advisory Board; an additional member will be selected in consultation with IES staff upon funding. See Budget Narrative for a full list of individuals and their institutions: Abedi, Bennett, Cizek, Cook, Erpenbach, Kame'enui, Linn, Plake, Vaughn, and Ysseldyke. Our Teacher Advisory Group will consist of 7 members from schools participating in the Multiple Measures Study.

Partnering Test Company Psychometricians & Co-Investigators

NWEA- *Steven Wise*, *Ph.D.* is the Vice President of Research and Development for the Northwest Evaluation Association (NWEA). Before joining NWEA in 2008, he spent 10 years as a Professor of Psychology at James Madison University in Assessment and Measurement. Prior to his time at JMU, Dr. Wise taught for 15 years in the Department of Educational Psychology at the University of Nebraska-Lincoln. Dr. Wise has published extensively on applied measurement and testing. *Yeow Meng Thum*, Ph.D. is a Senior Research Fellow at NWEA. He received his Ph.D. from the University of Chicago in Quantitative Psychology and has published in leading journals on issues in statistics, educational assessment, value added and longitudinal modeling, and accountability.

EasyCBM is operated from a technology transfer agreement with the UO arranged by Dr. Tindal. It is supported by his research with grants primarily from the Institute of Educational Sciences. The system is supported with three programmers, one whom will be funded by the Center (*Aaron Glasgow, a Programmer Analyst* working at .20 FTE for 12 months). In addition, *Paul Yovanoff, PhD*, Associate Professor and Research Associate, will be funded .25 FTE during the academic year and .20 in the summer, helping analyze the Interim Assessments with state tests (including Alternate Assessments).

Personnel Summary: Experienced, Competent, Dynamic, and Ready to Lead

In summary, the proposed Center's Leadership Team is an experienced and highly competent group of educational scientists with deep knowledge of large-scale assessments, special education, and performance measurement using growth modeling methods. Their intellectual curiosity about student achievement has driven their empirical investigations for more than two decades. The challenge and opportunities to continue this assessment research on a national scale with the IES is exciting. The PIs are proven national leaders. As a result of their established competence and leadership, they have attracted and assembled a team of measurement and assessment experts who have already, in many cases, collaborated to complete multi-site, data-intensive projects. These team members all share the vision for the Center's work operationalized by our goals and proposed studies. In cooperation with IES, this team is ready to work to provide answers to longstanding questions about achievement growth of students with disabilities and to lead in the dissemination and appropriate use of evidence-based research and growth modeling methods that can enhance special education accountability.