# An Accountability Framework For Providers Of Supplemental Education Services Under The *No Child Left Behind Act of 2001*

Original Draft, August 2002
Updated November, 2004

***Accountability for Results, Not Regulation***

**Theodor Rebarber & Richard W. Cross**

**Education Leaders Council**
**& AccountabilityWorks**

# Executive Summary

- The provisions of the No Child Left Behind Act of 2001 (NCLB) authorizing supplemental education services represent a clear break from the past. The new program is *parent-driven.* It is also focused on *accountability for results*, not on compliance with regulations for process or inputs.

- Consistent with the spirit of the Elementary and Secondary Education Act (ESEA), the new supplemental services provisions focus on those schools and students most in need—low-income students in schools that have failed to make Adequate Yearly Progress (AYP) for three or more years. NCLB provides the parents of such students an opportunity to seek out quality educational opportunities for their children.

- Accountability is a critical theme in the supplemental services program, as it is throughout NCLB. Three actors play major roles in accountability under this program: states, school districts, and parents.

- States must evaluate the contribution of each approved provider of supplemental services to increasing student academic achievement. Providers that fail to make a positive contribution to student achievement for two consecutive years are to be barred from providing such services within the state.

- Challenges to carrying out the state role include properly assessing achievement increases for students who are likely to be very low performing as well as disentangling the contribution of providers from the contribution of schools and teachers. Various options exist for solving these challenges. We recommend a gain model using assessments that are sensitive to the achievement of low-performing students, that use methods for isolating the providers' contribution from that of the schools.

- The state evaluation can provide a fairly precise measure of the contribution of large providers and a rough sense of the contribution of mid-size providers, but it is unlikely to measure accurately the contribution of small providers.

- LEAs are required to work with parents and providers in establishing individual student learning goals and in determining whether the provider succeeded in accomplishing these goals. Providers must offer regular progress reports during the course of services. For non-performance, an individual student's supplemental services agreement with a provider may be terminated or not renewed.

- Challenges for LEAs include providing user-friendly support to parents, assessing the achievement of students who are low-performing, and evaluating reliably the success of individual students. For students whose learning goals are appropriately captured in the state Proficient or Basic achievement standards, we recommend LEAs use these in an evaluation model focusing on absolute performance. For students with different, or narrower, learning goals for supplemental services, LEAs should identify the most

appropriate assessment instrument and benchmark, while considering all available evidence of student learning. Gain models alone are too unreliable at one-year increments for individual students.

- States and LEAs have the responsibility to assist parents in implementing parental accountability. Research has demonstrated that educational providers improve achievement when parents are able to choose among multiple viable options.

- States can provide information on all approved providers in a format accessible to parents, such as on the state website. In addition to descriptive information, states should include performance data on providers, such as ranking all providers by the size of their achievement gains, providing the frequency of parent withdrawals for each provider, and reporting the results parent satisfaction surveys.

- LEAs have even greater responsibilities, including assisting parents directly in interacting with providers. Unless data are unambiguous, LEAs should generally defer to parents on decisions regarding supplemental services for their child, but they should assist parents in making wise decisions.

- LEA contracts with supplemental service providers should be designed to allow parent discretion in switching between providers throughout the school year. Parents should be free to drop an ineffective provider in favor of one that is more effective, though LEAs may want to set some limits on the number of changes in any one year.

# Contents

## Introduction

This document is designed to assist state and district policymakers in designing effective accountability systems for supplemental education services under the *No Child Left Behind Act of 2001* (NCLB). The supplemental services provisions of NCLB (Title I, Section 1116(e)) represent a clear break with past practice on the best way to help recipients of federal Title 1 funds in persistently failing schools. Unlike other federal K-12 education program, the operation of this program is designed to be *parent-driven*. Under the new legislation, eligible parents are empowered to choose among providers of tutoring or other supplemental academic services for their children and districts are obligated to contract with the providers selected by the parents. The services, which may be purchased with federal Title 1 funds from either public or private providers, are to be offered outside of regular school hours.

Consistent with the strong theme of accountability that runs through the entire legislation, NCLB also insists on accountability for supplemental services. Supplemental education providers, in order to continue to participate in this program, must demonstrate results—in particular, academic achievement. Implementing these accountability provisions involves multiple actors at the state, district and even parent levels, each with his or her role to play. Yet, aligning the federal and district accountability requirements of this program with each other and with the central role of parents requires careful calibration. So do the many challenging technical hurdles facing states and districts as they carry out these responsibilities.

To assist, first this document identifies some of the key legislative goals that embody the spirit, as well as the letter, of the statute. Following that, this document does *not* offer a detailed blueprint or a particular approach to implementing accountability for this program. Instead, we offer a framework of key principles, analyses, options and some recommendations for designing accountability at each level. Also included are some illustrative scenarios of the real-world implications of different models and assumptions, including, unfortunately, some of their limitations.

Despite the challenges, the supplemental education services program offers a unique opportunity for creative state and local policymakers to harness apparently contradictory impulses—private sector entrepreneurialism (is this a word?) and government accountability—toward the common goal of improved student achievement. There is no one right way to do it—or, at least, it is far too early to know what it might be at this time. But this framework is designed to provide you with what we hope are some useful tools to help you start on the right track.

**Legislative Goals**

Consistent with the broad statement of purposes for Title I of the *No Child Left Behind Act of 2001* (the goals of the supplemental services provisions are based on the principle accountability for results, not inputs or processes.

Four of the legislative purposes of Title I (Section 1001) are especially relevant:

> (3) closing the achievement gap between high- and low-performing children, especially the achievement gaps between minority and nonminority students, and between disadvantaged children and their more advantaged peers;
> (4) …identifying and turning around low-performing schools…while providing alternatives to students in such schools to enable the students to receive a high-quality education;
> (7) providing greater decision-making authority and flexibility to schools and teachers in exchange for greater responsibility for student performance;
> (12) affording parents substantial and meaningful opportunities to participate in the education of their children.

Raising achievement, especially among low-performing and minority students is critical. Offering alternatives to students in low-performing schools, focusing on results, and a affording parents a meaningful role in their children's education are all animating principles.

**Eligibility for Supplemental Education Services**

The following schools and students are eligible to participate:

- Schools that are in their second, or later, year of "school improvement" under NCLB (i.e., schools that have failed to make "adequate yearly progress" (AYP) for three years).
- Students within those schools whose families are classified as "low-income".

We can expect that the great majority of participating students will be low-performing and functioning significantly below grade level.

**Input/Process requirements**

While the overall focus of accountability should be on results, some input/process requirements are legitimate, either because they comprise fundamental expectations of fairness, safety or fiscal prudence, or because they enable state, district, or parental accountability. Such requirements may include:

- Follow basic health, safety, civil rights statutes and regulations
- Provide parents with regular updates on student progress against instructional objectives (at minimum monthly—though could be triggered by instructional

milestones—e.g., every 5 lessons mastered—or by calendar—e.g., monthly or weekly), as specified in the agreement between providers and parents.
- Enrollment and attendance reporting requirements
- Testing requirements
- Requirements for return of (unspent) funds in cases where a parent leaves before the year/course is completed
- Instruction should be neutral and free from bias or religion

However, states and districts should avoid requirements that prescribe the organization, management or operations of supplemental services providers. Input/process/provider requirements to *avoid* at either state or LEA levels include such things as:
- organizational/governance requirements
- curriculum specifications, but the provider should offer assurances that the program is designed to improve students' performances relative to state academic content and achievement standards
- restrictions on instructional methods and services to be provided
- mandating a specific model for reporting student progress to parents and the LEA (though guidelines, such as at least one report per month, are appropriate)
- certification requirements for instructors
- maximum instructional group size
- restrictions on expenditure of funds
- excessive restrictions on minimum number of hours (some providers will be more efficient than others, though some minimum may be appropriate)
- applying public school facility code requirements to off-site providers

**NCLB Accountability Requirements for States:**

> 1116(e)4(D). A state education agency shall…develop, implement, and publicly report on standards and techniques for monitoring the quality and effectiveness of the services offered by approved providers…and for withdrawing approval from providers that fail, for 2 consecutive years, to contribute to increasing the academic proficiency of students served under this subsection…

Determining whether providers contribute to student achievement requires the state to distinguish between the impact of the supplemental services providers on academic achievement and the impact of the normal school day/teachers on achievement. Absent this distinction, the provider effect could be artificially inflated, or artificially deflated, by the school or teacher effect.

Evaluating the "increase" in academic achievement suggests the use of a gain, or "value-added," evaluation model. Alternatively, an evaluation could seek to identify the provider's contribution to increasing the percent of students reaching the "proficient" level of achievement. However, we recommend a gain model as the preferred approach. We do *not* do so because we perceive a gain model to be inherently superior to an absolute model. Indeed, in the section of this document addressing LEA accountability

we encourage absolute models for accountability where such assessments and achievement levels are valid measures of individual student learning goals.

With respect to state-wide evaluations of provider effectiveness, though, a gain model is likely to be superior for two reasons:

- First the design of the supplemental services program (i.e., low income students in persistently failing schools) ensures that most participating students will enter very low in academic achievement. Since the proficient level is intended to denote a high level of academic mastery, it is likely to miss a substantial portion of the learning gains of students in this program.
- Second, even if the lower level of achievement (i.e., Basic) were to be incorporated into the absolute model, many changes in student performance would go unnoticed unless students happened to cross the Basic or Proficient levels. Increases or declines that left students within a category would not be incorporated into the results. Given the challenge in discerning which providers are contributing to increases in student achievement (described on p.10-11), as well as potential legal challenges from disqualified providers, the most valid model takes advantage takes advantage of all available data in order to make the most accurate evaluation.

The remainder of this section on the state accountability model, discusses the key features of a gain model, including for states that currently assess grade-by-grade as well as those that currently do not.

**Identifying The Assessment For The State Evaluation**

The key issues involved in identifying the appropriate assessment(s) for the state evaluation are: a) is the assessment administered to comparable students who are *not* receiving supplemental services? b) is the assessment well-suited to measure the gains of students receiving supplemental services?

- One of the key issues in identifying the assessment instrument for the state evaluation is the extent of its use throughout the state (or in districts, for states where there is no state assessment). Specifically, it is helpful if students who receive supplemental services and students who do not receive such services both take the same test.
    - The reason this is important is that it is necessary for distinguishing between the impact of the provider and the impact of the school/teacher. In states with a state-wide assessment, the state assessment would best satisfy this criterion. Where there is no state assessment, this criterion could also be addressed through district-wide assessments, such as norm-referenced tests (in such cases, the gains for individual providers would have to be aggregated and compared across the different district assessments; the methodology for this analysis is commonly used to compare gains across multiple studies in research meta-analyses). If a test is not used for students other than those served by the provider, it will be

more challenging to identify the impact of the provider unless the provider has undertaken studies to demonstrate the correlation between its test and the state test or district tests. (See below regarding the state evaluation design.) If the state wishes to permit providers to select the assessment on which they should be evaluated (presumably, from a state list of approved valid and reliable assessments), the results from such "criterion validity" studies would permit the necessary comparisons.

- The other major set of issues in selecting the test is its ability to assess the learning gains of students served by the provider. Three issues are relevant: a) whether the test is designed for measuring gains, rather than simply absolute performance, b) whether the test is administered grade-by-grade, and c) whether the test is designed to accurately assess the skills of students functioning substantially below grade level.
    - To measure gains, it is very helpful if the test is designed around a *vertical scale* that covers multiple grade levels. The results of a test designed this way may be translated into numbers on a common scale, even for students at different grade levels. For example, on a test with a vertical scale ranging between 200 and 500, the typical performance for a 4th grader may be 310 and the typical performance for a 5th grader may be 342. Using such a scale, the difference between this student's performance in 4th grade and his performance in 5th grade, 32 points, is the gain. A few state standards-based tests are designed with such a vertical scale, though many are not. All major norm-referenced tests are designed with a vertical scale.[i] (For tests lacking a vertical scale, an alternative method of determining the "value-added" by calculates the expected performance on the higher grade test for students performing at a given level on the lower grade test. Then it determines whether students receiving supplemental services exceeded the expected performance. This alternative approach is possible, but less desirable.[ii])
    - The second issue, whether the test is capable of assessing gains for very low-performing students, is important because most students served under this program are likely to be functioning substantially below grade level.
        - State standards-based tests are usually designed to measure student learning of grade-level skills, including whether they are meeting a "proficient" achievement standard with respect to such skills. Many standards-based (or "criterion-referenced" tests) are not designed to assess the progress of students substantially above or below grade level, lacking an adequate number of test questions designed to tap these skills. If a test designed to assess students at or near grade level were used to assess the gains of students far below grade level, substantial gains could well be missed. States with vertically-scaled tests can utilize out-of-level testing.[iii]
        - Norm-referenced tests (e.g., the SAT-10, the ITBS) are typically designed to be sensitive to student skill levels within 1.5 academic years of the targeted grade, with some ability to assess skills beyond those limits. For example, a student operating at a reading

grade level of 4.6 in April of 5th grade—or 5.8—may be considered 1.2 years below grade level, well within the accuracy range of most NRTs.

- Another alternative is a computer adaptive norm-referenced test. A computer adaptive test (such as the STAR-Reading test) is able to assess students functioning at a wide range of grade levels by adapting the test questions based on how well the student performs during the assessment. A student who performs poorly on grade level questions at the beginning of the test would be assigned below grade level questions later in the test; similarly, a student who aces grade level questions at the beginning of the test would be assigned more above-grade questions. In either case, the test will provide a substantial number of questions at or near the student's skill level, resulting in an accurate assessment despite the fact that the student's skill level is far from grade level.[iv]

**State Evaluation Design and Analysis of Results**

As in the state's main accountability system for all schools and students, LEA's play an integral role in implementation of the state accountability model for supplemental services, including test administration and other forms of data gathering. Given their importance, LEAs should receive guidance from the state outlining their responsibilities.

Initial state guidelines for LEA implementation should address at least the following:

- Ensuring that districts provide a neutral and level playing field for all providers—including for-profit providers, non-profit providers and their own public schools—in evaluation.
- Recording with the assessment responses for each student: the provider, the student's LEA, school and teacher. These are necessary to allow for identifying the individual provider contribution and separating it from the school or teacher contribution.
- Requiring that districts ensure the integrity of state evaluation protocols, including: the administration of any additional assessments required of providers or LEAs; proctors for fall, spring, and, if possible winter, test administrations. Test administration should be overseen by district staff, or district-hired staff, rather than provider staff.

To determine whether the students served by a particular provider achieved a gain, the following requirements will need to be filled:

- The test scores of the same cohort of students, before and after the supplemental instruction, are matched. Only students present for both pre and posttests are included (but great effort should be made to ensure that all or nearly all students enrolled in the program and in school are included). A database containing all participating students, including individual student identifiers, is very helpful in properly matching and tracking student test scores. States that do not have such

databases at the state level may need to establish guidelines for developing such databases suitable for gain analyses at the district level and reporting the results to the state.

- A statistical model is used to calculate the gains and determine which gains are significant (apparent gains may not be statistically suitable for making decisions due to measurement and sampling limitations). Due to the nature of hierarchical school data (e.g., teacher gains nested within school gains), it is recommended that newer, more sophisticated statistical models be used for these calculations rather than the standard ANOVA models. Statistical models designed for this type of analysis include William Sander's version of the Mixed Model and Anthony Bryks' HLM model.[v]

The state evaluation must be designed to determine three things.

- First, it must determine if there was an increase, or gain, in student academic proficiency for students served by a particular provider. Since there is almost always *some* increase in student knowledge and skills (as measured on a test's vertical scale), the type of achievement gain we are seeking is a gain *above and beyond* the typical student gain. Since the goal of the *No Child Left Behind Act* is for all students to achieve proficiency in challenging subject matter—a high bar—it is reasonable to expect students receiving supplemental services to make gains above and beyond "typical" learning gains.
    - On state standards-based assessments, the state will need to determine what constitutes a typical gain in order to establish the higher "reference gain" that will denote success for the purposes of this evaluation. Equally defensible approaches include requiring that the reference gain exceed slightly the state average grade level gain for that year or, more generously, the average ("norm") grade level gain in the original administration of the state test. While a reference gain based on exceeding the annual state gain is likely to be more demanding (since most states experience achievement increases after a new test is introduced), it is also likely to lead to the exclusion of more providers and, therefore, fewer choices for parents. Such a decision may be viewed as a matter of state policy. On the other hand, it may be difficult to justify an even higher reference gain—one that exceeds both the original norm gain and the state average gain by more than a marginal amount—given the state role identified in the federal statute. The NCLB describes the state role in terms of ensuring that all providers contribute to some level of gain—not ensuring that all providers achieve a large gain. Beyond a state-ensured "floor" of adequacy, the statute seems to view parents as the ones to make decisions regarding which provider is *best* for their child. As we noted previously, there is reason to think that such parent accountability can be effective. We discuss ways in which states, and districts, can assist parents in wielding this power effectively later in this document.
    - On the widely used norm-referenced tests, which serve as an easily transferable illustration, the reference gain may be described in terms of the Normal Curve Equivalent (NCE) scale, where 50 is the national norm,

99 is the level achieved by the top 1% of students, 1 is the level achieved by the bottom 1% of students, and each unit in between is of equal size. On this scale, a group of students performing at the 40th NCE at the beginning of a school year—before receiving supplemental services—and then performing at the 40th NCE at the end of the school year—after receiving supplemental services—would have made only the typical gain (based on the original test norm sample) and 0 gain beyond that. On the other hand, students moving from the 40th NCE to the 41st NCE in one year would have achieved measurable, though perhaps not impressive, improvement beyond the norm gain. Therefore, a state may define a 1 NCE gain as demonstrating the minimally acceptable gain and evaluate providers against that. Alternatively, as with a state-developed test, the state may define the reference gain as exceeding the annual average state gain. If the state achieved a state-wide average gain of 2 NCE points, for example, the reference gain might be 3 NCE points.

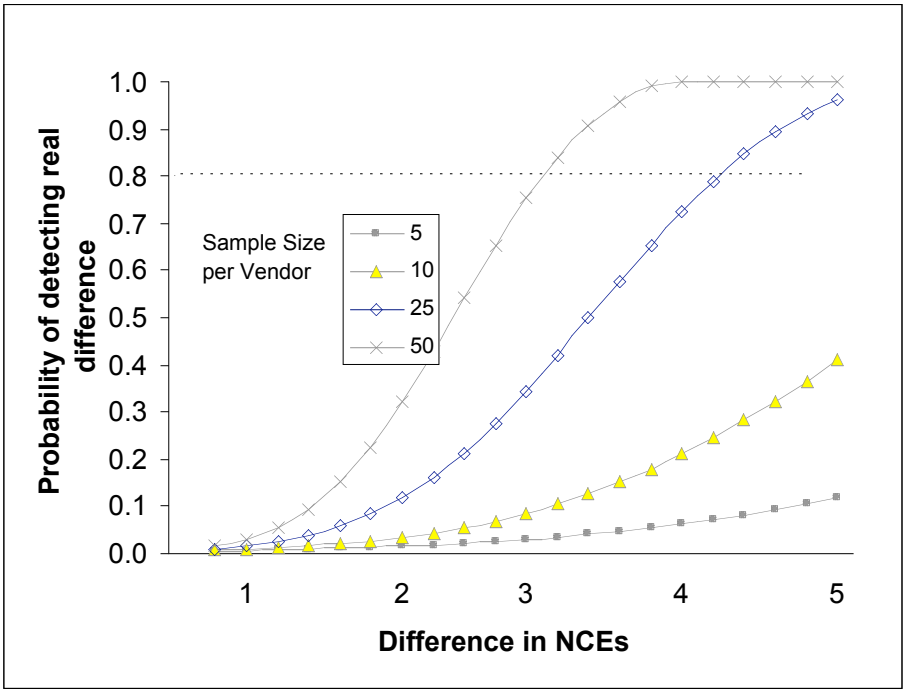
- Second, the state evaluation must distinguish the contribution of the provider from the contribution of the regular day school/teacher. In other words, did the provider's supplemental instruction result in a gain that met or exceeded the reference gain after separating out the contribution provided by the instruction of the school/teachers? For example, an overall student gain of 0 NCE points could be due to a school/teacher contribution of –1 NCE and a provider contribution of +1 NCE, resulting in the provider meeting a state reference gain set at +1 NCE. Or an overall gain of +2 NCE could be due to a school/teacher contribution of +3 NCE and a provider contribution of –1 NCE, resulting in the provider failing to meet the state reference gain. In order to isolate the provider impact on the gain, we must identify the school impact and subtract it from the overall gain. This can be accomplished in multiple ways, with some approaches more appropriate to a particular state context than others. One way to determine the school impact is by comparing the gains of students being served by a provider with the gains of similar (eligible) students in the same schools who are *not* receiving supplemental services (perhaps because they chose not to participate or because there were insufficient funds to serve all eligible students). In simple terms, the gains of the students in the same schools not receiving supplemental services would be subtracted from the gains of students who are receiving supplemental services from a provider, thus isolating the contribution of the provider. This presumes that state guidelines for test administration, discussed earlier, require that student assessment records identify for each student the provider (if any), the school, teacher, and income status/eligibility for supplemental services. The challenge is to ensure that this comparison group is indeed comparable to the students served by the vendor, with no systemic differences.
- Third, the state evaluation must determine if providers failed to meet the state reference gain for two consecutive years. NCLB indicates that only those providers failing to contribute to increases in achievement for two consecutive years are to be removed from the list of approved providers.[vi]

Apart from the design of the evaluation, states also needs to take into account the inherent challenge of identifying failing providers that serve relatively few students. (This challenge is impacted by other factors, including the total number of providers and the distribution of students among providers.) We offer two graphs on pp. 13-14 that illustrate in simple terms the extent to which the state evaluation will be able to identify providers that are failing for two consecutive years. The horizontal line (at 0.80) in each graph intersects the curved lines representing different-sized providers at the point indicating how far below the state reference gain the providers will need to perform before the model picks it up. As the graphs illustrate, the evaluation will likely be useful in eliminating just about all non-performing large providers (those serving 300+ students each year), and it may be useful in eliminating the worst among mid-size providers (those serving between 25 and 250 students). It will not, however, be useful in eliminating weak, or even terrible, small providers (those serving 10 or fewer students).

The first graph describes a scenario with 25 providers serving a total student population ranging between 75 and 1,250. The graph indicates, for example, that we could only be confident that a provider serving 5 students (assuming 75 students across all vendors state-wide) had failed to meet the state-established reference gain if the provider contribution to the gain fell short by approximately 15 or more NCE points. For a reference gain of +1 NCE, we could only identify such providers if they contributed well below –14 NCE points or less to the overall gain; clearly, a statistical analysis is not going to be a great help in evaluating such vendors. On the other hand, for a provider serving 50 students (assuming 1,250 students each year across all vendors state-wide) we could be confident that the provider had failed two years in a row to meet the reference target if the provider contribution fell short by approximately 3 NCE points; at this size, the analysis would identify providers that contributed a –2 NCE gain or less if the reference gain were +1 NCE point. The analysis may, therefore, be useful in eliminating the worst providers at this size, but will still miss many weak ones (those achieving a contribution of 0 NCE gain or –1 NCE gain).[vii]

**Graph 1: Small Providers, Small Total Population**

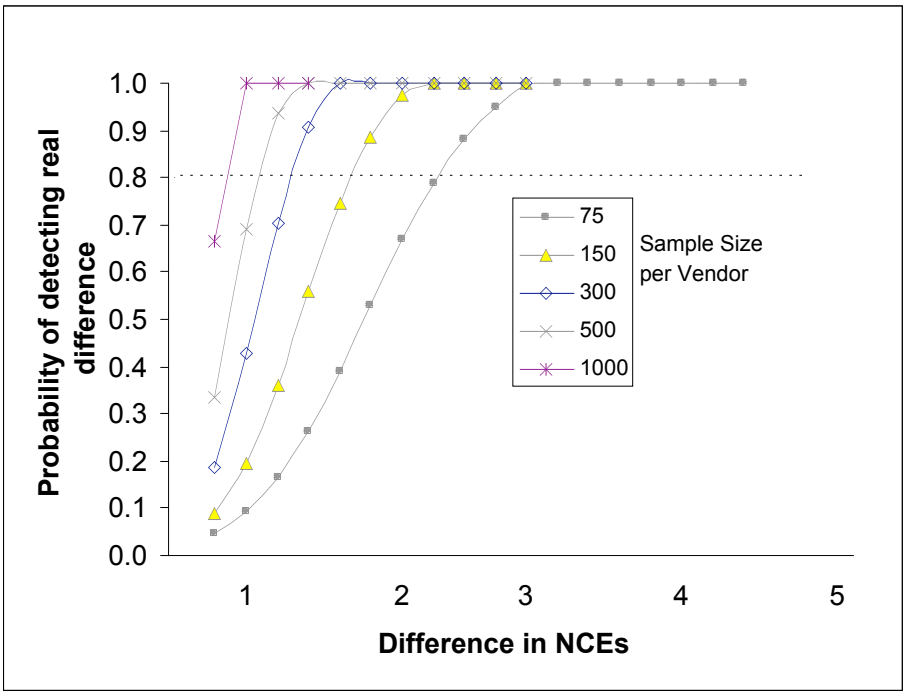
Probability of detecting a real difference from a minimum gain (or another reference gain set by a state), by NCE difference and sample size. Student sample size per provider per year for 25 small-scale providers.




(Approximate difference below the reference gain necessary to identify failure for different sized vendors: 50 students, 2.9 NCE points below reference gain; 25 students, 3.9 NCE points below reference gain; 10 students, more than 7 NCE points below reference gain; 5 students, more than 15 NCE points below reference gain.)

**Graph 2: Large Providers, Large Total Population**

Probability of detecting a real difference from a minimum gain (or another reference gain set by the state), by NCE difference and sample size. Student sample size per provider per year for 10 large-scale providers.




(Approximate difference below the reference gain necessary to identify failure for different sized vendors: 1000 students, 0.7 NCE points below reference gain; 500 students, 1.1 NCE points below reference gain; 300 students, 1.2 NCE points below reference gain; 150 students, 1.7 NCE points below reference gain; 75 students, 2.5 NCE points below reference gain.)

(SEE ENDNOTES FOR TECHNICAL SPECIFICATIONS)

**NCLB Accountability Requirements for LEAs**

1116(e)3(A). In the case of the selection of the selection of an approved provider by a parent, the local educational agency shall enter into an agreement with such provider. Such agreement shall…require the local educational agency to develop, in consultation with parents (and the provider chosen by the parents)—

- a statement of specific achievement goals for the student,
- how the student's progress will be measured, and
- a timetable for improving achievement that, in the case of a student with disabilities, is consistent with the student's Individualized Education Plan (IEP)

In addition, the agreement must—

- Describe how the student's parents and the student's teacher or teachers will be regularly informed of the student's progress
- Provide for the termination of such agreement if the provider is unable to meet such goals and timetables

The statute *requires* the LEA to enter into a contract with a state approved supplemental services provider selected by the parent of an eligible student. In effect, the LEA is acting as the parent's agent contracting for the services of the instructional provider. This arrangement highlights a key element of the relationship envisioned by the statute between the parent and LEA under this program. In addition to acting as agent, however, the statute also appears to envision a role for the LEA as expert advisor to parents, even their guarantor in obtaining high quality supplemental educational services. Whereas the state evaluation focuses on a provider's general effectiveness across all students it serves, the LEA evaluation focuses on provider success with each individual student receiving supplemental services. In the LEA evaluation of providers, success is determined on the basis of meeting agreed-upon achievement goals; such goals may be expressed either as increases or as absolute targets. There appears to be broad discretion in the design of these goals.

**Student Achievement Goals and Timetable**

The LEA has final authority for the development of the specific achievement goals and timetable for accomplishing the goals, an area where the LEA is presumed to possess professional expertise. Nevertheless, the LEA is required to develop these goals and the timetable for achieving them in consultation with the parent as well as the supplemental services provider. LEAs may wish to encourage (or require) all providers to propose learning goals that incorporate appropriate benchmarks on a valid and reliable assessment. Beyond such general guidelines, however, we expect (and recommend) that LEAs typically exercise this authority in the form of a final review and approval regarding goals and a timetable drafted by the provider with the input and agreement of the parent. This sequence will allow the provider to perform an evaluation of the skills of the student and develop a recommended course of instruction, with an accompanying timeline, based on the provider's own instructional model. The LEA should review this recommendation and, to the extent possible, make any necessary changes through a consensus-based process involving the parent and the provider. Where possible, the LEA should avoid alterations that simply reflect an differing instructional viewpoint or sequence, recommending (or insisting) on only those changes essential to achieving appropriate end-of –year learning goals for the student.

**Measurement of Student Achievement Goals**

Some of the same considerations apply to measurement. Measurement may usefully be divided between a valid and reliable end-of-year assessment of the learning goals and

more frequent curriculum-based assessments that are tightly aligned with instruction throughout the year. Where the state accountability assessment is a valid measure of a student's learning goals for supplemental instruction, it makes sense for the LEA to use results from the state assessment in evaluating provider success with individual students. In cases where the state assessment may not be a valid measure of the learning goals—such as where the supplemental instruction will focus on just a subset of the grade-level skills, or in cases where the student is functioning far below grade level—a more valid assessment should be selected if it is available (this assessment too should meet high standards of reliability). For example, if an evaluation of the skill deficits of a very low reader leads to a supplemental instructional regimen focused on oral fluency, the state accountability assessment may not be the most valid measure of whether the student is meeting goals for oral reading speed and accuracy (although one would expect increases in fluency to have a positive impact on the student's performance on the state reading assessment).[viii]

**Regular Reporting on Progress**

The LEA's responsibility to ensure that supplemental services providers "regularly inform" parents and teachers of each child's progress is a more important requirement than it may appear. A system-wide minimum reporting requirement for all providers, with waivers for unusual cases, may be helpful in ensuring that all providers, parents, and teachers have the information they need in time to use it; at least one report each month would be a reasonable requirement (some providers will likely provide even more frequent reports, such as weekly). Ensuring regular progress reports is essential for at least three reasons:

- First, regular updates on student learning for parents and teachers will assist them in ensuring that students benefit and are reinforced on those skills in other contexts, such as at home or during the regular school day.
- Second, documentation of student progress in such reports, including performance on curriculum-embedded assessments—such as quizzes or unit tests—can be useful information at the end of the year, in conjunction with standardized tests, in determining whether to renew a provider's contract (of course, such provider-created assessments should be interpreted with care).[ix]
- Third, regular reports to students' parents or guardians enable them to determine when to intervene if progress is inadequate. Such intervention may consist of supporting the provider or others in improving learning, or it may consist of switching the student from a less effective provider to a more effective provider.

**Evaluating End-Of-Year Performance**

Determining whether to renew a contract with a provider at the end of they year presents certain challenges. The most significant for evaluation purposes is the reliability of assessments at the individual level. As noted above, the NCLB appears to be neutral on whether individual student learning goals for supplemental services should be established in terms of gains or absolute targets.

- Standards-based tests (or other types of criterion-referenced tests) designed to report whether students are scoring at certain assessment cut points or levels—such as Basic, Proficient, and Advanced—should be usable for evaluating whether individual students achieved those levels of achievement (evaluation or testing specialists can review the "categorical reliability" of particular tests to determine if they can indeed be used for this purpose). If so, the most significant question is whether the specified levels constitute appropriate supplementary instructional goals for each student; as noted previously, some students may be functioning far below one of the assessment levels while others may be receiving supplementary instruction focusing on only a subset of the standards. Other points on the scale of such a test will typically not be useful as targets since the test is likely to be relatively unreliable in determining whether individual students have reached those points. If a specific achievement level is identified as an appropriate supplementary instructional goal for a student, the LEA should still retain the option of considering other evidence, such as from sound curriculum-based tests, in special circumstances.

- Other types of goals and assessments may also be used to establish targets and measure progress for individual students, such as a gain analysis. As described previously in the state accountability section, the assessment must be well suited for the analysis of gains of low-performing students. Since this may be overcome in a variety of ways, the greater challenge is the limited reliability of annual gain scores for individual students. The range of possible true scores for any given reported score are so large that one simply would not be able to rely solely on an individual student's reported gains to evaluate whether to renew most supplemental services agreements. The table below illustrates this for several norm-referenced tests (the SAT9, Terra Nova, and Woodcock Reading Mastery Test), a computer adaptive norm-referenced test (the STAR reading and math tests), and the Florida standards-based test (the FCAT, one the most reliable and trustworthy state assessments).

  For example, at accepted levels of confidence (90% chance of being correct), an individual student's true gain on the SAT9 may range up to 10.8 NCE points higher or lower than the reported gain in any one year. So for a reported gain of +4 NCE points, a gain significantly greater than the norm (the norm gain is 0 NCE), the student's true gain would range from a –9.5 NCE drop to a +17.5 NCE increase. The full battery of the Woodcock Reading Mastery Test—Revised (WRMT-R) is more accurate at +/- 6.8 NCE, but a reported gain score of +4 NCE would still represent a range of possible true gains between –2.8 and +10.8.

**Table: Confidence Intervals For Individual Student Gains on Different Tests**

| Test | Test Reliability (0-1) | +/- Gain Score Range (90% confidence interval) |
|---|---|---|
| WRMT-R (Full battery) | 0.98 | 6.8 |
| SAT-9 | 0.95 | 10.8 |
| FCAT | 0.92 | 13.5 |
| STAR-Reading | 0.94 | 11.8 |
| STAR-Math | 0.85 | 18.4 |
| Terra Nova (CTBS-5) | 0.95 | 10.8 |

Multiple data points would help in identifying trends and reducing uncertainty with respect to individual student gains. But given the one-year duration of student supplemental education services agreements, options are limited for increasing the number of data points to allow for trend analysis. Most state-developed assessments are designed to be administered annually in the spring, allowing for two data points. NRTs can typically accommodate a winter testing "window," as well as fall and spring administrations, thus permitting up to three data points.[x] The Woodcock Reading Mastery Test-Revised (WRMT-R) may be administered even more frequently throughout the year, but it requires a trained individual to administer the test to every student.

**State and LEA Support for Parent Accountability**

In part due to these challenges, parents turn out to be key players in ensuring accountability for supplemental services providers. Research by Caroline Hoxby has demonstrated that the more parents have a viable opportunity to choose among multiple educational providers, the greater the impact on student achievement (for a summary of findings, see Hoxby in the Winter, 2001, issue of *Education Next*). Parents are able to monitor closely their own child's performance. While parent perceptions are not always correct, informed and involved parents with the power of educational dollars riding on their decisions certainly represent an empowered constituency. Both states and LEAs have responsibilities to assist in empowering parents to exercise choice among supplemental services providers.

States should make available to parents a variety of descriptive and performance information on approved providers through the state website or by other means. This type of information may include, but need not be limited to:

- Information drawn from providers' applications to the state and other provider sources, including evidence of past effectiveness as well as overviews of the instructional model and philosophy.
- A ranked ordering of the achievement gains of all supplemental services providers, public and private. This data should be presented in a manner that

makes clear to parents which differences are statistically meaningful and which are not, especially with respect to smaller and mid-size providers. State may determine not to publicize the results of the smallest providers if it endangers individual student privacy.

- How often parents switch away from each vendor, especially if information can be included on how often parents dropped a vendor due to dissatisfaction with the service (otherwise, providers will have a disincentive to serve high mobility populations). States can require that LEAs collect such basic information from parents as part of the procedure when parents switching providers during a school year or upon termination of a service agreement at the end of a school year. Data should be presented in a form that takes into account the number of students served by a provider (e.g., percentages).
- The results of parent satisfaction surveys completed by parents about their provider. The collection of such information does not have to be time-consuming and expensive. If entered through a secure web-based questionnaire at the local school or LEA, for example, results could be updated instantaneously on the state website. To ensure a high return rate, the state (or LEAs) could require all participating parents to fill out a short questionnaire on their provider(s) when switching, terminating or registering for another year.

LEAs have even greater responsibility, including:

- Providing user-friendly assistance to parents in selecting among providers, including parents with limited English proficiency and parents of students with special needs.
- Working with parents and providers to develop individual student achievement goals, timetables, and other required components of the service agreement.
- Ensuring the integrity of any special testing to be used in accountability for supplemental services, such as overseeing administration or even arranging for proctors where feasible.
- Assisting parents who are dissatisfied and interested in switching providers mid-year. LEAs should ensure that their administration of the supplemental services program is "user friendly" to parents interested in switching between providers during the school year. Contracts with providers should be designed to permit such terminations at parent discretion during the year. An LEA may, however, may wish to place an upper ceiling on the number of such switches to prevent chaos or excessive disruption for individual students (e.g., no more than three switches during one school year).
- Working with parents to determine whether to renew an agreement with a provider for another year; unless the objective data are unambiguously negative, we recommend that LEAs generally defer to parents on this decision.
- Collecting and disseminating survey results, or other types of information on providers, if the state is not doing so or if the state is not covering key areas of interest to parents.

Endnotes:

[i] The major off-the-shelf NRTs and some states that test consecutive grades have continuously scaled tests. To enhance reliable consecutive estimates of achievement, tests should be horizontally equated--assuring that different forms of the test at the same level give comparable results--and they must be vertically equated--linking performance of the test at one year, with the performance in the subsequent years.

[ii] Tests that are not linked will usually have reduced correlations, which lowers overall reliability.

[iii] Out-of-level testing also has the virtue of improving the alignment between the test instrument and the state achievement standards. The alignment of off-the-shelf NRTs with state standards would be diminished, given differences in content coverage requirements between a state CRT and the NRT. However, augmented NRTs that improve alignment with state standards would be a reasonable compromise test that provides somewhat better coverage for lower performing students.

[iv] The higher-quality Computer-Adaptive-Tests such as the STAR system are often designed to minimize testing time, with some sacrifice in reliability (the STAR math test as an average correlation below 0.90.) Longer high-quality CATs should have significantly higher reliability.

[v] Each of these models are especially designed to look at longitudinal gains with nesting. See for references. Sander's model is especially well-suited to determining the sizes of gains over two or more years were there are missing data.

[vi] The two-year requirement provides a significant advantage in setting up statistical analyses. The problems that can plague decisions based upon one year of data, such as measurement error and sampling error, are attenuated significantly with use of longitudinal data.

[vii] A power of 0.50 means that if there is an actual main effect, there is a 50-50 chance of detecting it. Research standards recommend approximate sample sizes to detect real differences 80% of the time (Power = 0.80). Because of the number of assumptions on the quality of the measures and the nature of the statistical designs that are ultimately used, estimation of statistical power is necessarily an approximation within an order of magnitude. In practice, these estimations give reasonable results to approximate whether *n* or *2n* subjects are needed, but not whether *n + 2* subjects is better than *n + 1*.

Technical Assumptions for Graphs 1 and 2: The graphs chart estimates of curves of constant group sample size for planned comparisons (one-tailed) of main effects between a vendor gain and a reference gain. Using a different statistical design, such as employing the Analysis of Covariance (ANCOVA), or substantially altering the number of comparisons, would yield different results. Calculations here are based upon the following parameters: (a) The pre-post test gains correlate 0.75. (b) NCE gain standard

deviation is 12 points. (c) The comparisons are planned, that is, each vendor gain is compared to the reference gain only. (d) Experiment-wise error rate is held at 0.05, and the Bonferonni adjustment for multiple (planned) comparisons is set at .005 for the large vendors (10 contrasts), and .002 for the small vendors (25 contrasts). (e) The reference gain for these power estimates is derived from control group matched in size of students academically comparable to provider samples.

Several factors that could well be encountered under real-world application will put downward pressure on these estimates, including unbalanced comparison groups, lower test-retest correlations, and larger gain variance. Lower test reliabilities would indirectly have an impact on the small vendors studies. Substantially reducing the number of planned comparisons would increase the power estimates.

[viii] Similarly, NRTs should be carefully evaluated for the appropriateness of their content alignment with suitable instructional targets.

[ix] Vendors should be encouraged to provide data that allows outside evaluators to conduct studies on the links between curriculum embedded tests and the evaluation instruments such as the NRT or state CRT.

[x] These NRT reliability estimates will be lower for abbreviated versions of the tests. For example, the SAT-9 reading in the full battery averages 0.95 across grades, but in the abbreviated battery reading reliability is somewhat lower at approximately 0.92, which yields a 90% prediction estimate of 13.5 points.

References:

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. Journal of Educational and Behavioral Statistics, 29(1), 37-66.

Bryk, A. S. and S. W. Raudenbush (1992). Hierarchical linear models: applications and data analysis methods. Newbury Park, Calf.:Sage.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum.

McCaffrey, D., Lockwood, J. R., Koretz, & Hamilton, L. (2003). Evaluating value-added models for teacher accountability. RAND Corporation: Santa Monica, CA.

Oakes, J. M. and H. Feldman (2001). Statistical Power for nonequivalent pre-test-post-test designs; the impact of change-score versus ANCOVA models. Evaluation Review, vol. 20(1). Pp. 3-28.

Sanders, W. L., A. M. Saxton, et al. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. Grading Teachers, Grading Schools: Is student achievement a valid evaluation measure? J. Millman. Thousand Oaks, Calf., Corwin Press: 137-162.