

Assessment of ELL Students under NCLB: Problems and Solutions¹

Monty Neill, Ed.D., Co - Executive Director, FairTest
July 2005

I. Problems with testing ELL Students²

There are a variety of problems with the testing of ELL students, problems that predate but are often compounded by the federal No Child Left Behind Act (NCLB, 2001). Here is a brief summary of some of the most salient issues. (For a history of federal policy and implications of NCLB for ELL students, see Wright, 2005). There are various possible solutions to these problems, some of which require federal, perhaps legislative, action; but some can be acted on at the state or local levels. I suggest possible solutions after each problem; the next sections of this paper focus in more detail on solutions.

A. *Unequal resources available to ELL students*

These students are disproportionately low-income and more likely to attend lower-resourced schools. SES remains strongest predictor of test scores, overall. ELL students must become proficient in English and learn in the subject areas. It is unlikely students who are not proficient in English can progress in content areas taught primarily in English as rapidly as do native English speakers. Also, there is a chronic lack of bilingual educators. NCLB's "highly qualified teacher" provisions may make it harder for schools to use bilingual teachers (if they must also be content-area certified); in addition, requirements for para-professionals appear to be driving bilingual paras out (though I have not seen hard data on this). In general, ELL students are likely to have far less opportunity to learn the content expected by state standards defining "proficient."

Solutions: resource adequacy and equity; extra resources for ELL students (advocates in NYC say funding ratios of under 1.5 are inadequate). This will help, but will not overcome fact that students are learning English while learning academic content, and thus the pace of progress may not be as fast as that of middle-income native English speakers.

B. *Different starting points:*

Schools with larger numbers or percentages of ELL students, which are often more ethnically diverse and more low-income, typically start behind in the "adequate yearly progress" (AYP)³ race (Neill and Guisbond, 2004). They have to catch up and make the same progress as others. This is, in general, unlikely. The more groups that are counted in the AYP process, the less likely the school is to make AYP – the diversity penalty. Many ELLs are also racial-ethnic minority and low-income, some have disabilities, meaning they may count in two or more groups' AYP

¹ This paper was first prepared for presentation to Iowa educators through the Iowa Department of Education. I have retained specific references to Iowa, but the material often will be relevant to other states.

² I use the term "English Language Learner (ELL)" for convenience and its widespread acceptance. NCLB uses "limited English proficient." "Bilingual" or even ESL (English as a second language) might be more appropriate, but fully bilingual students are not ELL or LEP.

³ AYP refers to the score gains students must make if their schools are to avoid sanctions.

results. ELLs typically vie with students with disabilities for the lowest scores, further contributing toward multiple chances for a school to not make AYP.

Crawford (2004) points out that ELLs themselves are a "highly diverse population in terms of socioeconomic status, linguistic and cultural background, level of English proficiency, amount of prior education, and instructional program experience. That is, the students themselves start out from very diverse places, and appropriate education must respond to that diversity.

Solutions: do not expect those behind to catch up and make additional AYP in a short time span (the "safe harbor provisions do this to some extent, but also appear to be unrealistic); use growth models that employ multiple measures (not just standardized tests) toward reasonably achievable ends (unlike 100 percent proficient by 2014; see below). The issue here is whether establishing "reasonable" levels will lead to establishing too-low expectations. The fundamental long term solution is to meet the varying needs of the actual students.

C. Changing composition of the ELL group

Even with federal regulations now allowing students who reach English proficiency to be included in the group for two additional years for purposes of measuring AYP, the basic problem remains that higher-scoring students exit the group while new, while non-English proficient students enter the group. This is compounded by the problem that students take 5-7 years to attain proficiency in academic achievement. Under this scenario, it is not reasonable to expect that in any given year all the students in the group will score proficient on tests. This guarantees that if there are enough such students for the group to count toward AYP, the school or district will fail.

- *Solutions:* for AYP, allowing "once in LEP always in LEP" will help but not solve the problem of group instability. Measure (with multiple indicators) success in moving children to English acquisition at reasonable rates (3 years is not enough) and in their making reasonable progress in academic achievement. In any event, the 100 percent proficient goal is not achievable. These will probably require changes law rather than in regulation, though perhaps some can be done via regulation (Sec. Spellings recently said that measuring progress, not only passing, is important for ELLs, and indicated there will be changes in regulations).

D. Inconsistent LEP classification

Older language proficiency tests have found to have low reliability in their use for determining language proficiency classifications, and results vary greatly across different tests (that is, one may say is English proficient while another says is not). In addition, Abedi (2004a, b) found correlation of only .223 between scores on Language Assessment Scales and LEP classification codes across the grades, and similar or lower correlations between LEP classification codes and scores on standardized tests such as Stanford 9 and ITBS. The tests may be particularly weak at the high ends, making them especially poor tools for evaluating exiting students. The Iowa ELL Handbook notes that language proficiency tests are often poor measures of progress in attaining proficiency. Iowa's report to the feds on Title III also noted the weaknesses in these tests (e.g., academic English; progress). The question is whether the new tests – ITELL and ELDA – are adequate.

Abedi (2004a, b) reports that the concepts underlying different proficiency tests vary. Also, states are involved in creating new tests, suggesting dissatisfaction with current tests.

However, coalitions creating new tests are not talking with each other, and it is not clear if they are ensuring sound underlying concepts or learning from problems in older tests.

Solutions: use an array of measures to identify students; work for consistency across the state; careful examination of reliability and validity of standardized instruments. New proficiency exams must have a clear operational conception of English proficiency and its acquisition, rooted in relevant scientific domains (Abedi, 2004 a, b; Gottlieb); whether this has been done well for ITELL and ELDA, the new tests developed by consortia of states? Consider use of classroom-based performance assessments, as Wisconsin is using (Gottlieb, 2003), for professional development and part of placement decisions. (Gottlieb said it has problems with reliability, but results for professional development and improved learning have been excellent; personal conversation, May 9, 2005.)

Also, as the IA guidelines note, be especially careful in identifying ELL students with disabilities so as not to over-identify.

E. Flaws in achievement tests used with ELL students:

ELL students are subject to tests of language proficiency, required under NCLN Title III, and to achievement tests, required under NCLB Title I. With all their flaws as noted above, the language proficiency tests are less problematic because most states have attached far more reasonable growth expectations than are attached to the achievement tests (see *Irrational Sanctions*, below). Achievement tests include both commercial, usually norm-referenced tests, and state tests that are supposed to be based on state standards.

1. Norm-referenced tests (NRT) are typically not normed with ELL students in the norming group. If the tests are going to be used on LEP students, they should be re-normed with such students in the norming group. James Crawford recommends that it would be best to determine at what level of English proficiency the English-language tests (e.g., with accommodations) become meaningful; then include students who are at or above that level of English proficiency in the norming group (personal communication, June 6). Note also that state exams may also lack sufficient numbers of ELL students in the groups on whom the tests are tried out during development.

Solution: re-norm standardized achievement tests as necessary, noting Crawford's recommendation. Do not use the tests on students for whom they are not normed. The tests themselves are likely to continue have weak accuracy and will remain inappropriate for many of those required to take them simply because their level of English proficiency is low.

2. Lower reliability of items for LEP students. Abedi, et al.(2003) report on analysis of the Stanford 9 (an NRT) showing substantially lower internal test reliability for LEP students than for FEP (LEP students who have become proficient) or for English-only (for whom reliability was higher than for FEP). The correlations between test scores and other academic measures were also substantially lower for LEP students. Other statistical and structural analyses produce similar results. Linguistic factors thus affect test reliability, meaning results for ELL are less certain than for English-speakers and achievement of LEP students may be underestimated, making it even more likely schools and districts will not make AYP.

- Solution: don't use or do not attach stakes to the use of these tests on LEP students; wait until they attain language proficiency (using several measures to ascertain academic language proficiency). More generally, use multiple measures of student achievement in content areas.

3. *Language complexity*: the complexity of the language used means students who may know the content may not be able to understand the test questions. Tests in English are reading tests as well as content tests. This is, for example, understandable in math, with word problems. But tests often use unnecessarily complex or unclear language – "superstandard English," an exaggerated form of English that may be used on tests since it helps sort test-takers, even though it is likely to be construct irrelevant and thereby contribute to test bias (Hoover, Politzer & Taylor, 1987).

Gottlieb (2003) sums up the evidence on accommodations as mixed. She argues that accommodations are largely used to retro-fit tests that are invalid for ELLs, but evidence is needed to determine if this has been a proper procedure. Most accommodations are taken from those developed for students with disabilities.

Solutions: language simplification, "rewording test items to minimize construct-irrelevant linguistic complexity" (Abedi, 2004a). Studies of reworded NAEP items consistently showed that appropriate linguistic simplification resulted in higher scores for grade 8 LEP students, but not for grade 4 (Abedi, *et al.*, 2003). (Abedi [2004, p. 7] provides details on nature of linguistic complexity.) ITBS/ITED tests do not allow this option, but should; at a minimum, the state should request the ITBS producers to evaluate language complexity of these tests.

Some studies suggest more time or use of dictionaries can help; other studies find these are not helpful (Abedi, *et al.*, 2003; Gottlieb, 2003). Reading items aloud may also help for students with stronger auditory than reading proficiency in English. These accommodations are acceptable to the ITBS makers.

More generally, use multiple measures, as Iowa's own guidelines suggest.

4. *Mismatches in native language, language of instruction, and language of assessment*. A student may receive content instruction primarily or wholly in English. S/he may have insufficient language to be fairly assessed in English. However, assessment in native language (if available) may not be helpful if the student is not literate in native language, or if the student has learned content in English and cannot access the knowledge in native language. I have heard of some assessments allowing students to take test and respond in both languages, but I don't think this is common and I don't know of any studies done of such efforts (this would not work if student is not literate in native language). Related is that students acquire language structures in any language well past age of school entry, so even if minimally literate in native language, level of that literacy may be inadequate for older students taking state exams.

All that said, for students who are receiving content instruction in a language other than English, tests should be available in that language. As Crawford points out, this would have the additional beneficial effect of helping schools resist pressure to eliminate bilingual education as they would not face preparing students to pass tests in English (personal communication, June 6, 2005).

F. Irrational sanctions under NCLB

If it is certain that ELL students as a group cannot meet the AYP requirements (true in reading by definition, unless all students can attain English academic proficiency within 3 years), then, under NCLB, progressively more severe sanctions will ensue. However, there may in fact be no problem that can be solved with the NCLB-mandated sanctions, since those sanctions are unlikely to enable ELL students to gain English proficiency at double the normal speed. More,

not making AYP may be based on results for one or two groups, yet sanctions will be applied to the entire school or district. This could lead to resentment toward groups that do not make AYP.

Solutions: change the way AYP and sanctions are conceptualized in the law; use multiple measures and indicators to better define problems (if there are real problems) so as to ensure properly tailored solutions; make sanctions fit the problem; apply sanctions to schools that focus on assisting only groups not making AYP.

II. Alternative Approaches

A. To sum up solutions proposed above, they include:

Establish common standards for definition of LEP and proficient that are rooted in up-to-date research in relevant areas – if this has not already been done.

Develop/acquire new, well-grounded tests of English language proficiency.

Revise and re-norm standardized achievement tests, and/or create simplified language tests; find test provider who will do so (persuade ITBS maker).

Ensure the use of multiple measures of language proficiency and content achievement and use those multiple measures in all decision making, and not in a way that only increases the likelihood of failing to make AYP. The former can be done by districts and guided by the state; changes to AYP must be made by feds: influence your delegation.

Change the AYP formula to address unequal starting points, ensure appropriate response to inevitably changing composition of ELL group in a school or district, and replace the expectation of 100 percent proficient with realistic goals in a realistic time frame. These are largely issues of the law itself.

Make sanctions fit actual problems. Example, under governmental reconstitution (final stage), there is a "something else" clause: this can be tailored to meet actual issues and needs, which can be determined through careful investigation, rather than standardized responses to test results.

Ensure adequate and equitable resources, including bilingual teachers and paraprofessionals, and additional supports for students who need them. While the federal government has inadequately funded NCLB, states themselves have too often allowed inadequate and inequitable schooling. States should take their own steps.

Don't be controlled by or fixate on test scores. That is, don't become a test-prep program (not teaching untested subjects; turning tested subjects into test-prep programs) for ELL or any children. This is an essential point: research across the nation shows that this is in fact happening frequently. The pressures to boost scores immediately are severe, but too much attention to the tests will probably backfire, not succeed in making AYP or in improving learning.

B. An alternate assessment approach

Since my main task is to address assessment issues, I will turn to some brief suggestions for the development and use of alternative approaches to assessment.

1. *Use multiple measures, including standardized tests but also classroom-based evidence of student achievement* (e.g., portfolios, work samples, grades; see Wisconsin for classroom measure of language proficiency; Gottlieb 2003). In using multiple measures for achievement and AYP purposes, do not allow tests to become a sole hurdle (conjunctive approach), but use a composite (combine different measures into one score) or compensatory approach (a high score on one measure overcomes lower score on another). Additionally, the relationship between summative measures (used for AYP) and formative measures (used to guide instruction and improvement, especially at individual level) needs to be clear: formative measures should not be frequent, mini-summative measures. Districts can take steps in this direction on their own; state support would help a lot.

2. *Using classroom-based evidence in reporting and accountability presents its own requirements, problems and complications*, including extensive professional development for teachers (especially in use of formative assessment), accuracy of measures, means of verifying results, means to incorporate multiple and diverse measures into a decision. There is a large literature on this, and I will here confine myself to a few suggestions:

a. Some experts emphasize the need to standardize classroom-based assessments (e.g., tightly define the content to be in the portfolio), while others focus on standardizing the scoring process. I am in the latter camp, though clearly guidance on what is to be in a portfolio is necessary in order that it contain material that can demonstrate achievement in the needed domains. For example, the Learning Record (n.d.) specifies the kinds of information that are needed, but leaves it up to the teachers to select the material. There is good evidence this approach can work, eliminating any purported need for detailed content mandates. In addition, tightly mandated approaches often produce substantial resistance from teachers.

b. While evidence on the value of standardized testing for improving achievement is thin at best, there is powerful evidence on the positive impact of formative assessment (Black and Wiliam, 1998). However, teachers are not well-prepared for using formative assessment. Educators who use formative assessment well also ensure the production of work by students that can be included, as appropriate, in portfolios and summative data. Thus, proper professional development in this area is vitally important.

c. It is infeasible to rescore across a whole state or even a district all student portfolios. It is also unnecessary. A more reasonable approach is to rescore many or all at a school level, some at district or regional, and some at a state level. Those rescored should be randomly sampled. Teachers should be paid to do this work. Results are used to provide feedback to originating teachers and as a check on the system. Teachers who are not accurate in scoring their own students or who do not know how to appropriately select student work need further education.

d. This approach assumes that the results are not immediately used for high-stakes decisions. The point is to create space for teachers to learn to use such a system and to provide them with help as needed. Careful attention needs to be given to how and when to use such evidence in a potentially punitive manner. Backwash effects (do teachers distort curriculum and instruction to fit the assessment, in ways that undermine good education) need to be carefully monitored.

e. If high stakes are not attached to individuals or schools based heavily or solely on scores on portfolios, then inter-rater agreement levels can be sufficient to use the portfolios in an overall evaluation program designed to unearth issues. This approach is of course counter to NCLB.

3. NCLB will need to be changed. The law links tests, progress requirements and sanctions in dangerous ways. The underlying theory of action (punitive) is unwarranted. An alliance of now more than 50 national education, civil rights, religious, children's rights and civic groups produced a "Joint Organizational Statement" on NCLB (on the FairTest website). It states, "*Overall, the law's emphasis needs to shift from applying sanctions for failing to raise test scores to holding states and localities accountable for making the systemic changes that improve student achievement.*" Implementing multiple measures and using formative assessment are tools for improving achievement as well as ensuring greater accuracy in assessment and thus in any subsequent actions.

C. A proposal for Iowa⁴

Iowa should implement a high-quality assessment program for ELL students and use that process to begin to implement such a program for all students. In doing so, IA should pay close attention to Nebraska and focus on building a local assessment system to meet both local and state needs. Nebraska allows districts to develop their own assessment programs to meet state standards (or comparable local standards), provided that the local assessments meet state quality standards. Nebraska requires districts to use an NRT in three grades, but does not use results from those tests in its accountability program. This is a good basic model. Within that approach, several possible additions should be considered:

a. Require that since the state already requires some standardized testing, the local assessments have to include assessment methods that go beyond those in the state exam. It could consider requiring districts to substantially rely on classroom-based measures.

b. A stronger means of establishing comparability across districts could be considered. "Moderation" sessions (e.g., scoring samples of portfolios) could be included.

c. It is possible for a state to have districts also use some performance tasks; such tasks can be obtained from banks of tasks, and they can be included in portfolios.

d. A guideline-driven portfolio can be supported for use with ELL students, then generalized for all students.

Classroom and school-based assessment within a framework, plus standardized tests, can be used for determining proficiency as well. All of this requires substantial, ongoing professional development.

⁴ Crawford (2004) has an interesting proposal for authentic accountability based on the Castaneda decision.

In addition to using standardized tests and classroom-based information, an accountability system should include the following:

a. Periodic, in-depth reviews by qualified teams, similar to accreditation or the school quality review processes used in New Zealand, England, and Arizona.

b. Other measures such as dropout and graduation rates, grade promotion rates, etc. These need to be carefully defined and established, and carry significant weight in an evaluation process.

c. Measures of other indicators deemed important to schools, beyond academic measures, such as evaluations of school climate (see "Draft Principles for Authentic Accountability," FairTest 2004).

A final word: For many immigrant students, schools raise complex issues of culture and relationship to family and community. Many perceive schools as undermining identity, in part through devaluation of native language. This can, among other harmful consequences, undermine student achievement. Thus, schools should to the extent possible support students home, community, culture and language. Among other things, teachers and paraprofessionals from the community should be employed. Since in the end the point of assessment and accountability is to improve teaching and learning, work on assessment must interact with work on instruction. Among other things, states and districts should consider support for two-way bilingual programs.

Bibliography

Abedi, J. 2004a. The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher* 33(1), 4-14.

Note: A summary of more detailed reports, listed in references, many available on the CRESST website. A shorter version of this article is Abedi and Dietel (2004).

Abedi, J. 2004b. Inclusion of Students with Limited English Proficiency in NAEP: Classification and Measurement Issues. Los Angeles, CA: CRESST, UCLA, CSE Report 629; May. http://www.cse.ucla.edu/products/reports_set.htm

Abedi, J., and Dietel, R. 2004. Challenges in the No Child Left Behind Act for English language learners. *CRESST Policy Brief 7* (Winter). <http://www.cse.ucla.edu/products/newsletters/policybrief7.pdf>

Abedi, J., Courtney, M & Leon, S. 2004. Effectiveness and validity of accommodations for English language learners in large-scale assessments. Los Angeles, CA: CRESST, UCLA, CSE Report 608, September. http://www.cse.ucla.edu/products/reports_set.htm

Batt, L, Kim, J, & Sunderman, G. 2004. Limited English proficient students: Increased accountability under NCLB. Cambridge: Civil Rights Project at Harvard University, Policy Brief, February. http://www.civilrightsproject.harvard.edu/research/esea/lep_policy_brief.php

Black, P., and Wiliam, D. 1998. Inside the black box. *Phi Delta Kappan*. October, 139-148.

Crawford, J. 2004. No Child Left Behind: Misguided approach to school accountability for English language learners.
http://www.nabe.org/documents/policy_legislation/NABE_on_NCLB.pdf (accessed July 7, 2005).

FairTest. 2004. Draft principles for authentic accountability. Cambridge, MA: author.
<http://www.fairtest.org/nattest/Authentic%20Accountability/Draft%20Principles.html>

FairTest. 1995. Selected annotated bibliography on language minority assessment. Cambridge, MA: author. <http://www.fairtest.org/bilingbib.html>

Note: quite extensive, including critiques of testing and discussions of alternatives; now somewhat outdated.

Gomez, E. 2000. Assessment portfolios: Including English language learners in large-scale assessments. Washington, DC: Center for Applied Linguistics, Digest EDO-FL-00-01, December. <http://www.cal.org/resources/digest/0010assessment.html>

Gottlieb, M. 2003. *Large-Scale Assessment of English-Language Learners: Addressing Educational Accountability in K-12 Settings*. Alexandria, VA: Teachers of English to Speakers of Other Languages, TESOL Professional Papers #6.

Hoover, M.R., Politzer, R. L, & Taylor, O. 1987. Bias in reading tests for Black language speakers: A sociolinguistic perspective. *The Negro Educational Review*, Vol. XXXVIII, Nos. 2-3, pp 81-98.

Joint Organizational Statement on No Child Left Behind (NCLB) Act. 2004.
<http://www.fairtest.org/joint%20statement%20civil%20rights%20grps%2010-21-04.html>

Learning Record. See http://www.fairtest.org/Learning_Record_Home.html

Note: An assessment tool and practice, first developed in London, England, in part for use with bilingual students. Complex to use, very rich, useful for formative and summative, individual and program evaluation.

Neill, M., and Guisbond, G. 2004. *Failing Our Children*. Cambridge, MA: FairTest. On the web at http://www.fairtest.org/Failing_Our_Children_Report.html

No Child Left Behind Act. 2001. P.L. 107-110.
<http://www.ed.gov/policy/elsec/leg/esea02/index.html>

Wright, Wayne E. 2005. Evolution of Federal Policy and Implications of No Child Left Behind for Language Minority Students. Tempe: Arizona State University, Education Policy Studies Laboratory, Language Policy Research Unit, EPSL-0501-101-LPRU; January.
<http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0501-101-LPRU.pdf> (The LPRU has a variety of valuable policy briefs.)

- Note that other works provided helpful background for me. I listed above mostly works that are accessible to the non-expert and are mostly readily available (e.g., by internet). The bibliographies in many of these works are themselves of great use.

Some of the work performed in producing this paper was done under contract for the Iowa Department of Education. The opinions and conclusions are my own and do not necessarily represent those of the Iowa Department of Education.

Thanks to participants in workshops in Iowa and in North Carolina, as well as comments from several readers, all of which I believe improved this paper.