A STUDY DESIGN TO EVALUATE STRATEGIES FOR THE INCLUSION OF L.E.P. STUDENTS IN THE NAEP STATE TRIAL ASSESSMENT

Kenji Hakuta and Guadalupe Valdés

School of Education
Stanford University

10/18/94

Paper prepared for the National Academy of Education Panel on NAEP Trial State Assessment. Correspondence address for both authors: School of Education, Stanford University, Stanford, CA. 94305. Fax for Hakuta: 415-723-7578; for Valdés: 415-725-7412.

A STUDY DESIGN TO EVALUATE STRATEGIES FOR THE INCLUSION OF L.E.P. STUDENTS IN THE NAEP STATE TRIAL ASSESSMENT

Kenji Hakuta and Guadalupe Valdés
Stanford University

BACKGROUND

The issue of the inclusion of L.E.P. in NAEP is best understood in the context of the current era of systemic reform (especially the passage of Goals 2000 and the implications of this movement toward outcome based accountability for students whose educational attainments cannot be readily assessed. In some ways, L.E.P. students can be "caught between a rock and a hard place." In the absence of adequate inclusion in the system of assessment, one fear is that L.E.P. students will not be "counted," thus not given equitable access to resources. On the other hand, if they are assessed, thus counted, but assessed inappropriately, the concern is that they will be penalized, given the increasingly high-stakes nature of the assessment. Although NAEP is not a high-stakes assessment for the student, it is a national measure of educational attainment that will be taken seriously, and thus any data that will be reported for the subgroup of students known as L.E.P. can be anticipated to have broad impact on policies and attitudes toward this population.

Recently, a prominent and diverse group of experts on the education of L.E.P. students developed a consensus document with recommendations regarding *Goals 2000.*[1] Although they were more concerned about state level assessments, their recommendations can be directly extended to NAEP. We quote their recommendations at some length to provide the policy context for the study we propose:

---

[1] *"For All Students": Limited English Proficient Students and Goals 2000.* Recommendations based on a series of meetings sponsored by the U.S. Department of Education (OBEMLA), the Carnegie Corporation of New York, and the MacArthur Foundation. Stanford University, School of Education, September 2, 1994.

In most states, however, L.E.P. students are not assessed for accountability purposes until they have acquired a certain level of English proficiency and/or have been in a school system for a specified period of time. As a result, L.E.P. students are often exempt from testing for accountability purposes. Even when L.E.P. students are included in assessments, scores are often not reported by L.E.P. status. Thus, the data on how L.E.P. students are progressing against the standards of a particular school, district, or state are quite limited and/or not easily accessible. The result is that no one is ultimately responsible for ensuring that L.E.P. student receive high quality instruction comparable to that provided to their English-speaking peers.

If the reform process is to make a difference in the education of L.E.P. students, they too must be included in assessments. However, for L.E.P. students, assessments which rely on standardized norm-referenced tests in English have historically been problematic. These assessment instruments actually assess both content concepts and language ability -- in particular, reading comprehension and writing. The interconnection of language and content makes it difficult to isolate one feature from the other. As a result, it is almost impossible to determine whether a student is unable to demonstrate knowledge because of a language barrier or because he or she does not know the content material being tested. Adding to the problem is that such assessments are generally not aligned with the school curriculum. Furthermore, they are usually normed on non-L.E.P populations and thus scores cannot be interpreted for L.E.P. students. In short, traditional assessments are not designed with L.E.P. students in mind. Often they simply become measures of L.E.P students' language proficiency rather than measures of content knowledge, as they are intended to be.

An assumption implicit in Goals 2000 is that new assessments such as performance-based measures and portfolios will change the nature of the teaching/learning process and that these new assessments will enable students to more aptly demonstrate what they know and can do. However, even with new assessment technologies, equity is still a key concern for L.E.P. students. For example, many new assessments emphasize English communication skills and subject matter knowledge and thus place a heavy demand on the English skills of L.E.P. students. Moreover, as with traditional assessments, L.E.P. students continue to be exempted from these assessments until they reach a certain level of English language proficiency, thus maintaining the issue of lack of progress and accountability data for these students.

If L.E.P. students are not assessed, no one can really be held accountable for what these students know and can do in important content areas. Thus, we recommend that states develop performance assessments that are appropriate for L.E.P. students.

L.E.P. students who are instructed in their native language, should be assessed in that language. L.E.P. students who are better able to demonstrate content knowledge in their native language, even though they have not received native language instruction, should also be assessed in their native language.[2] The native language assessments should parallel content assessments and performance standards in English. States with substantial numbers of L.E.P. students in given language groups should include a process in their state plan for developing or borrowing (from other states or entities such as large school districts with

[2] Such assessments are particularly important for students who have been educated in other countries and thus are able to demonstrate content knowledge in their native language.

substantial L.E.P. students) content area assessments in languages other than English. This process might also involve cooperative efforts among two or more states, or the development of multi-state item banks, and should include persons knowledgeable about the assessment of L.E.P. students and systems serving them.

Modifications in assessments and assessment procedures should be encouraged to enable L.E.P. students to take content assessments in English. These modifications might entail: altering the procedures used to administer the assessments (e.g., giving instructions in the native language, allowing students to respond in their native language, using think-aloud techniques); modifying the assessment itself so it is more comprehensible to L.E.P. students (e.g., decreasing the English language demands); using alternative assessments (e.g., portfolios to collect the student's best work over time); and employing computer-assisted assessments that are tailored to the language needs and content knowledge of L.E.P. students. In all instances, however, it is important to ensure that assessments are equivalent in content and rigor to those used to measure the progress of fluent English speakers.[3] It is not imperative that these assessments be the same as those given to fluent English speakers. However, to gauge the progress of L.E.P. students, the assessments must remain comparable over time.

Until the psychometric issues underlying these new assessments have been addressed, and until mechanisms to ensure opportunities to learn have been fully implemented, these assessments should not be used in high stakes testing for students.

---

[3] There will have to be considerable research and development in the construction and evaluation of these instruments before this becomes a reality.

The thrust of the policy discussion is clearly to maximize inclusion as well as equitable assessment. The inevitable tension caused by these often two converging pressures will surely lead to an information vacuum about alternatives and their viability. This paper will begin this exploration by identifying the some options that exist for state NAEP to include L.E.P. students, and offer a preliminary study design to shed light on the wisdom of the options.

PRELIMINARY DISCUSSIONS WITH STATE NAEP TRIAL ASSESSMENT PANEL

At the last meeting of the State NAEP Trial Assessment Panel meeting in June, we met with the panel and NCES staff to discuss the broad outlines of our charge. Based on the discussion, the following working principles were identified as important in developing strategies for L.E.P. inclusion:

Maximal Inclusion Principle

Ideally, every student in each state should have an equal probability of being included in the assessment sample.

Continuum of Strategies Principle

Looking for a single strategy to provide the solution is unrealistic, i.e., "one size fits all" will not work. Rather, the appropriate view is that there is a continuum of options available to support assessment, ranging from valid to not valid. These options should be treated as a working set, with ongoing attempts to (1) maximize the number of students who are offered options on the valid end of the continuum, and (2) increase through R&D the validity of options on the not-so-valid end of the continuum. Using the entire range of the continuum would enable inclusion of all students, even though some of the students would only be included through the use of assessment strategies of non-comparable validity.

Use of supportive and alternative assessment strategies must be supported by research that shows comparability. For example, assessment in the native language for students who

receive instruction in that language, assessment in English using special administrations; and alternative assessments that might include ratings or portfolios.

Reality Principle

Only options that are realistic in the context of policy and NAEP should be considered. This principle would lead to the choice of group-administered over individually-administered assessments whenever possible. The principle further requires clear groundrules and criteria that trigger the different assessment support strategies. In addition, assessment supports and alternative assessments must take into account an "urban reality principle" of teacher stresses and teacher turnover, such that in cases of special administrations, the burden should be on the NAEP assessor, rather than on the teacher.

Based on these principles, then, the task is to identify a parsimonious set of alternatives that would optimize the number of students that would be validly assessed yet minimize the number of alternatives, and to keep the decision flow simple and realistic within the NAEP context.

In addition to these principles, the following considerations were raised in the discussions as important features of any inclusion strategy:

Categories of Students

The three major factors are the grade level, native language, and prior educational experience of the students (including amount of formal educational experience and enrollment in bilingual education programs).

Content Area and Domain of Assessment

Clearly, some content areas being assessed are more dependent on language than are others (for example, reading versus math). However, the current trend in assessment is tapping into knowledge that are increasingly language-based (for example, requiring an explanation for

a solution to a mathematics problem), leading to an artificial separation between language proficiency and demonstration of content knowledge. Lurking in the background of the assessment problem for L.E.P. students is the question of the extent to which language proficiency and content knowledge are separable.

CONSIDERATIONS OF STUDENT BACKGROUND

There are many different kinds of L.E.P. students in terms of their educational experiences and access to English. A simplistic view of L.E.P. students, unfortunately prevalent even among educational experts, maintains the following:

> Students speak L1 at home in infancy, enter 1st grade, are served by bilingual education programs and receive instruction in L1 in grades 1-3, have access to parallel curriculum as mainstream children. If they are exited from bilingual program and placed in English medium instruction in grade 4: they can be assessed in English at grade 4. If they are not exited and still classified as L.E.P., the best language for assessment would be Spanish.

The reality, however, much more complex. Even in grades 1-4, students enter both all English instructional programs or bilingual programs at different points and shift between programs. What needs to be avoided is the presumption that non-English-background children remain in the same kind of program during their early schooling experience (grades 1-4). Because of high family mobility, the following patterns of movement are typical:

Grade 1: Bilingual education program- Access to instruction in L1 Grade 2 All English program with ESL support
Grade 3: Bilingual education program-Access to instruction in L1 Grade 4 All English program

Grade 1: School in Spanish in home country
Grade 2: All English program with ESL support
Grade 3: Bilingual education program- Access to instruction in L1 Grade 4 All English program with ESL support

Thus, in selecting language of assessment for Grade 4, recency and extent of instruction in L1 needs to be determined.

The situation is even more complex in grades 6-8 and 9-12, since there is generally little or no L1 instruction available, and students enter the U.S. at different ages. Thus, 8th and 12th grade classes of L.E.P. kids include:

- newly arrived immigrants with high literacy skills and good L1 school experiences;
- newly arrived immigrants with low literacy skills and limited L1 school experiences;
- students schooled exclusively in US and instructed in both L1 and L2 or only in L2.

An 8th grade student for example, may have arrived in the U.S. in grade 5. While highly literate in L1 and schooled from grades 1 to 4 in L1, she has received no instruction through Spanish in the U.S. Instruction in grades 6-8 has taken place exclusively in English. She has been enrolled in 3 periods of ESL, art, cooking, P.E., L.E.P. computers, and L.E.P. math. Testing such a student in Spanish may be very problematic. The problem here is recency of instruction in L1.

Additionally, different schools offer different types of access to English. An 8th student schooled exclusively in English since grade 2 in a predominantly Latino urban school may, in spite of such instruction, still be very limited in his English language abilities. However, neither will he have developed his ability to use Spanish for academic purposes. Again, testing this student in Spanish may be very problematic.

Sample Student Questionnaire

We offer here a set of questions, similar to those used with children by *Linguistic Minorities Project* in England,[4] as a starting point to gather information on the complexities of student background described above.

Questions about language preference for academic tasks

Language preference (questions of following type)

Language I read most comfortably in

[4]Institute of Education (1983). Linguistic Minorities in England: A Report by the Linguistic Minorities Project for the Department of Education and Science. London: Institute for Education.

Language I write most comfortably in
Language I think in when I do math
Language in which I explain math problems best
Language in which I do best on tests

Questions about languages used in instruction

Grades in which you received instruction through Spanish

Check all that apply

___1 ____2 ____3 _____4 ____5 ____ 6 ____ etc.

Recency of instruction in Spanish

Present grade_____
Last grade in which I received instruction in Spanish_____

Grades in which you received instruction through English

Check all that apply
___1 ____2 ____3 _____4 ____5 ____ 6 ____ etc.

Questions about language background

Age of arrival
Years in US
First language
Age at which English acquisition began
Language(s) spoken at home
Sources of English outside of school

Questions about language use

Frequency of use
Range and mode of use home/neighborhood/school

Responses to these questions can be used to categorize students into the taxonomy of background and instructional characteristics as schematically outlined in Table 1.

Insert Table 1 here

Treatment Design Rationale

In establishing the usefulness of various strategies to maximally include L.E.P. students in NAEP, there are two approaches.

One would be to make a set of options available to the test administrator or a person in the student's school, and to let that person make the choice. This approach would, in one sense, match the reality of the field -- for example, currently, it is left up to the test administrator, based on information gathered at the site and consultation with school authorities, whether or not to include or exclude an L.E.P. student. This approach, however, will not to provide useful information on the relationship between the particular accommodation strategy and test performance, since the choice of accommodation strategy will itself be confounded with test performance.

A more rational approach would be to randomly assign the test conditions to students. Unless a large number of schools are sampled, it is important to randomly assign conditions to students within schools, rather than to employ the same strategy for any given school. Obviously, the law of convenience would press in the direction of assignment by school, but the large school-to-school variability in L.E.P. student characteristics would pollute the power of the comparisons.

1. Native Language Assessment in Spanish, and Exploration of Bilingual Assessment

For Spanish, it is realistic to expect the development of an assessment in the native language. But one problem is that most native speakers of Spanish are instructed in English, so an assessment just in Spanish may not be universally appropriate for L.E.P. students whose native language is Spanish. Perhaps fairer would be an assessment that includes both languages.

However, bilingual assessment is not universally favored among experts.[5] Exploration of systematic differences in performance between Spanish and bilingual side-by-side versions is needed. Currently, ETS is conducting a study for NAEP that assesses the scalability of math items that are administered in Spanish and in side-by-side translation versions. This study is important in revealing the psychometric properties of items under these conditions, and will tell us whether the minimum terms of comparability are met. However, it will not address whether such modifications result in *improved* performance for L.E.P. students.

2. Adaptations of English Assessments

About 25 percent of L.E.P. students are speakers of languages other than Spanish. It is not realistic to assume that native language assessment will become available for these students any time in the foreseeable future. Thus, for these students, unless modifications are made to English assessments to make them more accessible, they will continue to be excluded from NAEP.

The universe of possible candidate modifications is large, and there is little basic research in this area to help inform our guesses. However, the universe may initially be divided into those that provide support during administration of unmodified items, and those that involve actual modification of the items (hence possibly affect their validity). Thus, along the continuum of validity, these two approaches may be ordered as the former being more valid than the latter. In light of the importance of having L.E.P. students participate as much in the ``real thing" as possible, it would be a useful principle to place greater emphasis on strategies that offer support, with a secondary emphasis on those that provide modifications. However, serious R&D effort

[5]For example, a group of experts convened by the California State Department of Education wrote: ``Bilingually structured assessments, defined her to mean a single assessment instrument or procedure administered during a single time period in two languages, are extremely difficult to design and almost impossible to evaluate in any meaningful way. In most cases, such assessments are unlikely to reveal anything more informative than would be obtained from separately administered tests in two languages. Because of the problems associated with developing, administering, scoring, and interpreting results as well as financial constraints associated with mixed language assessments, their use is not recommended as a general practice for large scale assessments of language or academic matter." *Assessing Students in Bilingual Contexts: Provisional Guidelines* (p. 9). Bilingual Education Office, California State Department of Education. July, 1994 (Prepublication Edition).

should continue to vigorously investigate the approaches that involve modification with the assumption that it is remains technically imaginable -- and empirically investigable -- that the best of the modified approaches will not compromise validity.

With respect to the universe of possibilities that offer support during administration of unmodified items, we recommend piloting a format that provides additional, clarifying information at the end of the booklet. This approach would not be feasible for the 4th grade sample, but should be within the competence of the 8th and 12th grade samples. One addendum might be an English-Spanish glossary for vocabulary for which difficulty may be anticipated. A second addendum might be an English annotation for the same words. This modification would require an increase in test-taking time to allow students the time to use the information in the addenda. The amount of increase can be determined in during pilot testing.

Another approach might be to have a bilingual administrator available to support administration and answer questions during the test. However, preliminary findings from such an approach in an experimental administration of California's CLAS (Kopriva, personal communication) indicates considerable cueing by the administrators. Thus, we would not recommend this approach, subject to a more thorough inspection of the final analyses of the CLAS results.

Linguistic modification of NAEP math items has been investigated in an on-going study for the TRP by Jamal Abedi of UCLA's CRESST. Although the results are still available only in preliminary form (a Progress Report dated circa September, 1994), they are instructive. One phase of this study analyzed grammatical features of some items, and found huge effects of L.E.P. status on performance on items that were linguistically complex ($F[1,1170]=56.42$). However, the report does not indicate the degree of interaction, i.e., whether the effect of L.E.P. status was significantly less on the linguistically simple items. The second phase, more relevant to our interest, identified items with linguistic complexity, including ``familiarity/frequency of non-

math vocabulary, length of nominals, voice of verb phrase, conditional clauses, question phrases, and abstract or impersonal presentations". Items were then simplified, while keeping the content of the items intact. In an interview substudy, they verified that L.E.P. students in fact reported the simplified items to be easier to understand. However, the actual performance study was disappointing, yielding no statistically significant effects related to linguistic modification. Abedi shows figures that suggest an interesting trend that students in the low to intermediate math levels tend to profit from the modifications. However, the effects are not significant, and even a visual inspection of the graphs requires an active imagination to appreciate the trends.

Based on the preliminary findings of Abedi's study, it would seem, at least on the math items, that simple grammatical modification is not sufficient. Lexical and semantic modifications may be worthy of exploration. Creative ideas, followed up with good pilot work, are needed.

3. Capturing the Remainder through Unconventional Alternatives

Assuming that there will still be a sizeable proportion of the L.E.P. student population that is left out of assessment even with the availability of Spanish assessment and some modifications,[6] it would be informative to collect data on the excluded students, even if the data may not be fully valid and reliable. One method is to obtain ratings from teachers knowledgeable about the students (or, if this is not available, from NAEP staff who interview the students) as to how they would have performed on this test. For example, a teacher may be asked to ``imagine that the student was fully English proficient, and took the test today." The teacher would then be shown a range of possible responses to each item, and to identify the one that would most likely have been made by the student. Scores can then be assigned to the students as if they had taken the test. Considerable background work is needed to determine the conditions under which such ratings are robust.

[6]The *Prospects* study (Abt Associates), in their oversampling study of L.E.P. students, offered the possibility of administering students achievement tests in math and reading using the Spanish *SABE*, considered roughly equivalent to their primary outcome measure in English, the *CTBS*. Even when this possibility was available, approximately 25% of L.E.P. students were excluded from either assessment.

4. Evaluating Interaction of Adaptations with Item Type

The effect of the modification treatments will most likely interact with the type of item, *especially the extent to which the item requires a high level of proficiency in English.* Item selection, if conducted strategically, can afford measurement of within-subject effects of item types and provide a window into specific interactions of language proficiency and item characteristics.

5. Comparison Sample

For comparison purposes, it would be important in sample language minority, non-L.E.P. students to better understand some of the treatments. Targeted comparisons can determine whether (1) the English language modifications are useful even for English-fluent students; and (2) the teacher/staff ratings are accurate. One might consider two categories of LM/non-L.E.P. students: (1) those whose native language is English, and (2) those whose native language is Spanish and who were at some point classified as L.E.P.

IDEALIZED DESIGN

There are at least 5 versions of the test: (1) English; (2) Spanish; (3) bilingual side-by-side; (4) modified administration, but items unmodified; (5) modified administration and linguistically modified items.

These conditions would be between-subjects, randomly assigned within school, to all L.E.P. students in the chosen grades. In addition, LM/non-L.E.P. students in the school would be randomly assigned to conditions (1), (4), and (5).

There is one within-subjects factor, item type (the levels of which will be determined, but definitely will vary according to language-loadedness). In addition, all subjects will be rated by their teacher/others for how they might have performed, blind to their actual performance on the test. Thus, whether rating-actual performance difference varies as a function of administration condition can be assessed through the interaction effect in the ANOVA.

Other design variations that make more powerful comparisons are also possible, subject to the reality principle. A repeated measures design for the main conditions would more powerfully test the differences.