Improving the Equity and Validity of Assessment-Based Information Systems*

CSE Technical Report 462

Zenaida Aguirre-Muñoz and Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

December 1997

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

* This document is excerpted from Z. Aguirre-Muñoz and E. L. Baker, "Improving the Equity and Validity of Assessment-Based Information Systems," in *Challenges Minorities Face in Educational Testing and Assessment*, ed. M. Nettles (Boston: Kluwer, in press).

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

IMPROVING THE EQUITY AND VALIDITY OF ASSESSMENT-BASED INFORMATION SYSTEMS

Zenaida Aguirre-Muñoz and Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

This report focuses on issues of validity and equity of assessments as they guide educational policies and practices for the education of limited English proficient students. Although estimates of the number of students who are English language learners (ELLs) vary, from self-reports in the 1990 census (U.S. Bureau of the Census, 1990) to surveys conducted of schools districts (Fleischman & Hopstock, 1993), their proportion is rising and may reach 10% by the end of the century. Although Spanish is the primary language for about three fourths of these students, Asian group languages—Vietnamese, Hmong, Cantonese, Cambodian, Tagalog, Laotian and Korean—are represented in large numbers. Navajo and Russian are also significantly represented. The case of limited English proficient students is particularly instructive, for it illustrates the unprecedented challenge posed by the educational reforms of the 1990s: the simultaneous call for higher standards of performance in content areas and the inclusion of children of all backgrounds in the reform movement. Although this expanded set of requirements may be regarded by some as little more than optimistic rhetoric, state and federal legislation has been enacted to create policies and practices intended to raise the attainment of limited English proficient children. The challenge is twofold: to change the perceptions of the public and teaching personnel so that these goals may be accepted; and to achieve the twin goals of increased attainment and expanded participation. In the case of students who are not fluent in English, the situation is complicated by diverse public perceptions on the use of primary language in school. At the heart of much of the discussion is the role of language in student achievement and the expectation by a majority of the public that learning English should be a priority. Controversy exists, for instance, on the degree and length of time of maintenance of primary language in instruction. There is also a strong basic education movement in some sectors of the public, exemplified by the pressure for

computationally oriented mathematics and phonics-based reading programs. These advocates take the position that the education system should demonstrate that it can teach children fundamentals before it tackles higher standards and more ambitious goals. The great success of the American system, its retention of more students through high school, is also its downfall, for the lack of demonstrable skills for many of these students is unacceptable. As the proportion rises of students in school who have home languages other than English, pressure increases for better approaches to teach and assess their learning.

Our focus is not on the desirability or relative empirical merits of bilingual, immersion, or other approaches to develop language proficiency in English. In education, there are rarely main effects. Instead, we address the problem of assessing students in content areas other than English in order to determine their levels of attainment in the subject matter. Even an approach that assists children to display mathematics competence without unnecessary language interference, for example, is controversial. But putting aside for the moment the problem of public credibility, there is considerable difficulty in meeting the technical challenge of assessment, and we must trust the results of tests if we are to act upon them. The technical attributes of assessments, especially approaches to make them fair and accessible to students who are not fully competent in English, are of great interest nationally. Since 1990, the teaching and assessment communities in the United States have been intensely exploring new forms of assessments, and the forms of these assessments greatly complicate the problem of testing in other than native languages. The newer, performance-based assessments require longer tasks, more complex cognitive processing, and deeper subject matter knowledge. They also require elaboration or explanation typically displayed through speaking or writing. These assessments are thought to be a major method to operationalize higher expectations for American children and youth. In their design, they consolidate new knowledge about student learning processes, subject matter expectations, and extrapolations from analyses of examination systems abroad. They also add new demands for students who are learning English.

Requirements for assessments of challenging content are embodied in legislation designed to impact disadvantaged children, the Elementary and Secondary Education Act (1965), Improving America's Schools Act (IASA, 1994). This legislation also requires that children who have not yet developed English

competency be included in the evaluations of schools that receive federal funding. Moreover, in an effort to include students whose performance has not been reported in the past, the policies clarifying the IASA provisions require that assessment of student progress be accomplished using, where indicated, linguistically appropriate assessments for English language learners (ELLs)—students whose native language is not English and who demonstrate low English proficiency (see section 200.1, 200.4 Exec. Order No. 12866, 1995). Although the decision of which type of assessment to administer is largely left to local school districts, many are using combinations of multiple-choice and performance tests to meet these requirements.

Performance assessments by design require both linguistic and content-related skills for their successful completion. For native speakers of English, differences in performance among different groups of students exist because of differences in familiarity, exposure to the content of the test, instruction, and motivation to complete the task (Linn, Baker, & Dunbar, 1991). ELLs have an added difficulty, as the language on the test may place them at a disadvantage (Baker & O'Neil, 1996).

The extent and nature of the impact of language skills on performance assessments remains elusive due to the paucity of research in this area. However, lessons can be learned from what is known about the impact of language skills on standardized test scores. Studies that explored the impact of language background on standardized test scores have found three factors affecting ELLs' test performance: (a) limited second language skills (Figueroa, 1990), (b) limited background knowledge of the implicit meaning of the text within a test (Hafner & Ulanoff, 1994), and (c) limited access to content-specific knowledge, such as assignment into classes with narrow curricular coverage (as argued by the work of Oakes, 1990, and Stanovich, 1991).

The first two factors influencing ELLs' test performance involve linguistic factors that provide the basis for the accommodation requirements in the IASA legislation. High-quality procedures are not available to assist in identifying students who should receive linguistic support during testing. The typical approach is to administer some measure of English language proficiency. Yet many widely used English proficiency measures have weak validity and reliability data (see Zehler, Hopstock, Fleischman, & Greniuk, 1994, for a list of reviews). These measures often lack clear construct definition, adequate scoring

directions, and high-quality norms. In some cases, such as with the Bilingual Syntax Measure, validity information is not reported at all (Valdes & Figueroa, 1995). In a recent summary of research progress in the area of inclusion (Olson & Goldstein, 1997), Cheung, Clements, and Miu were reported to have documented a wide range of methods to identify and monitor ELLs' progress, including the review of archival records, home language surveys, observations, and interviews (1994, as cited in Olson & Goldstein, 1997). Yet, Hopstock and Bucaro (1993) report that 83% of local districts used English language proficiency testing to determine, in whole or in part, students with English language limitations.

Even if English proficiency identification were perfect, there would be a need to provide tests that enable ELLs who do not pass an English proficiency test to display their competence in school subjects. Why are linguistic accommodations necessary when assessing ELLs' content understanding with complex performance assessments? Two issues related to the cognitive demands of performance assessments are important to note. First, the design of most performance assessments demands higher levels of understanding of content-specific language, or academic knowledge (conceptual as well as factual information). Attempts to assess ELLs' content understanding may result in underestimating their knowledge. Unlike basic conversational English, there is convincing empirical evidence indicating that the ability to use English for academic purposes takes several years (approximately 5 to 7 years) to develop (see Collier, 1987, 1989). While this academic language is developing, students will need help in demonstrating their knowledge acquisition, a point to be expanded upon in the next section. This situation is exacerbated by the curriculum access issue. If lower performing students are less apt to receive complex curricula, ELLs are less likely to be exposed to the academic language necessary to do well on complex performance assessments. Therefore, not only is it important to identify students appropriately and to provide ELLs with linguistic accommodations in testing situations, it is also necessary to interpret and account for educational experiences, particularly the amount and quality of content coverage and the availability of instructional resources.

Language Development Issues

Knowledge about second language acquisition is essential in the development of a system of linguistic accommodations directed at varying levels of English proficiency. Scholars interested in language learning generally agree

that the developmental processes of primary language (L1) and second language (L2) acquisition are interrelated. The debate lies in the extent to which one language influences the other and in the methods used to measure their interrelatedness (Ascher, 1991).

Many educators and policy makers believe that children's control over the surface features of English (their fluency in conversational English) is a sufficient indicator of all aspects of English proficiency (Cummins, 1980, 1994). Once the child exhibits mastery over conversational English, efforts are made to place the child in an English-only classroom. This misunderstanding of English proficiency has had a great impact on the organization of bilingual educational programs in the U.S. In one major school district, for example, a student is placed in an English-only program if she passes an oral English test. Exiting the bilingual program is also primarily based on an oral proficiency measure.

This practice is problematic in many ways. First, it reduces English proficiency to the oral command of the language. Oral proficiency is necessary but not sufficient for educational achievement. Second, this practice suggests that conversational and academic language proficiency are synonymous. There is growing evidence that academic success is also tied to cognitive academic language proficiency (academic proficiency), also known as the cognitive demands of communication (e.g., Collier, 1987; Cummins, 1981; Wong-Fillmore, 1991), and has a different developmental trajectory.

The third concern is related to the development of academic language proficiency. Conversational English has been found to take only 2 to 3 years to master (Cummins, 1980; Gonzalez, 1986). Most bilingual programs transition students to English instruction in the third or fourth grade. Academic English proficiency, on the other hand, is acquired in approximately 5 to 7 years, depending on when the child enters the school system (Collier, 1987; Cummins, 1981). Placing ELLs in an instructional setting where English is the only language of instruction may not allow the students sufficient time to develop the academic language skills necessary for educational success.

Cummins (1994) argues that there are two principal reasons why there are major differences in the length of time needed to acquire conversational and academic language proficiency. In conversation, the learner utilizes contextual cues to facilitate the communication of meaning, cues that are largely absent in

most academic settings, which depend on decontextualized literacy skills and manipulation of language for successful task completion. Cummins (1994) uses an interesting example to illustrate the impact of linguistic cues:

> a cohesive device such as *however* coming at the beginning of a sentence tells the proficient reader (or listener) to expect some qualification to the immediately preceding statement. Lack of experience with or sensitivity to such linguistic cues will reduce students' ability to interpret meaning in decontextualized settings where interpersonal or non-linguistic cues are lacking. (p. 10)

Traditional assessment situations, in most instances, represent the most decontextualized contexts. Even performance assessments that attempt to provide intrinsic meaning in their assessment tasks may be decontextualized and cognitively taxing in large-scale contexts, for it is difficult for on-demand assessments to present the contextual cues found in interpersonal communicative situations. Given that most on-demand performance assessments also involve a great amount of English reading and writing, underestimates of students' content understanding are likely unless particular steps are taken to support linguistic task demands.

What factors facilitate the acquisition of academic proficiency? Based on her review of the literature, Collier (1989) proposed several generalizations about optimal age, L1 cognitive development, and L2 academic achievement. Cognitive development of L1 appears to be necessary for L2 academic language proficiency for both communicative and academic purposes regardless of age and number of hours of second language instruction. (See Collier, 1989, for review of literature.) Collier also found compelling evidence supporting the claim that negative cognitive effects in L2 acquisition are likely if L1 development is discontinued before it is complete. It is therefore reasonable to assume that L2 learning, particularly the learning of academic language, is dependent on the nature and level of L1 development (Saville-Troike, 1991).

Given the interrelatedness of L1 and L2 proficiency in academic contexts, it may also be necessary to determine the extent to which L1 proficiency impacts achievement on complex performance assessments. Information from such analysis may be useful when considering accommodation strategies for ELLs.

Accommodation Strategies

The term *accommodation* denotes an adjustment to be made to the testing situation to allow the test taker to display more adequately his or her competency. The National Center for Educational Outcomes (NCEO, 1995) categorized accommodations offered for students with disabilities into four categories: (a) timing, (b) setting, (c) response format, and (d) presentation. Common accommodations for students with special needs involve modifying the testing conditions, so that more time, for example, may be available for students who have difficulty in reading. In a survey of testing practices, all but 7 states report that they provide accommodations for ELLs, and 36 states permitted ELL exclusion if it is judged that the student has insufficient English competence to respond to the test (Bond, Council of Chief State School Officers, & North Central Regional Educational Laboratory, 1996). Seventeen states reported language accommodations of the following type: separate scheduling and testing settings, multiple or extended testing opportunities, and small-group administration. Linguistic supports were reported by some states: simplification of directions (11 states), audiotaped instructions or questions (9), use of dictionaries (9), audiotaped responses (4), other languages (4), and an alternative test (3) (Council of Chief State School Officers & North Central Regional Educational Laboratory, 1996).

Expanding upon the NCEO and the Chief State School Officer (CCSSO)/ North Central Regional Educational Laboratory (NCREL) categories, let us consider a continuum of accommodation anchored at one end by students responding unaided to English language assessments and at the other by students excluded from the testing situation (see Table 1).

The list in Table 1 pertains to tasks that are intended to be the same for all students, that is, a written response to a provided poem, a set of mathematical word problems using one unknown, or the completion of a science experiment according to specified procedures. Using these accommodations creates a series of concerns related to validity and its important subset, fairness. To begin, there is the need to determine and logically weight the advantage provided to the student against at least two factors. First, the testing authority needs to consider the cost of including the student by involving multiple accommodations against the risk of excluding the child. In order to make a better decision, at least two

Table 1

Continuum of Accommodations for English Language Learners

Standard examination:	Comparable materials, setting, conditions, instructions, and response formats
Modified setting:	Reduced distractions by administering test alone, in carrels, or in small groups
Modified time:	Extended testing period
Modified directions:	Simultaneous oral directions in English
	Simplified English directions
	Non-verbal supports in directions (pictures and schematics)
	Translated directions into l1; bilingual directions provided in writing and/or orally
	Demonstration of sample item
Multiple trials:	Permitting more than one administration opportunity
Adapted test materials:	Dictionaries or special-purpose glossaries; simplification or partial translation (key words or concepts); fully translated stimulus materials, including texts, problems; culturally adapted materials, using comparable content
Response options:	Oral response in English; written response in English; written Spanish response; oral Spanish response; supplement to written response by oral explanation
Modifications in scoring and interpretation combinations:	Special scoring procedures, rubrics, adjustments; special identification of students receiving accommodations incorporating one or more adjustment or accommodation

pieces of information should be considered. One is the degree of proficiency in English, in order to determine whether relatively minor accommodations will suffice. There is no reason to provide a full range of accommodations that take additional time and that separate students from their peers if only modest assistance is needed. In addition, the students' L1 proficiency should be known. One of the most needed diagnostic devices is a tool to establish L1 proficiency for ELLs. There is no sense in administering fully translated test materials if the student has low levels of L1 literacy.

Second, the use of combinations of accommodations raises the question of comparability of results for a given test taker. If test results are normed, the likelihood is small that results will be directly interpretable if multiple accommodations are used for a given student. In the case of norms available for translated or bilingual versions, it is incumbent upon the testing authorities to

determine that the norming group is appropriate for interpreting the results of the tested students. Norming for these tests may be based on calculations from tests administered in English or on tests administered in other countries (e.g., Mexico or Spain). This assumes that the population of students in those countries is similar to the population of students in the United States and that the contexts are also comparable (Figueroa, 1990; Valdes & Figueroa, 1995). Even when norming data for U.S. populations is available for students with Spanish language backgrounds, they are less likely to be at hand for students from many other language backgrounds. Related to interpreting performance with respect to norms is the more global issue of validity. Even when norms are not used, there is a question of validity—that is, whether it can be demonstrated that the same construct has been measured by accommodated forms, and whether it can be shown that factors irrelevant to the construct are not intruding in the estimates of student competence. No easy solutions exist to these problems, particularly for performance tasks where explanation and elaboration are intrinsic parts of performance. Similarly, validity inferences may be compromised for tasks that in themselves are based in relatively sophisticated language skills, whether they involve encoding (i.e., reading comprehension; analysis of historical, literary, or scientific text) or the construction and expression of competence in writing and speaking. Validity interpretations also become particularly troublesome if modifications in scoring rubrics are made. The difficulty of demonstrating that the measures are assessing the same construct in such cases is formidable and unlikely to be realized on the schedule needed for regular assessment of progress. Finally, there is the question of credibility of results and the perception of fairness. Students who are not provided accommodations (such as extended time, dictionaries, or oral directions) may perceive themselves at a disadvantage. Students who are provided accommodations may not wish to have their records marked to indicate that special adjustments in testing materials or conditions were provided.

Research on Accommodations

Published studies that examine the impact of linguistic factors on performance assessment scores are few. This work is concentrated in classroom-level assessments (e.g., Pierce & O'Malley, 1992; Rosebery, Warren, & Conant, 1992) and therefore provides strategies that are not feasible for large-scale testing. Nevertheless, the demand for alternative assessments has grown among

language and content educators who want more accurate measures of their students' knowledge (LaCelle-Peterson & Rivera, 1994; Short, 1993). Educators who want to measure students' content knowledge are faced with the difficulty of disentangling linguistic factors from content knowledge as well as dealing with the problem of extensive variation in language proficiency.

One study that directly examined the effects of linguistic complexity of test items on ELLs' test performance was conducted by CRESST on NAEP math items (Abedi, 1994). Although NAEP items are not performance assessments, this study is noteworthy in that it attempted to identify and modify specific linguistic features of test items that may contribute to content-irrelevant difficulties ELLs often encounter. In the first phase of this study, linguistic complexity of the 1992 NAEP items was examined. ELLs' test scores were lower than the scores of other students, and the differences were greater for those items identified as being more linguistically complex. In this part of the study, students were also asked to complete background questionnaires. Abedi found that students who reported that they received English as a Second Language (ESL) instruction had considerably lower math scores than other students. No significant differences were found between this group and other students on other background variables such as socio-economic status. Although the exact nature of the impact of language on performance was indeterminate, these findings provide evidence for the assertion that the language of test items may contribute to underestimates of ELLs' content knowledge.

In the second phase of this study, one linguistic accommodation was tested, the simplification of linguistic features contained in the items. Students who enrolled in average- and low-level mathematics courses and who received the standard version of the math assessment scored significantly lower than those who received the linguistically modified version. Differences were found across categories of ethnicities; however, when type of mathematics class was statistically controlled, these differences were not significant. A hierarchical linear modeling analysis was then conducted; interesting patterns emerged. Language-related variables were shown to be more effective than the model with the original item score as the outcome variable. These patterns, however, did not reach statistical significance.

The lack of significance was attributed to three limitations. The first limitation is related to the limited number of NAEP items available for the

study. Subscale analyses could not be conducted in this field test because only 10 items were used and the *p* values indicated that some of these items were either too difficult or too easy. Moreover, they did not obtain a good range of the types of linguistically complex items and the range of linguistic features that would be more desirable for this type of study. The third limitation was the variation in how students were classified as ELLs. School district information about students' language background was also incomplete, outdated, or invalid in many instances.

Beyond Accommodation

It is possible that manipulating surface features of tests and making minor modifications of test conditions are the best that can be accomplished with the present state of the art. Nonetheless, we believe that for these approaches to be successful, substantial problems of interpretation, practicality, and credibility need to be overcome.

As an alternative, we propose an approach that moves beyond accommodations to a consideration of theories related to knowledge representation and cognitive structure as a strategy to address the task of assessing students from varied language backgrounds. Knowledge representation theories in general capture the related propositions that knowledge in memory is organized into structures and that these structures influence the selection, encoding, interpretation, and use of novel information (Anderson, 1984; Glaser, 1984; Mayer, 1984; Rumelhart, 1980; Seel, 1995). Jonassen, Beissner, and Yacci (1993), for example, argue that human memory is organized into schemas, or networks of interrelated concepts, that contain and invoke not only related conceptual and factual knowledge about a concept but also information about situations where such knowledge can be used.

The development and wide acceptance of various types of knowledge representation theories suggest an alternative to essay construction as a measure of conceptual understanding; that is, the use of concept maps as measures of cognitive structure. Concept maps emerged from research on structural knowledge (Jonassen et al., 1993; Rumelhart & Norman, 1988; Rumelhart & Ortony, 1977) and thus purport to represent a student's knowledge structure in a given domain. A concept map is a graph of a given content domain (or subset

thereof) consisting of nodes that represent important concepts or ideas and labeled links that depict the relationship between a pair of concepts (nodes).

Concept maps are designed to be less discourse-dependent than essays. Although studies that directly examine this assertion have not been conducted, there is some evidence indicating that some students' knowledge may be underrepresented by their performance on multiple-choice tests or essays (Baker, Niemi, Gearhart, & Herman, 1990). This situation is particularly true for ELLs.

Early research with concept maps was originally designed to elicit learning strategies in students (e.g., Anderson, 1979; Anderson & Armbruster, 1981; Dansereau & Holley, 1982) and has since then demonstrated its usefulness in instructional and assessment settings (e.g., Herl, Niemi, & Baker, 1996; Horton et al., 1993; Lambiotte & Dansereau, 1991; Novak 1995; Ruiz-Primo, Schultz, & Shavelson, 1996).

In recent years, there has been a growing interest in the use of concept maps in assessment settings. Lomask, Baron, Greig, and Harrison (1992), for example, used teacher-constructed maps from student essays and judged them on the basis of their match with expert maps. Training teachers to draw concept maps from student essays has a practical advantage for large-scale contexts in that students do not have to be trained to draw their own maps. However, this practice raises cognitive-theoretical and methodological issues concerning, for instance, the degree to which teacher maps reflect the structural representations of students (Shavelson, Lang, & Lewin, 1994).

An alternative and more common procedure for map construction are paper-and-pencil tasks that ask students to construct their own maps. Herl et al. (1996), for example, instructed eleventh-grade U.S. history students enrolled in general and Advanced Placement courses to construct concepts maps on the Great Depression and compared them to concept maps generated by experts in the field. Herl et al. found that experts performed higher than either of the two student groups and had higher structural scores (a measure of the similarity between clusters of concepts in a semantic network) than either group. A similar pattern was found between Advanced Placement students and students enrolled in general history courses.

When interpreting the scores of concept maps, issues related to comparability should be considered. Comparability of administration conditions

refers to the additional demands placed on the test administrator that emerge from the performance task (e.g., group work, experimentation, etc.). Concept mapping represents a response modality and a set of cognitive demands that reduce the dependence upon complex discourse. However, at the outset, concept mapping tasks may pose greater variability in administration because this kind of task includes an instructional lesson, and more questions are likely to arise due to the novelty of the task. How the test administrator handles student questions is largely left to the individual, thereby increasing variability in testing administration.

The key test will be whether concept mapping introduces or restricts construct-irrelevant variance. If test takers do not possess the necessary ancillary or enabling skill requirements of a given performance task, a test can be said to be biased if particular groups of "examinees are deficient in a test's ancillary abilities" (Haertel, & Linn, 1996, p. 63). Ancillary abilities refers to the set of skills or abilities required for successful completion of a task that are not explicitly part of what is to be assessed. Among these skills are students' understanding that it is important to show their best work; their willingness to do so; ability to understand the task requirements; and their mastery of the communication skills necessary to produce measurable responses (Haertel, & Linn, 1996; Linn et al., 1991).

Ability to understand the task requirements and mastery of communication skills are critical areas when testing ELLs in English. To the extent that concept maps are dependent on English, scores for ELLs may not be comparable to scores for native speakers of English because the relative contributions of distinct ancillary abilities to the construct measured may depend on the language background of the test taker (Haertel, & Linn, 1996). We believe that there are ways to provide stimulus materials in forms that maintain their cognitive complexity and avoid bias. Nonetheless, our hunches are vastly insufficient to recommend an approach. Research is underway to determine the extent to which language background characteristics impact scores on concept maps and other forms of performance assessments, to assess the feasibility of the application of a concept mapping approach to ELLs, and to determine whether valid inferences of ELLs' content understanding can be drawn from their scores on performances assessments.

References

Abedi, J. (1994). *Language background as a variable in NAEP mathematics performance. NAEP TRP Task 3d: Language background study*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Anderson, T. H. (1979). Study skills and learning strategies. In H. F. O'Neil. Jr., & C. D. Spielberger (Eds.), *Cognitive and affective learning strategies*. New York: Academic Press.

Anderson, J. R. (1984). Spreading activation. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower* (pp. 61-90). San Francisco: W. H. Freeman.

Anderson, T. H., & Armbruster, B. B. (1981). Studying. In P. D. Pearson (Ed.), *Handbook on reading research* (pp. 657-680). New York: Longman.

Ascher, C. (1991). Testing bilingual students: Do we speak the same language? *PTA Today*, March.

Baker, E. L., Niemi, D., Gearhart, M., & Herman, J. (1990, April). *Validating a hypermedia model of knowledge representation*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Baker, E. L., & O'Neil, Jr., H. F. (1996). Performance assessment and equity: A view from the USA. [CD-ROM]. *CRESST: 5 Years of Research.* Los Angeles: University of California, Center for the Research on Evaluation, Standards, and Student Testing.

Bond, L. A., Council of Chief State School Officers, and North Central Regional Educational Laboratory. (1996). *Statewide assessment of students with disabilities*. Oak Brook, IL: North Central Regional Educational Laboratory.

Cheung, O., Clements, B., & Miu, Y. C. (1994). *The feasibility of collecting comparable national statistics about students with LEP*. Washington, DC: National Center for Education Statistics.

Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, *21*, 617-641.

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly*, *23*, 509-532.

Council of Chief State School Officers and North Central Regional Educational Laboratory. (1996). *1996 state student assessment programs database*. Oak Brook, IL: North Central Regional Educational Laboratory.

Cummins, J. (1980). Entry and exit fallacy in bilingual education. *NABE Journal*, *4*(3), 25-59.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, *2*, 132-149.

Cummins, J. (1994). Primary language instruction and the education of language minority students. In C. F. Leyba (Ed.), *Schooling and language minority students: A theoretical approach* (2nd ed.). Los Angeles: California State University, Evaluation, Dissemination and Assessment Center.

Dansereau, D. F., & Holley, C. D. (1982). Development and evaluation of a text mapping strategy. In A. Flammer & W. Kintsch (Eds.), *Discourse processing*. Amsterdam: North Holland.

Elementary and Secondary Education Act of 1965., 20 U.S.C. §§ 236 *et seq*., 821 *et seq*.

Exec. Order No. 12866, 34 C.F.R. 200, 201, 203, 205, and 212 (1995).

Figueroa, R. A. (1990). Assessment of linguistic minority group children. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 671-696). New York: Guilford Press.

Fleischman, H. L., & Hopstock, P. J. (1993). *Descriptive study of services to limited English proficient students, Volume 1: Summary of findings and conclusions*. Arlington, VA: Development Associates, Inc.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, *30*, 93-104.

Gonzalez, L. A. (1986). *The effects of first language education on the second language and academic achievement of Mexican immigrant elementary school children in the United States*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59-78). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Hafner, A. L., & Ulanoff, S. H. (1994). Validity issues and concerns for assessing English learners: One district's approach. *Education-and-Urban-Society*, *26*, 367-389.

Herl, H. E., Niemi, D., & Baker, E. L. (1996). Construct validation of an approach to modeling cognitive structure of experts' and novices' U.S. history knowledge. *Journal of Educational Research*, *89*, 206-218.

Horton, K. J., McConney, A. A., Gallo, M., Woods, A. L., Senn, G. J., & Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. *Science Education*, *77*(1), 95-111.

Hopstock, P. J., & Bucaro, B. J. (1993). *A review and analysis of estimates of the LEP student population*. Arlington, VA: Development Associates, Special Issues Analysis Center.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment: Policies for English language learners. *Harvard Educational Review*, *64*, 55-75.

Lambiotte, J. G., & Dansereau, D. F. (1991). Effects of knowledge maps and prior knowledge on recall of science lecture content. *Journal of Experimental Education*, *60*, 189-201.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Lomask, M., Baron, J., Greig, J., & Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. Symposium presented at the annual meeting of the National Association of Research in Science Teaching, Cambridge, MA.

Mayer, R. E. (1984). Aids to prose comprehension. *Educational Psychologist*, *19*(1), 30-42.

National Center on Educational Outcomes. (1995). *Compilation of states' guidelines for accommodations in assessments for students with disabilities* (Synthesis Report 18). Minneapolis: University of Minnesota, College of Education, National Center on Educational Outcomes.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: RAND.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments*. Washington, DC: U.S. Government Printing Office.

Pierce, L. V., & O'Malley, J. M. (1992). Performance and portfolio assessment for language minority students. *NCBE Program Information Guide for Bilingual Education Series 9*. Washington, DC: National Clearinghouse for Bilingual Education.

Rosebery, A. S., Warren, B., & Conant, F. R. (1992). Appropriating scientific discourse: Finding from language minority classrooms. *The Journal of the Learning Sciences*, 2(1), 61-94.

Ruiz-Primo, M., Schultz, S. E., & Shavelson, R. J. (1996). *Concept map-based assessments in science: Two exploratory studies* (CSE Tech. Rep. No. 436). Los Angeles: University of California, Center for the Research on Evaluation, Standards, and Student Testing.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., & Norman, D. A. (1988). Representation in memory. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. Duncan Luce (Eds.), *Stevens' handbook of experimental psychology, Vol. 1: Perception and motivation; Vol. 2: Learning and cognition* (2nd ed., pp. 511-587). New York: John Wiley & Sons.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of* knowledge (pp. 97-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Saville-Troike, M. (1991). Teaching and testing for academic achievement: the role of language development. *NCBE Focus: Occasional Papers in Bilingual Education*, *4*. Washington, DC: National Clearinghouse for Bilingual Education.

Seel, N. M. (1995). Mental models, knowledge transfer, and teaching strategies. *Journal of Structural Learning*, *12*, 197-213.

Shavelson, R. J., Lang, H., & Lewin, B. (1994). *On concept maps as potential "authentic" assessments in science* (CSE Tech. Rep. No. 388). Los Angeles:

University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly*, *27*, 627-656.

Stanovich, K. E. (1991). Discrepancy definitions of reading disability: Has intelligence led us astray? *Reading Research Quarterly*, *26*(1), 7-29.

U.S. Bureau of the Census. (1990). *The foreign born population in the United States*. Washington, DC: U.S. Government Printing Office.

Valdes, G., & Figueroa, R. A. (1995). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Wong-Fillmore, L. (1991). When learning a second language means losing the first. *Early Childhood Research Quarterly*, *6*, 323-346.

Zehler, A. M., Hopstock, P. J., Fleischman, H. L., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students* (Task Order Rep. No. D070). Arlington, VA: Development Associates, Inc., Special Issues Analysis Center.