

Language Testing

<http://ltj.sagepub.com>

Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas

Frances A. Butler and Robin Stevens

Language Testing 2001; 18; 409

DOI: 10.1177/026553220101800406

The online version of this article can be found at:
<http://ltj.sagepub.com/cgi/content/abstract/18/4/409>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Language Testing* can be found at:

Email Alerts: <http://ltj.sagepub.com/cgi/alerts>

Subscriptions: <http://ltj.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations (this article cites 4 articles hosted on the SAGE Journals Online and HighWire Press platforms):

<http://ltj.sagepub.com/cgi/content/refs/18/4/409>

Standardized assessment of the content knowledge of English language learners K–12: current trends and old dilemmas

Frances A. Butler and Robin Stevens *University of California, Los Angeles*

Within the context of accountability for US schools, standardized achievement tests are being used for increasingly 'high stakes' decisions for all students including those for whom English is a second language, even when their English language skills are not adequate for the task. This article discusses approaches to the standardized assessment of content knowledge for English language learners (ELLs),¹ including testing in the student's first language, the use of test accommodations, and measuring growth in English as an alternative for accountability until student control of English is sufficient to assure validity of test scores. Limitations of current research on the use of standardized content assessments with ELLs are presented and alternative approaches suggested.

I Overview and problem

One of the most critical challenges facing educators today in the USA is the ever-increasing enrolment of students in all grades K–12 for whom English is a second language. During the 1996–97 school year, approximately 3.4 million students in the USA were designated as limited English proficient (LEP), a 7% increase from the year before (Macias, 1998); by the 1999–2000 school year that number had increased to approximately 4.2 million students (National Clearinghouse for Bilingual Education, 1999). Of the students who are classified as LEP, approximately 80% are enrolled in any one of a range of specially designed instructional programs, including transitional bilingual programs, sheltered content classes and pull-out English classes (Kopriva, 2000). In some states, children are simply sent into

Address for correspondence: Frances A. Butler, UCLA CSE/CRESST, 300 Charles E. Young Drive North, GSE&IS Bldg., 3rd Fl/Mailbox 951522, Los Angeles, CA 90095-1522, USA; email: butler@cse.ucla.edu

¹The expression 'English language learners' is used henceforth in this article to refer to students who are acquiring English as a second language; the term includes a range of English proficiency.

the mainstream classroom with their native English-speaking peers, where teachers must attempt to meet their specialized needs in addition to the needs of the native speakers. Depending on a multitude of factors – such as age of arrival and prior formal education – many children take a number of years to master the language of education of the USA, academic English, before they can participate and learn equally alongside native English speakers (Cummins, 1981; Collier, 1989; Hakuta *et al.* 2000).

For all students, an integral part of the American education system is participating in the various levels of assessment, from teacher-designed classroom assessments to national assessment programs such as the National Assessment of Educational Progress (NAEP). The ‘Improving America’s Schools Act of 1994’ states that all students, including limited English proficient (LEP) students, should be included in state and district assessment programs and that comparable information about student progress should be obtained for students who are excluded for educational or psychometric reasons. Thus, the stakes have been raised to include all students in all assessments. At the same time, however, standardized testing programs at the district and state levels are becoming increasingly driven by accountability, making tests such as the Stanford Achievement Test Series, ninth edition (Harcourt Brace Educational Measurement, 1996), currently mandated for use in the state of California, and the Iowa Tests of Basic Skills (The University of Iowa & The Riverside Publishing Company, 1994) more ‘high stakes’ than in the past. In fact, the trend now in many states, such as California, is to use statewide assessment results as a part of a reward–punishment system, rewarding schools that have high test scores with additional funds (Helfand, 2000).

While there is a definite need to show the public what students are learning in schools and to hold schools accountable for the education of students, the use of the same standardized content assessments with all groups of students is problematic and may not be the best approach to accountability. Test results for all students are often widely misinterpreted or misused. For example, states do not always sequence their curriculum in exactly the same way, and although test publishers attempt to take these differences into account, test topics do not always align with the curriculum students have been taught. This ‘opportunity to learn’ (OTL) issue is a critical one in that students may be tested on material that they have not yet been exposed to (Stevens *et al.*, 2000).

Furthermore, the test scores are not always reliable and valid for all of the purposes for which they are used or for all the students to whom they are given. This is particularly true for students who speak English as a second language. Commercially developed large-scale

content assessments were produced and intended for students who are native speakers of English or highly proficient non-native speakers. Indeed, very few English language learners (ELLs) are included in the norming samples of the most widely used standardized tests (Davidson, 1994). At the present time, however, across states, large numbers of ELLs are expected to take standardized assessments with their native English-speaking peers, regardless of how long they have lived in the USA or how well they have mastered English (Holmes and Duron, 2000). For students whose English proficiency is still developing, the tests often pose significant reading challenges that interfere with the assessment of the content they have learned, making their test scores invalid as indicators of content knowledge or achievement.

Many politicians and the public at large do not fully understand the impact of this validity issue; yet test results are being used for high-stakes decisions about school programs despite their potential inappropriateness. Contrary to what some of the test results indicate, '[ELLs] . . . may understand and know much more than they are allowed to demonstrate under the confines of large-scale tests designed for mainstream use with students sharing common cultural experiences' (Kopriva, 2000: 5). The challenge, then, is to determine how best to tap what an ELL student knows in specific content areas.

In this article, the assessment of ELL content knowledge is framed against the backdrop of accountability. The question of inclusion is critical. Current approaches to inclusion are discussed, and unresolved research issues related to the assessment of ELL content knowledge are raised.

II Approaches to standardized assessment of content knowledge

In the past, the answer to the standardized testing dilemma was to simply exclude students from testing who had been in the USA for a limited number of years or who were deemed not to have the English proficiency needed to participate in the testing process. While there is some logic to this approach, there are also problems associated with it. Probably the single biggest concern about this approach is the lack of representation for these students. Excluding them from assessment is in essence the same as excluding them from educational programs, in that assessment is used as a tool to help to understand and identify student needs and progress. If ELLs are not tested, information on their achievement is effectively absent from the decision-making processes that affect their academic futures. This position is untenable.

With growing national emphasis on accountability, educators have

attempted a variety of approaches for guaranteeing the inclusion of ELLs in assessment processes. One approach involves testing subject matter in the student's first language. A second approach allows the use of test accommodations to make the content more accessible to the test-taker. A third approach measures and reports growth in English until a student has the requisite English language skills for taking a standardized achievement test in English. A discussion of each of these approaches follows.

1 Testing in the first language

One method of including students in testing programs has been the use of content assessments in the student's first language. California, Minnesota, New Jersey and New York offer tests in some first languages as part of their accountability systems (for the specific tests, see Holmes and Duran, 2000). This approach has recently been classified as a type of test accommodation.

Generally, however, there are two approaches to developing first language assessments for ELLs:

- 1) developing an assessment parallel to the English assessment where specifications help to assure that the same concepts are being tested through the use of authentic materials in each language; and
- 2) translating the English assessment into the first language(s) (Butler and Stevens, 1997a).

While the first approach is preferred, both approaches are problematic. Parallel development is expensive. It requires comparable academic materials in all of the languages. Translation, the most commonly used method, assumes an equivalency of linguistic features, such as word meaning and syntax, that does not always exist.

Assuming parallel tests are available, the issue of student literacy and control of academic usage of the non-English language must also be considered. If students are, or have been, enrolled in dual bilingual programs that stress the academic use of both languages, or if they have received significant content instruction in the language other than English outside the USA, then they may be able to demonstrate what they know about the content area on a test in their native language. If, however, content instruction has been in English with little prior content instruction in the first language, then testing in the first language would yield misleading results by suggesting that students do not have specific content knowledge, when in fact they may be able to demonstrate that knowledge in the language of instruction.

Another important consideration remains with regard to testing in

the first language. That is, since appropriate assessments are generally not available for all of the languages represented in schools in the USA, to offer standardized content assessments in some languages and not in others in itself creates another type of inequity. Thus the use of large-scale first language assessments should be approached with caution and other means for assuring inclusion carefully examined.

2 Accommodations

A growing body of research is exploring the use of test accommodations with ELLs on content assessments (Butler and Stevens, 1997a; Abedi *et al.*, 1998; Castellon-Wellington, 1999; Abedi *et al.* 2000b; Aguirre-Munoz, 2000). This approach with ELLs follows from the use of test accommodations for students with disabilities. Accommodations theoretically offer ELLs support in the test situation with the goal of facilitating their ability in order to demonstrate what they know in content areas.

Accommodations for English language learners on large-scale content assessments refers to the support provided students for a given testing event . . . to help [them] access the content in English and better demonstrate what they know. (Butler and Stevens, 1997a: 5)

Test accommodations fall into two categories: modifications to the test itself and modification to the test procedure. Table 1 provides types of accommodations that fall into each category. Test accommodations have been used by school districts and states for a number of years in an attempt to level the playing field. A recent survey by Rivera *et al.* (2000) documents the use of accommodations in the USA. A problem, however, with the use of test accommodations is

Table 1 Two categories of accommodations for English language learners

Modifications of the test	Modifications of the test procedure
<ul style="list-style-type: none"> ● assessment in the native language; ● text change in vocabulary; ● modification of linguistic complexity; ● addition of visual supports; ● use of glossaries in native language; ● use of English Glossary; ● linguistic modification of test directions; ● additional example items/tasks. 	<ul style="list-style-type: none"> ● extra assessment time; ● breaks during testing; ● administration in several sessions; ● oral directions in the native language; ● small-group administration; ● separate-room administration; ● use of dictionaries; ● reading aloud of questions in English; ● answers written directly in test booklet; ● directions read aloud or explained.

Source: Butler and Stevens, 1997a: 6

that they have been implemented with good intentions but without an empirical base that demonstrates whether using them makes any significant difference in the performance of ELLs on large-scale content tests (Spicuzza *et al.*, 1996; Butler and Stevens, 1997a). Also, there is very little uniformity in how they are used or in who makes the decisions about their use (Abedi and Leon, 1999).

It has been argued that:

decisions about the type(s) of accommodations to use with assessments must be made systematically on the basis of both (a) the content and nature of the assessment and (b) the characteristics of the subpopulation(s) being tested. (Butler and Stevens, 1997a: 6).

An accommodation that is appropriate for ELLs at one level of English language proficiency, for example, may not be appropriate or necessary for ELLs at another level. In other words, students may benefit differentially from accommodation support depending on their levels of English language proficiency. Guidelines are clearly needed to assist decision makers in their selection or development of accommodations appropriate to their specific situations.

Research is beginning to shed light on the potential impact of test accommodations with ELLs on test scores. A study by Abedi *et al.* (1998) showed that although ELL scores increased slightly when given accommodations – such as extra time, glossaries or tests with modified English – the gains were not statistically significant. Interestingly, native speakers in the study also had slightly increased scores when they received accommodations. Disaggregating ELL performance data according to English language proficiency classifications was not possible due to inconsistencies of classifications across schools and districts, so the impact of differential English language proficiency levels could not be evaluated.

In a study that investigated whether student scores improved if given the choice of two accommodations, extra time or reading the test items and instructions aloud, Castellon-Wellington (1999) found that student performance was not significantly improved with either of the accommodations, even when students received the accommodation of their choice.

Aguirre-Munoz (2000), in a study looking at the use of accommodations with performance assessments, found that the strength of the relationship between English proficiency and performance on complex history-explanation tasks differed depending on the type of accommodation a student received. The type of accommodations used were Spanish-only versions of the test materials and modified English versions. ELLs with the lowest English proficiency benefited most from the Spanish-only accommodation, while students with intermediate proficiency in English benefited more from the modified

English version. This finding supports the need for differential use of accommodations based on English language proficiency.

A pilot study by Abedi *et al.* (2000b) used three types of accommodations – English dictionary, bilingual dictionary, and a linguistically modified test version – with NAEP science items at the fourth and eighth grades (ages 10 and 14, respectively). The results show differences in ELL² performance across the accommodations at the two grades. At the fourth grade there was a significant difference between the accommodated and unaccommodated conditions on the science items for the English dictionary and the bilingual dictionary. There was no significant difference in performance between the group that received the linguistically modified items and the group that received the original items without accommodation. At the eighth grade the results were the opposite, with a significant difference in the performance of the ELLs between the accommodated and unaccommodated conditions with the linguistically modified test accommodation. The performance differences between the accommodated and unaccommodated conditions with dictionaries were not significant. Accommodations did not seem to make a difference in non-ELL performance in this study.

The mixed results from these studies, along with issues regarding the feasibility of systematically implementing accommodations with ELLs, suggest that other options should be explored even as research on accommodations continues. Although some types of accommodations show promise for some groups of ELLs, the use of the same test accommodations with all ELLs is not likely to be an effective solution because of differential student needs based on level of English language proficiency and other variables. Therefore, until there is a clearer picture of which ELLs benefit from which accommodations, the widespread use of accommodations must be undertaken with care.

3 Measuring growth in English

Many US states are examining the option of providing a separate measure of growth or progress in English as an alternative approach to inclusion or even as a supplemental accountability measure to help monitor the success of instructional programs. For example, California has developed an English Language Development (ELD) test that is aligned to state ELD standards. The new test will be administered

²In this study, the term ELLs is used in the narrower sense to signify students with limited English proficiency.

annually as a supplementary accountability measure and, upon intake, to help identify students for services in ELD programs.

Other states are obtaining measures of growth in English as an alternative accountability measure. Illinois is measuring student achievement in English when English proficiency is not adequate for participation in large-scale assessments of content knowledge. ELLs from Grades 3 to 11 (aged 9–17), who have been in the specialized state English programs between six months and three years and are exempted from the state content assessments, must take the Illinois Measure of Annual Growth in English (IMAGE) (Illinois State Board of Education, 1996) instead. This allows the state to report on gains in English for those students who are unable to take the content tests, thus providing a measure of accountability for them. After three years, the students are then required to take the state content assessments. In spring 2000, Texas began collecting baseline data on a similar assessment for students who are identified as LEP (Texas Education Agency, 2000).

These developments seem promising since, in this way, students whose English language proficiency is limited can still be included in the accountability picture and better-informed decisions can be made to improve the programs that serve these students. Additionally, providing ELLs the opportunity to participate in large-scale language assessments may help prepare them for taking large-scale content assessments later by giving them practice with standardized test formats.

4 Discussion

The approaches to inclusion discussed above demonstrate attempts to respond to state and national demands for accountability. The first two approaches are especially problematic. Testing in the student's first language is impractical on a large scale because of the great number of first languages represented in US schools and the difficulties in producing parallel tests. The use of other test accommodations, while offering more promise as a means of support for ELLs, is rife with difficulties as well. Logistical issues must be resolved for each accommodation and for the use of multiple accommodations, according to student needs in a specific testing situation. In addition, future research must help to identify language proficiency benchmarks associated with specific types of accommodations, so that guidelines can be prepared for their use under specified conditions. Finally, measuring and reporting growth in English until a student has been in state programs long enough to make gains in both English and content is a potentially attractive approach for assuring the participation of all

ELLs in the accountability process. This approach, however, is dependent on having effective state programs in place for ELLs. Further, for national testing in this vein, research on the alignment among state programs for ELLs is necessary to measure national gains in English proficiency. Indeed, research is sorely needed to help bridge the gap between the assessment of English language proficiency and content knowledge for ELLs. The discussion now turns to a focus on research.

III Rethinking research

Within the current educational context in the USA, immediate questions for educators are:

- 1) When is it appropriate to give standardized content assessments to ELLs? That is, when are the inferences made about the performance of ELLs on standardized content assessments valid?
- 2) Until it is appropriate to give these assessments to ELLs, how do we provide accountability and assure equity?

Unfortunately, research directed at these issues is falling short. First, there is a lack of uniformity in defining subgroups of ELLs. While the literature is beginning to acknowledge the need for focusing on the language variability (both first and second) among ELLs, a systematic approach for doing so has not emerged. Districts do attempt to articulate language differences by testing ELL students on intake and assigning designations such as fluent English proficient (FEP) and limited English proficient (LEP). A variety of language proficiency tests are used, some of which are commercially available, such as the Language Assessment Scales (Duncan & De Avila, 1988) and IDEA[®] Proficiency Tests (IPT, 1991). Others are of their own making. In part because different tests are used, the designations are not uniform across districts. Thus, the data on ELLs across districts and states are difficult to interpret. What we know in general terms is that students classified as LEP perform less well on standardized content assessments than non-LEP students (Abedi and Leon, 1999; Abedi *et al.*, 2000c; Butler and Castellon-Wellington, 2000). For purposes of intervention and in order to help assure test validity for these students and for research, this information is not specific enough to be useful.

Educators, then, remain in need of a means for operationalizing labels such as LEP, FEP, ELL and 'bilingual' in a uniform way, so that the terms have consistent meaning linked to language proficiency across all educational environments. Inconsistency in the use of the terms makes generalizability extremely difficult, even when the

broadest groupings (e.g., ELL as opposed to non-ELL groups) are used for analyses.

Secondly, currently available language proficiency tests are not tapping the more academic language of the classroom and of content assessments. They are also failing to differentiate adequately among ELLs at the upper end of the proficiency continuum. These limitations restrict their overall usefulness in research that looks at ELL performance on content assessments. Language proficiency tests are needed that match the language demands of the content assessments. Such tests would also provide a mechanism for defining subgroups of ELLs according to their language proficiency. Work by Bailey (2000) and Stevens *et al.* (2000) is just beginning systematically to characterize and analyse the differences between the language on the two types of tests.

Since the emphasis on large-scale content testing is to gauge the effectiveness of school programs, research should be directed at determining accessibility to the content for ELLs. To facilitate such a research focus, intensive work in three areas is needed. First, the development of appropriate English language assessments that tap the language of content tests is crucial. Secondly, the development of methods for determining whether all students have had the opportunity to learn the content being assessed – language, maths, social studies, etc. – is essential for identifying curriculum-accessibility issues that are impacting student performance. Finally, focused research is needed on some of the more promising recent approaches to inclusion. Specifically, if accommodations are to continue to be implemented in testing programs, additional research should further explore the utility of the various types, their effectiveness, and the impact of their use on the reliability and validity of student scores. Research should also be directed at developing measures of growth in English, since English is a major content area for ELLs. These three research issues are discussed below.

1 Academic language assessment

In assessing the content knowledge of ELLs, academic language emerges in many ways as a critical factor because it is the vehicle for acquiring the knowledge being assessed. Indeed, academic language proficiency measures are needed for a variety of reasons, including but not limited to:

- 1) making judgements about a student's English language readiness for taking standardized content tests;

- 2) for use in research on test accommodations in order to help determine the appropriacy of specific accommodations for students at various levels of proficiency; and
- 3) for use in measuring student growth in academic English.

Existing language proficiency measures are generally inadequate to these tasks due to the mismatch, mentioned above, between the type of language being assessed on such tests and the more academic language of the classroom and of the content tests. Because researchers have not been able to control adequately for language proficiency, findings to date are not specific enough to allow for effective intervention or decision making. Aguirre-Munoz (2000: 125) makes the point that:

the usefulness of . . . accommodations depends to a great extent on the ability of the school site administrators to effectively identify ELLs and classify them into precise levels of English proficiency. Therefore, valid measures of English proficiency are critical for making decisions about what accommodations to assign ELLs.

In general, then, there is a need to understand more precisely the relationship between the language of the classroom, including the language used in textbooks and supplemental materials, the language of standardized content tests and the language assessed in language proficiency tests. Stevens *et al.* (2000) found that while the language proficiency and content tests used in their study overlap in terms of high-frequency general vocabulary and grammatical structures, the language of the content test is more complex in vocabulary and sentence structure containing nonspecialized academic words, specialized content words, more embeddings and test language.

In related research on the language used in tests, Bailey (2000) found that standardized assessments contain linguistic demands that are not captured by English proficiency measures often used by school districts. She states that 'the test-taking routine is a conventional script with specific structures that need to be learned . . . ' (p. 89). Stevens *et al.* (2000) also identify 'test language' as a linguistic register commonly found in standardized content assessments, but not in measures of language proficiency. In a study of vocabulary used in test questions, Cunningham and Moore (1993) found that performance was significantly increased when 'test language' was modified.

Research is needed that will lead to descriptions of the language used in a range of academic contexts. Since academic language is developmental, research must include each grade level or grade-level cluster. Classroom observations and systematic review of materials across subjects and grade levels will help identify subject-matter-specific language and the more general cross-subject-matter language

that are both important in the acquisition and assessment of academic language. Descriptions of the language could then be used in a framework that will generate test specifications for specific testing needs. Work along these lines is underway at CRESST (Butler *et al.*, 1999). Resources such as the Cognitive Academic Language Learning Approach (Chamot and O'Malley, 1994) are contributing to the content base for developing measures of academic language.

2 Opportunity to learn

All students, native speakers and ELLs alike come to school with a variety of educational backgrounds and experience. ELLs, in particular, vary on many dimensions, including language proficiency and educational experiences in the USA and abroad (for a discussion of variables related to individual student characteristics, see Butler and Stevens, 1997a: 14–20). The variation in exposure to content alone points to inherent problems with using large-scale content assessments with ELLs. Standardized assessments are developed under the premise that all students at a given grade level receive a relatively homogenous curriculum. Figueroa and Garcia (1994: 12) state that 'standard assessments, and particularly tests, have operated under a key, robust assumption. They assume equal or comparable exposure to the content of the assessments prior to the assessment'.

The varied programs serving ELLs across the USA are in themselves barriers to the delivery of 'homogenous' experience or curriculum in school. Since states use a wide variety of instructional methods to serve ELLs, there is little continuity across the nation in the amount or quality of content that they receive. Language proficiency is often singled out as the primary reason for the poor performance of ELLs on standardized content assessments. However, opportunity to learn (OTL) may be as strong a contributor to poor test performance for some ELLs as language. Stevens *et al.* (2000) found that ELLs who performed best on a standardized content assessment, but were considered Limited Readers on a widely used language proficiency test, exhibited response patterns similar to native speakers who also took the content test, even when they selected the incorrect answers. This finding suggests that language proficiency was not the only barrier to performance since these ELLs read and responded to the items like the native speakers, whereas the lowest scoring ELLs showed an opposite random response pattern. The performance of the highest scoring ELLs indicates that they may not have received the necessary content instruction to answer the questions on the test. Thus, the use of accommodations or other specialized testing approaches aimed at mitigating language factors might not be effective, since OTL rather

than language may be the critical barrier to higher achievement for these students.

Aguirre-Munoz (2000) found similar OTL effects in her accommodations research. She states that ‘consistent with previous research, prior knowledge and the extent to which students are exposed to the content of the test had the greatest impact on test results’ (p. 121). In a related, small-scale survey of curriculum coverage in the social studies content area at the seventh grade, students received anywhere from zero to seven weeks instruction in each state-mandated content area, depending on the class in which the student was enrolled (Butler and Stevens, 1997b).

These studies highlight OTL as an important factor in the content assessment of ELLs and raise a critical accountability issue. Accountability testing is intended to measure what students have learned for the purpose of evaluating the effectiveness of the programs that serve them. If ELLs are not receiving the content that is covered on large-scale assessments, then accountability data based on these assessments are neither valid nor reliable. Classroom-based research will help to determine whether ELL students are being exposed to the content they should be learning.

Abedi *et al.* (2000a) recommend controlled, small-scale case studies that include classroom observations across content areas to look at OTL, not only in terms of the delivery of content, but also the use of academic language. The language used to deliver content, the content materials used in class and exposure to the language of standardized tests all make up different types of OTL that may have profound effects on student performance on standardized content assessments.

3 Focused research on promising approaches to inclusion

Assessing the content knowledge of ELLs is a complex undertaking because of the great diversity in the population. This diversity argues for the use of a multiple-assessment approach to determine the effectiveness of school programs, of which the use of large-scale standardized assessments may be only a part. Currently no single assessment approach has proven to be an effective solution to the problem of measuring ELL content knowledge in English when student English language ability is weak. While new methods for providing accountability data for ELLs may be desirable, standardized assessments are an integral part of the current assessment landscape in the USA. Thus, ways of promoting their effective use in the existing accountability paradigm should continue to be evaluated.

Two approaches to inclusion discussed above have potential as a

part of a program of accountability for ELLs that includes standardized assessments. First, offering test accommodations to students for whom there is evidence that accommodations mitigate the effects of language and thereby increase validity of test scores is a possible alternative. Secondly, providing a measure of growth in English for students whose English language proficiency is not yet ready for the large-scale assessments and for whom accommodations are not useful allows those students to be included.

a Test accommodations: The use of accommodations should not be viewed as a solution for testing the content knowledge of ELLs. Instead, accommodations should be regarded as interim procedures that may provide more inclusiveness for some groups of ELLs. The ultimate goal is for ELLs to become proficient enough in English to be able to benefit from instruction and take assessments without accommodations. Thus, even if the validity of the constructs being measured remain intact when accommodations are used, resources should also be spent on ensuring that classroom teachers are well trained to facilitate the acquisition of English by the ELLs in their classrooms.

Since there are indications that some students may be able to show more accurately what they know through the use of accommodations such as providing dictionaries to students along with extra time or providing a glossary and extra time (Abedi *et al.*, 2000b), further research will help to verify the early findings and help to set parameters for their use. As indicated above, the results in most studies on test accommodations have been mixed. To help clarify the picture, future research should specifically control for both language proficiency and OTL on a variety of accommodations and with different subgroups of ELLs.

b Measures of growth in English: If through an eligibility screening process students are identified who do not have the English proficiency necessary to demonstrate what they know on a standardized content assessment, alternative measures of content should be obtained. For many students who do not have the requisite English proficiency, English as a second language is a major component of their studies. As the case of Illinois has demonstrated, an alternative approach to accountability for this content area, then, is to obtain a measure of English growth until they have been in school programs long enough to progress in both English and content learning. This approach would allow ELLs to participate in the testing process and to make them part of an accountability system that holds programs responsible for their effectiveness.

However, while using a measure of English growth may currently only be feasible on a state-to-state basis (due to the lack of homogeneity across state programs), a number of steps could be taken to implement such an assessment on a national level. Specifically, programs in English as a second language would need to be aligned to the language arts content area, possibly using standards already implemented in many states, such as the TESOL Standards for ESL Students (TESOL, 1997). Then a nationwide metric for defining the language ability of ELLs with clear, objectively established parameters for ranges of performance on both language and content measures within ELL populations could be developed.

IV Conclusion

The problem facing educators in the USA of evaluating the academic progress of students who are acquiring English as a second language is without question daunting. The sheer numbers of students who are struggling in US classrooms makes the issue one of the most visible and politically sensitive at the start of the twenty-first century. The solutions will not be easy. They will require multiple types of expertise. Kopriva (2000) suggests including more language experts in the development of content assessments and the use of techniques such as bias or sensitivity reviews, language accessibility reviews, and the use of differential-item functioning procedures for their potential in helping to reduce test bias. Johnstone (2000: 132) states, in discussing the educational context in Scotland, that 'if valid and reliable assessments are to be developed, "experts" from outside primary schools are likely to be required, desirably working in collaboration with primary school teachers'.

Applied linguists, language testers, psychometricians, classroom teachers, district, state and federal officials, and other experts in the field of education should join together in serious collaboration. While some work is underway within individual disciplines and contexts, solid progress will only be made through an integration of expertises. Language testers bring to the table a unique expertise that is rarely applied in K–12 environments. The knowledge of sound test development principles, coupled with a knowledge of how to best evaluate language ability, makes language testers and other applied linguistics colleagues invaluable participants in undertaking the research recommended above. In every instance – interpreting results, recommending action and so on – knowing about language is essential.

There is an urgent need to systematically operationalize academic language across content areas by describing the language used in mainstream classrooms and on content tests, and then translating that information into academic assessments and guidelines for teachers. It is difficult to see how this work can be accomplished without the long-term collaboration of language experts and content-area experts. This article constitutes a call for language testers to be proactive in these endeavours.

Acknowledgements

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, US Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the US Department of Education.

V References

- Abedi, J., Hofstetter, C., Baker, E. and Lord, C.** 1998: NAEP math performance and test accommodations: interactions with student language background. Draft Report. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Abedi, J. and Leon, S.** 1999: *Impact of students' language background on content-based performance: analyses of extant data*. Final Deliverable to OERI, Contract No. R305B60002. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Abedi, J., Bailey, A. and Butler, F.A.** 2000a: General discussion and recommendations. In *The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives*. Final Deliverable to OERI/OBEMLA, Contract No. R305B60002. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Abedi, J., Courtney, M., Mirocha, J., Leon, S. and Goldberg, J.** 2000b: Language accommodation for large-scale assessment in science: assessing English language learners. Draft Deliverable to Office of Bilingual Education and Minority Language Affairs, OBEMLA, Contract No. R305B60002. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

- Abedi, J., Leon, S. and Mirocha, J.** 2000c: Examining ELL and Non-ELL student performance differences and their relationship to background factors: continued analyses of extant data. In *The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives*. Final Deliverable to OERI/OBEMLA, Contract No. R305B60002; pp. 3–49. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Aguirre-Munoz, Z.** 2000: The impact of language proficiency on complex performance assessments: examining linguistic accommodation strategies for English language learners. Doctoral dissertation, University of California, Los Angeles, CA.
- Bailey, A.** 2000: Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives*. Final Deliverable to OERI/OBEMLA; pp. 85–105. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A. and Castellon-Wellington, M.** 2000: Students' concurrent performance on tests of English language proficiency and academic achievement. In *The validity of administering large-scale content assessments to English language learners: an investigation from three perspectives*. Final Deliverable to OERI/OBEMLA; pp. 51–83. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A. and Stevens, R.** 1997a: Accommodation strategies for English language learners on large-scale assessments: student characteristics and other considerations. CSE Technical Report 448. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A. and Stevens, R.** 1997b: In-house summary report, history/social science topics teacher questionnaire. Summary Report. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Butler, F.A., Stevens, R. and Castellon-Wellington, M.** 1999: *Academic language proficiency task development process*. Final Deliverable to OERI, Contract No. R305B60002. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Castellon-Wellington, M.** 1999: The impact of preference for accommodations: the performance of English language learners on large-scale academic achievement tests. CSE Technical Report 524. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Chamot, A.U. and O'Malley, J.M.** 1994: *The Calla Handbook: implementing the cognitive academic language approach*. Reading, MA: Addison-Wesley.

- Collier, V.** 1989: How Long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly* 23, 509–31.
- Cummins, J.** 1981: The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: a theoretical framework*. Los Angeles: California State University, National Evaluation, Dissemination and Assessment Center.
- Cunningham, J.W.** and **Moore, D.W.** 1993: The contribution of understanding academic vocabulary to answering comprehension questions. *Journal of Reading Behavior* 25, 171–80.
- Davidson, F.** 1994: Norms appropriacy of achievement tests: Spanish-speaking children and English children's norms. *Language Testing* 11, 83–95.
- Duncan, S.E.** and **De Avila, E.A.** 1988: Language Assessment Scales (LAS) reading and writing examiner's manual, forms 1A and B, forms 2A and B, and forms 3A and B. Monterey, CA: CTB/McGraw-Hill.
- Figuroa, R.A.** and **Garcia, E.** 1994: Issues in testing students from culturally and linguistically. *Multicultural Education*, 2, 10–19.
- Hakuta, K., Butler, Y.G.** and **Witt, D.** 2000: *How long does it take English learners to attain proficiency?* Stanford University, CA, The University of California Linguistic Minority Research Institute. Policy report 2000–1.
- Harcourt Brace Educational Measurement** 1996: *Stanford Achievement Test Series*. 9th edition. San Antonio, TX: Harcourt Brace.
- Helfand, D.** 2000, March 4: Merit pay proposed for L.A. teachers. *Los Angeles Times*, pp. A1, A19.
- Holmes, D.** and **Duron, S.** 2000: *LEP Students and High-Stakes Assessment*. Washington, DC: National Clearinghouse for Bilingual Education.
- Illinois State Board of Education, Division of Standards and Assessment** 1996: *Illinois Measure of Annual Growth in English*. Springfield, IL: Illinois State Board of Education.
- Improving America's Schools Act of 1994** 1994: Conference Report 103–761. Regarding Public Law 103–382, signed October 20, 1994, pp. 6–33. Washington, DC: House of Representatives.
- IPT[®] Reading and Writing Proficiency Tests**: 1991: Brea, CA: Ballard and Tighe.
- Johnstone, R.** 2000: Context-sensitive assessment of modern languages in Primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing* 17, 123–43.
- Kopriva, R.** 2000: *Ensuring accuracy in testing for English language learners*. Washington, DC: The Council of Chief State School Officers.
- Macias, R.F.** 1998: *Summary report of the survey of the states' limited English proficient students and available educational programs and services, 1996–97*. Washington, DC: National Clearinghouse for Bilingual Education.
- National Clearinghouse for Bilingual Education (NCBE)** 1999: The growing number of limited English proficient students. Poster

presented by NCBE, Washington, DC: National Clearinghouse for Bilingual Education.

- Rivera, C., Stansfield, C.W., Scialdone, L. and Sharkey, M.** 2000: *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999* (Final Report). Washington, DC: George Washington University, The Center for Equity and Excellence in Education.
- Spicuzza, R., Erickson, R., Thurlow, M., Liu, K. and Ruland, A.** 1996: *Input from the field on assessing students with limited English proficiency in Minnesota's basic requirements exams*. State Assessment Series, Report 2. University of Minnesota, MN, Minnesota Department of Children, Families and Learning.
- Stevens, R., Butler, F.A. and Castellon-Wellington, M.** 2000: *Academic language and content assessment: measuring the progress of ELLs*. Final Deliverable to OERI, Contract No. R305B60002. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- TESOL (Teachers of English to Speakers of Other Languages)** 1997: *ESL standards for pre-k-12 students*. Alexandria, VA: TESOL.
- Texas Education Agency, Student Assessment Division.** 2000: *Texas Reading Proficiency Test in English: spring 2000 baseline administration information guide*. [On-line] Austin: TX.
- The University of Iowa and The Riverside Publishing Company** 1994: *Integrated assessment program, technical summary I Riverside 2000*. Chicago, IL: The Riverside Publishing Company.