Accountability in Language Assessment of Adult Immigrants in Canada[1]

Bonny Norton and Gail Stewart

Abstract: This article addresses the challenges the authors faced in developing a task-based language assessment instrument for adult immigrants in Canada that is accountable to the needs of diverse groups of stakeholders, including learners, teachers, and administrators. The test, the Canadian Language Benchmarks Assessment (CLBA), is designed to place adult newcomers in language programs appropriate for their level of proficiency in English and to assess progress in these programs. The authors note that while stakeholders wanted the assessment tasks to be authentic and realistic, many were concerned that authentic tasks are culturally biased. These concerns were associated with stakeholders' theories of language, their conceptions of test bias, and their understanding of the purpose and use of the test. The authors suggest that when the low-stakes CLBA is used for its intended purpose, its results can satisfy the imperative for accountability in the context of large-scale, system-wide assessment.

Résumé: Le développement d'un test langagier canadien basé sur des tâches et destiné aux immigrants adultes a posé des défis aux auteures car cet instrument devait répondre aux besoins des groupes concernés : apprenants, enseignants et administrateurs. Le test, le « Canadian Language Benchmarks Assessment » (CLBA), sert à placer les immigrants dans des cours d'anglais langue seconde à leur niveau, et à suivre leurs progrès. Selon les auteures, les personnes concernées préfèrent les tâches langagières authentiques et réalistes, cependant, la possibilité d'un préjugé culturel dans ces tâches les inquiètent. Leurs théories du langage et conceptions de préjugés dans les tests, et leur compréhension des buts et usages du test influencent leurs inquiétudes. Les auteures pensent que lorsque le test à bas enjeu CLBA est utilisé à bon escient, il répond au besoin impératif de responsabilité envers les personnes concernées dans le contexte des tests de large envergure relatifs à tout un système.

© 1999 *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 56, 2 (December/décembre)

Introduction

To be accountable in language assessment is to take seriously the investments that different stakeholders have in any given language testing process (Norton, 1997). Stakeholder groups generally include learners, teachers, test developers, administrators, community agencies, public officials, and researchers. In recent years, calls for accountability have focused on the consequences of assessment practices for test takers, who have hitherto been a relatively powerless group in the field of language assessment (Lacelle-Peterson & Rivera, 1994; Norton Peirce & Stein, 1995; Raimes, 1990; Shohamy, 1993). The recognition that language assessment practices should be accountable to test takers is indicative of the broader trend towards democratization of educational assessment in general (Darling-Hammond, 1994; Gipps, 1994). The debate has included arguments that the consequences of a test should be considered an integral part of a test's validity (Messick, 1989).

As interest in accountability towards individual test takers has increased, there has also been a growing demand that assessment practices address system-wide accountability. In a time of economic and political uncertainty, schools, colleges, and universities are under pressure to inform the public about what they are teaching and how effective they are (Brindley, 1998). This trend is not unique to language assessment, but is characteristic of educational assessment in general (Darling-Hammond, 1994; Earl, 1995; Gipps, 1994). In this climate, there is tension between formative and summative assessment, what Gipps (p. 12) calls a 'paradigm clash.'

The tension between these orientations is exacerbated by the fact that, while much research exists to question the validity of large-scale, system-wide assessment (Alderson & Buck, 1993; Lowenberg, 1993; Matthews, 1990; Norton Peirce, 1992), there is little research to support the validity of more learner-oriented assessment, such as performance assessment and portfolio assessment (Balliro, 1993; Reardon, Scott, & Verre, 1994). In the field of language assessment, McNamara (1995) identifies a number of problematic features of performance assessment in the main models proposed in the second language assessment context, particularly with regard to the interactional aspect of performance. Hill and Parry (1994, p. 146) point to some of the problems with alternative forms of assessment, focusing on the extent to which they meet reliability criteria. Furthermore, as Darling-Hammond (1994) argues, alternative assessment methods that appear more accountable to learners are not necessarily equitable, and, as Gearhart and Herman (1995) demonstrate, attempts to integrate learner-oriented performance

assessment into large-scale, system-wide assessment pose many challenges for all stakeholders.

In this article, we describe and evaluate the challenges we faced as we sought to develop language assessment instruments that would integrate learner-oriented performance assessment in a large-scale, nationwide language benchmarks project in Canada. The instrument we developed, called the Canadian Language Benchmarks Assessment (CLBA), is a task-based assessment for adult newcomers to Canada (Norton Peirce & Stewart, 1997). Its purpose is to help place adult language learners across the country in instructional programs appropriate for their level of proficiency in English and to assess learner progress within these programs. The article focuses, in particular, on stakeholder contributions to the test development process and on our attempts to address the questions they posed. It is also, in part, a response to calls for wider reporting of studies on integrative testing:

> Perhaps the most important development will be in the more systematic reporting and dissemination of research and development in the field of integration arising from a growing professionalism in testing. At present, much of this work remains unpublished and unknown. (Lewkowicz, 1997, p. 128)

We have structured this article as follows: first, we examine large-scale assessment research in the Canadian context, with particular reference to language assessment of adult newcomers. This is followed by a description of the history and mandate of our project. Thereafter, we describe how we sought to incorporate a wide variety of stakeholders into the test development process, focusing on their different investments in the project. The bulk of the article is devoted to a discussion of the challenges we faced in addressing stakeholders' assumptions about task-based assessment, authentic tests, and cultural fairness in language testing.

The Canadian context

In Canada, the literature on the development of large-scale language assessment instruments does not do justice to the energy devoted to this industry in the country. There has been extensive work on large-scale assessment of French immersion programs (Lapkin, Hart, & Swain, 1991; Swain, 1985), benchmarks for elementary and secondary schools (Larter & Donnelly, 1993), and innovative assessment for pre-tertiary admission (Des Brisay, 1994; Wesche, 1987). The language assessment

instruments for these programs were developed with a view to providing more learner-centred performance-based assessment than had hitherto existed. The development of the CLBA sought to fill a significant gap with regard to the needs of adult immigrants, who had been largely neglected in the assessment industry.

Cumming (1994, 1995) raises two central issues concerning this particular group of learners that were taken into account in our attempt to develop assessment instruments that would be accountable to adult immigrants to Canada. First, Cumming (1994) poses the question of whether language assessment facilitates the participation of immigrants in Canadian society. In this regard, the goals of assessment should be to provide greater access to the symbolic and material resources of Canadian society, and not to create additional hurdles to successful integration. He views this as a fundamental criterion on which to judge government policies and educational practices for newcomers. Of particular relevance to our article is the following cautionary observation:

> Language assessment may be too limited in scope (i.e., narrowly construed, ad hoc, or unsystematic) to reflect the range and quality of language uses that are actually fundamental to participation in Canadian society. (p. 120)

Second, he argues (Cumming, 1995) that while there is a general trend in education towards accountability and outcomes-based curricula, English as a second language (ESL) programs lack information and empirical validation that would provide evidence on the standards of learning outcomes achieved. With reference to the latter issue, the development of the CLBA can be seen as a response to the limitations of adult ESL instruction and assessment outlined by Cumming (1995) and discussed in greater detail in Norton Peirce and Stewart (1997). The purpose of this article is to focus on the former issue: the challenges we faced in developing a task-based assessment instrument that would facilitate the integration of newcomers into Canadian society.

Background and mandate of the CLBA

The history of the CLBA extends as far back as 1991. In its annual report to Parliament in 1991, Employment and Immigration Canada (now Citizenship and Immigration Canada) indicated its intention to improve the language training offered to adult newcomers by improving language assessment practices and referral procedures (Immigration

Canada, 1991). An important innovation of this new policy was the emphasis placed on partnerships between the federal government and local organizations involved in immigrant language training (Rogers, 1994.) In this spirit, in 1992, Employment and Immigration Canada organized a number of consultation workshops to consider what potential benefits there might be in having a set of national language benchmarks to offer to ESL learners, teachers, administrators, immigrant serving agencies, and governments. In March 1993, the federal government established the National Working Group on Language Benchmarks (NWGLB) (Taborek, 1993) to oversee the development of a language benchmarks document that would describe a 'learner's abilities to accomplish tasks using the English language' (Citizenship and Immigration Canada, n.d., p. 3). The NWGLB represented learners, instructors, and administrators from across Canada.

Once the benchmarks document, *Language benchmarks: English as a second language for adults,* had been developed, Citizenship and Immigration Canada put out a call for proposals to develop listening/speaking, reading, and writing assessment instruments that would be compatible with the benchmarks document. In the call for proposals (Citizenship and Immigration Canada, 1995, pp. 5–6), particular emphasis was placed on the need for the assessment instruments to be sensitive to the diversity and maturity of the client population:

> The bank of tasks developed for the placement of adult immigrants along the continuum as a result of this contract should meet [inter alia] the following criteria:
>
> - be respectful of, and even 'friendly' to, the adult clients and look realistic and fair to them
> - be free from racial and cultural bias

The contractor was expected to ensure the extensive involvement of stakeholders, defined as teachers, settlement workers, immigrant advocacy groups, employers, educational institutions, and ESL learners. The contractor was to meet on a regular basis with the consultant hired to field test the Language Benchmarks document, as well as with CIC officials and the NWGLB. Over a period of 12 months (April 1995–March 1996) the assessment instruments were to be developed, field tested, and professionally produced, along with scoring guides and instruction manuals.

The Peel Board of Education in Mississauga, Ontario, was awarded the test development contract, with Norton and Stewart as test

developers and da Silva and Bergin as Peel Board project management. The mandate of the project, as represented in the contract between the government and the Peel Board of Education, can be summarized as follows:

1. To develop tasks that are benchmarks-compatible.
2. To develop tasks that can place learners on a continuum.
3. To develop tasks that are free from racial and cultural bias.
4. To develop tasks that are realistic and fair.
5. To develop separate instruments for listening/speaking, reading, and writing.
6. To develop both placement and outcomes instruments.
7. To develop assessment instruments that can be administered and scored in an efficient, reliable, and cost-effective way.
8. To develop assessment instruments that are accountable to the field of adult language learning and teaching.

When the contract was nearing completion, Pawlikowska-Walentynowicz was contracted to revise and field test the document later called *Canadian Language Benchmarks Working Document, 1996* (Citizenship and Immigration Canada, 1996), which is being widely disseminated.

Putting accountability into practice

Our experience with the development of the CLBA gave us much insight into the complexities of putting accountability into practice. In seeking to achieve both system-wide accountability and learner-centred accountability, we held consultations with various stakeholder groups in order to determine their expectations for the CLBA. Examples of this consultation process follow:

1. From the beginning of the project, stakeholders were involved in the development of the test model and in the analysis and revision of test items. We had regular meetings with the NWGLB, during which we took the opportunity to get feedback on the test development process as well as to share the challenges that we faced in attempting to meet the sometimes contradictory demands of our mandate. This group, comprising 19 members, was a very important stakeholder group, as it represented learners, teachers, and administrators from different parts of the country.
2. The project team established a Cultural Advisory Group (CAG), representing a variety of community stakeholders, who gave us

valuable feedback and suggestions on the cultural fairness of the tasks. While working with the CAG, we found our discussions to be more productive when we asked the group to provide criteria for cultural fairness, rather than asking for general impressions of individual tasks. For example, in a meeting with the CAG on December 4, 1995, we provided the group with a sample of photographs for possible use in the listening/speaking test. We then asked them to choose which of the photographs they thought most suitable for the purpose of the given task and asked them to explain the criteria they used in the selection process.

3. The reading and writing tasks underwent a field testing process and piloting process (see Norton Peirce & Stewart, 1997). In the field testing, we elicited qualitative feedback from teachers, learners, and consultants concerning the length, content, clarity, and relevance of the individual tasks. As well, during the course of the field testing, teachers were asked to keep a note of their observations of learners taking the tests and to record learner questions and comments. This exercise, we hoped, would give teachers as well as learners a stake in the process. As one teacher remarked,

> I think it was an interesting revelation for us as teachers as well. There was a lot of apprehension prior to the survey, but it was surprising how sportingly the students participated and were delighted at the idea of the whole process ending up in a final exiting test to determine level of proficiency.

During the piloting process, which involved more than a thousand learners from across the country, we also gathered quantitative data on learner performance that were subsequently analyzed and incorporated in the revision process (Nagy, 1996).

4. In the development of the listening/speaking test, we were mindful of the need to work with assessors who would be involved in the implementation process. This was particularly important for the listening/speaking instrument because it is administered on a one-to-one basis. As McNamara (1997, p. 135) notes, 'The interaction of rater characteristics and the qualities of the rating scales they are using has a crucial influence on the ratings that are given, regardless of the quality of the performance.' Throughout the development process, we worked closely with assessors and, at key points in the process, videotaped their interactions with learners. In addition, on February 14, 1996, the

> project team organized a workshop with 17 assessors and administrators from a variety of assessment centres in Ontario. We discussed the challenges they foresaw in the implementation of the CLBA listening/speaking assessment, and their suggestions were incorporated in the revision process.

The feedback from stakeholders was invaluable in helping us determine how best to construct instruments that would satisfy the requirements stipulated in our contract and our mandate. However, different groups had different investments in the project, and a great deal of time was spent in negotiating, explaining, listening, and debating. We were mindful of the fact that if the larger adult ESL community was not invested in the project, the CLBA would have little credibility in this community. It soon became clear, however, that stakeholders had diverse and often mutually incompatible needs. In this regard, 'stakeholders' cannot be understood as a single group, with a unified voice and common objectives. At the outset, the expectations of ESL teachers were that the instruments should accurately place learners in classes and help in determining progress or achievement within those classes. In addition, they hoped that the instruments would comprise varied, motivating, and interesting content that would encourage positive curricular washback and provide precise separate-skill diagnostic information on individual student strengths and weaknesses. Learners wanted instruments with a great degree of relevance to their daily lives. They wanted tasks that appeared authentic and that would allow them to be given credit for the full extent of their proficiency in each skill area. Furthermore, they wanted the instruments to be administered in a non-intimidating format, with ample time allotted for completing them. Administrators desired a time- and cost-efficient package that would be easy to administer in a variety of settings. They indicated a desire for reliable instruments that could be easily and quickly scored and which would be suitable for use by itinerant assessors.

While each of the stakeholder groups had differing investments in the CLBA, they all struggled with three related questions: (a) How can we reconcile task-based assessment with separate-skills assessment? (b) What are the limits of authenticity in language testing? (c) How can we create a test that is 'free from cultural bias'? Since we spent a great deal of time addressing these questions, which are perennial ones in the field of language testing, we will discuss each question in greater depth below.

How can we reconcile task-based assessment with separate-skills assessment?

Cumming's call for language assessment that reflects the range and quality of usage necessary for participation in Canadian society is reflected in the task-based approach defined in the call for proposals (Citizenship and Immigration Canada, 1995):

> the items must be task-based, i.e. they measure ability to accomplish realistic tasks using the English language. Particularly at the lower and intermediate ranges of the benchmarks a high proportion of the tasks should be reflective of those encountered by immigrants or refugees during initial settlement. (p. 5)

In this spirit, the Language Benchmarks document, which served as our test specifications, required that adult ESL learners be able to perform a variety of authentic tasks that are commonly carried out in the Canadian social and business context. The sample tasks listed within that document included such activities as reading menus, interpreting traffic signs, and filling out banking forms. These sample tasks reflect certain highly desirable features of the task-based approach in that they are authentic and meaningful, as well as bringing a high degree of relevance to assessment. At the same time, however, our mandate was to develop three separate instruments: one for listening/speaking, one for reading, and one for writing:

> The CLB recognizes three skill areas, treating speaking and listening as one domain. A single global scale of proficiency description for all skills for the CLB was rejected by the ESL field in the consultation and field testing process. Separate scales reflecting competencies within skill areas were favoured for many reasons. (Citizenship and Immigration Canada, 1996, p. 11)

Stakeholders soon realized, however, that many highly realistic and meaningful tasks are unsuitable for the purpose of separate-skills assessment. Authentic tasks are fluid and involve the simultaneous or rapidly alternating use of different language skills. For example, one of the specified writing tasks in the Language Benchmarks document involved taking a telephone message. This type of authentic task can be problematic for the purpose of assessing writing, since a successful note-taking exercise requires as much proficiency in listening as in

writing. A task-based approach to language assessment, therefore, presumes a certain degree of skills integration – which conflicts with the CLBA requirement that the instruments isolate the language skills to simplify administration and render the skill-by-skill diagnostic information sought by some stakeholders.

The problems associated with task-based integrative assessment are not unique to the Canadian context. As Lewkowicz notes,

> task dependence which is characteristic of integrated tests gives rise to what Weir (1990) refers to as 'muddied measurement.' This occurs when performance on one test task affects that of another, for example, when the success in completing a writing task depends on understanding a reading or listening extract. In such circumstances it may be difficult to determine where the process has broken down and accurately to profile candidates' strengths and weaknesses. (1997, p. 127)

Two approaches to the integrative task vs. separate-skills dilemma have been posited by language testers. Bachman and Palmer (1996), on the one hand, have rejected, for the purposes of language testing, the division of language proficiency into the four skill areas of speaking, listening, reading, and writing, focusing instead on the inseparability of ability and task characteristics. They note as follows:

> This approach to distinguishing the four skills treats them as abstract aspects of language ability, ignoring the fact that language use is realized in specific situated language use tasks. Thus, rather than attempting to distinguish among four abstract skills, we find it more useful to identify specific language use tasks and to describe these in terms of the task characteristics and the areas of language ability they engage. (p. 70)

Notwithstanding the widespread use of the model developed by Bachman (1990) and Bachman and Palmer (1996), Shohamy (1997, p. 146) notes that 'in the theoretical domain there is still no evidence for the validity of the Bachman model and its revised version in spite of its wide acceptability.'

The second approach to the dilemma is posited by Cummins (1999), who argues for the legitimacy of assessing a more generalized language proficiency, reminiscent of the work of Oller (1976). He draws on recent studies to argue as follows:

> Despite the fact that notions of 'general language proficiency' are not particularly fashionable at the moment, this construct emerges consistently in

> virtually all of the psychometric research on the nature of second language proficiency. There is also evidence for other more specific components of proficiency but the bulk of the variance in most batteries appears to be accounted for by a common general language proficiency factor. (Cummins, 1999, p. 36)

The important difference between his work and Oller's, however, is that as proficiency increases, the construct is more accurately conceptualized as reflecting what Cummins calls 'academic language proficiency' than as reflecting general or global language proficiency. As Cummins (1996, p. 59) argues, 'the essential aspect of academic language proficiency is the ability to make complex meanings explicit in either oral or written modalities by means of *language itself* rather than by contextual or paralinguistic cues (e.g. gestures, intonation etc.).'

Given the demands of our mandate, as well as the inconclusive literature on this topic, we sought to reconcile the often conflicting requirements of separate-skills imperatives with the task-based approach. Our strategy, somewhat reluctantly undertaken, was to narrow our selection of assessment-eligible tasks. We began by eliminating those tasks which introduced an obvious conflict between oral/aural and written components of language. As a result, for example, telephone messages were eliminated from the pool, along with lecture note-taking and any task involving the oral summary of print material. Next, we examined those tasks that required a combination of listening/speaking or reading/writing proficiency. In the case of the former combination, we recommended that the integrated structure of listening and speaking presented in the benchmarks document be maintained in the assessment. This enabled us to develop a listening/speaking assessment based on a face-to-face oral interview. We agreed with the CLB position that 'using speaking but not listening or vice versa in most oral communication/exchange situations is hardly conceivable, except for broadcast and lecture situations' (Citizenship and Immigration Canada, 1996, p. 11). When it came to reading and writing, teachers and administrators made the argument that many courses for adult immigrants are divided into reading classes and writing classes, and that the separation of these skills was highly desirable for placement into these classes. We accepted this argument, focusing on selecting and developing tasks that strongly emphasized one skill over the other. In our guidelines for task writers, for example, we stressed that reading tasks should require a minimum of writing on the part of candidates, and that writing tasks should minimize the focus on reading text.

What are the limits of authenticity?

ESL learners found the use of authentic tasks in the benchmarks highly desirable. In a survey conducted by Ontario Welcome House (1994), the following responses were typical:

> They [the benchmarks] describe all the steps and problems that ESL learners pass through.
> [They] cover all the important activities learners can do when they use English.

Notwithstanding such positive responses, we found that in the context of language assessment, authentic tasks pose a number of dilemmas for stakeholders. There was much discussion, for example, about a task that required candidates to complete a standard form – one of the writing tasks suggested in the Language Benchmarks document. As stakeholders indicated, the requirements for filling out a form on an assessment may not be the same as the requirements in a real life situation. In many real life circumstances, it might be quite legitimate for an immigrant to provide a name and address only, and respond 'not applicable' to all other questions. Stakeholders debated whether such a response on an assessment would render the task incomplete. We all agreed that in the real world, a reader of a form is interested in the accuracy and veracity of the information presented. It is, in fact, an offence to falsify information on many of the standard forms an immigrant might encounter in Canadian society. In a language testing situation, accuracy and veracity may be less important. We recognized, furthermore, that some learners might be intimidated by an official-looking form requesting personal data, while others might consider the task an invasion of privacy. As one teacher noted:

> A couple of students were concerned about the anonymity of the test. They asked if they had to use their real names to fill out the forms.

The Language Benchmarks document also included tasks based on authentic reading material such as classified advertisements, flyers, consumer catalogues, maps, and travel guides. Some of these materials raised important ethical issues for some stakeholders. One advertisement, for example, which dealt with the sale of automobiles, though relevant to life in Canada, was viewed by some stakeholders as promoting a consumerist philosophy and flaunting a product that many

recent immigrants might be unable to afford. Other advertisements were considered problematic because they made reference to specific products, real or fictitious: tasks based on real products were seen by some as free advertising, while those based on fictitious products were considered misleading. In terms of maps and travel guides, tasks based on real locations were viewed as giving unfair advantage to learners from those locations, while tasks based on fictitious locations were seen to be confusing to all candidates.

The real life experience of many adult immigrants is not always a felicitous one, and daily life in Canada sometimes involves rejection and disappointment (Goldstein, 1996; Morgan, 1998). In attempting to reflect as full a range of authentic tasks as possible, the original bank of trial CLBA tasks included, among others, a report on an automobile accident and a letter from a school principal regarding a child's difficulties at school. These tasks were criticized by some stakeholders as being too negative for assessment purposes. Other stakeholders were concerned that a broad range of authentic tasks cannot be easily accomplished in a test situation. The specifications in the Language Benchmarks document for the higher levels included some reading tasks of essays and journal articles. Though stakeholders felt it was important to include sufficient variety to allow learners to demonstrate the full scope of their language proficiency, they recognized that it was difficult to combine tasks of this length and complexity into an instrument that could take about an hour to complete.

In our stakeholder groups, we discussed the limits of authenticity in language testing. As Bachman (1990) notes, language testers have been preoccupied with this aspect of language testing for over a quarter of a century, devoting an entire issue of the journal *Language Testing* to this topic (volume 2, no. 1, June 1985). In a comprehensive discussion of authenticity, Bachman makes a useful distinction between what he calls the 'real life (RL) approach to authenticity' and the 'interactional/ability' (IA) approach. The RL approach, perhaps the most common understanding of the term, refers to the extent to which test performance replicates some specified non-test language performance. It is primarily concerned with the appearance of the test and how it may affect test performance and test use (sometimes referred to as 'face validity') as well as the accuracy with which test performance can predict future non-test performance. The IA approach, on the other hand, focuses not on non-test performance per se, but rather on the interaction between the language user, the context, and the discourse. As Bachman (1990) notes,

> test performance is interpreted as an indication of the extent to which the test taker possesses various communicative language abilities, and there is a clear distinction in this approach between the abilities to be measured, on the one hand, and the performance we observe and the context in which observations take place, on the other. (pp. 302–303)

Norton Peirce (1992) and Norton Peirce and Stein (1995) draw on their research with standardized reading tests to argue that attempts at 'authenticity' (in particular the RL approach, as defined by Bachman) are inevitably flawed because of the unequal relations of power between test makers and test takers. When a test taker reads a passage, they argue, the test taker's central concern is not, 'How do I make sense of this passage?' but rather, 'How am I *expected* to make sense of this passage?' They draw on the work of Kress (1989) to argue that a language test is a genre no less authentic, in and of itself, than any other social occasion. As Norton Peirce (1992) argues,

> while test makers have generally assumed that a standardized reading test is an aberration in the 'real world,' I wish to argue that it is no less authentic a social situation than an oral presentation or a visit to the doctor. In a standardized reading test, the value ascribed to texts within this genre is associated with a ritualized social occasion in which participants share a common purpose and set of expectations. (p. 685)

The central point they make, however, is not that language tests should be 'inauthentic,' but rather that there are a host of ethical issues associated with debates on authenticity. In this sense, they echo the position taken by Spolsky (1985):

> The criterion of authenticity raises important pragmatic and ethical questions in language testing. Lack of authenticity in the material used in a test raises issues about the generalizability of the results. Any language test is by its very nature inauthentic, abnormal language behaviour, for the test taker is being asked not to answer a question giving information but to display knowledge or skill. With examinees who do not know the rules of the game, or who are unwilling to play according to them, the results will not be an accurate and valid account of their knowledge. (p. 39)

In the development of CLBA, we were asked to develop tasks that would 'be respectful of, and even "friendly" to, the adult clients and look realistic and fair to them.' We therefore strove to develop tasks

that did have high face validity or, at least avoided what Messick (1989, p. 19) calls 'face invalidity.' At the same time, however, we made every effort to ensure that learners did know 'the rules of the game': the instructions on all tasks have been carefully developed to be as comprehensive and unambiguous as possible; candidates have a generous amount of time to complete the tasks; markers and assessors are encouraged to 'bias for best' (Swain, 1985); and candidates are treated with great respect in the listening/speaking assessment. In encouraging markers and assessors to 'bias for best,' we wanted assessors to give credit to learners for their strengths rather than giving disproportionate attention to their weaknesses. Lack of accountability in this regard would render the CLBA not only invalid, but unethical.

How can we create a test that is free from cultural bias?

In the test development process, a fundamental paradox emerged: for many stakeholders, the greatest strength of the task-based approach was also its greatest weakness. While all stakeholders agreed that the trial tasks were appealing, meaningful, and realistic examples of the tasks required in day-to-day Canadian life, they also regarded many of the same tasks as culturally biased. In the case of a reading task that centred around a menu, for example, a hamburger was considered to be a culture-specific food that, if included in a test, would disadvantage some candidates. A voting notice was thought to be too unfamiliar to learners from non-democratic countries. Some respondents viewed a hydro bill as gender-inappropriate, pointing out that in some cultures, women never pay, or in fact even see, bills of any kind. A fax memo format was believed to require knowledge of Canadian business practices, and a map an understanding of Western geography. In addition, there was disagreement with regard to the vocabulary and terminology used in the trial tasks. Suggestions that a 'washroom' should be referred to as a 'toilet,' and an 'apartment' as a 'flat,' or that a 'chain' of supermarkets is more commonly labelled a 'group,' gave rise to questions concerning what constitutes standard Canadian English. Some stakeholders requested that we develop tasks that were culturally 'neutral.' However, when we asked stakeholders to suggest tasks that would meet this criterion, they were unable to agree on a single culturally neutral task.

Test developers have long struggled with this critical aspect of accountability (Bond, 1995; Cole & Moss, 1989; Duran, 1989; Zeidner, 1986). In the CLBA project, we soon became aware that much of the conflict experienced by our stakeholders could be attributed to three

related issues: stakeholders' understanding of what it means to know a language; their conceptions about the nature of 'test bias'; and their uncertainty about the purpose and use of the test. With reference to the first issue, it is instructive to consider what one administrator said in relation to the CLBA trial reading task based on a menu:

> When I came to this country, my English was very good. But I would not have understood this menu.

Implicit in this observation is the view that knowledge of a language can be separated from knowledge of its use. This theme recurred frequently in the field testing process, as illustrated in the following comment from one of the teachers:

> Most of the tasks test not only the student's language competency but also his or her familiarity with the customs and practices common in Canada which may put some newcomers to this country at a disadvantage. They may perform the task below their language abilities because they lack the 'life experience.' For example, in their native country one is not required to produce a resume when one applies for a job. Similarly, they may lack the experience of renting a video because there are no video rentals in their country.

Distinctions drawn between knowledge of language and knowledge of its use included reference to practices such as reading maps and diagrams. As one teacher said:

> Many of the tasks are based on graphics and cannot be performed without the ability to read maps and diagrams. Some students may lack this ability because of little formal schooling or poor spatial orientation. As a consequence, the level of their performance does not reflect the level of their language proficiency.

As we discussed with our stakeholders, the task-based approach used in the CLBA was predicated on the model of language proficiency advocated by Canale and Swain (Citizenship and Immigration Canada, 1996, p. 13). In these well-established theories of communicative competence (Canale & Swain 1980; Canale, 1983), knowledge of a language comprises four competencies: linguistic competence, strategic competence, discourse competence, and sociolinguistic competence. Linguistic competence refers to the grammatical well-formedness of sentences; discourse competence addresses the extent to which sentences are

combined cohesively and coherently; sociolinguistic competence addresses considerations of appropriateness in interaction; and strategic competence refers to the extent to which speakers use strategies to deal with communication breakdown. The assumption made by the developers of the benchmarks documents is that while adult newcomers may indeed have some degree of linguistic competence in English, they also need to have strategic, discourse, and sociolinguistic competence for full participation in Canadian society (Citizenship and Immigration Canada, 1996, pp. 12–16). This conception of language is consistent with that advocated by Cumming (1994) and guided our decision-making throughout the project.

Many stakeholders remained ambivalent, however. On the one hand, they wanted the tasks to be authentic and realistic; on the other hand, they wanted the tasks to assess linguistic competence only. Tasks that addressed strategic competence, discourse competence, and sociolinguistic competence were understood by many as 'biased.' It was for this reason that we had many discussions on the nature of test bias. In the *Language Benchmarks Bias Review Report*, Dixon (1995) defines 'bias' as follows:

> A bias is an uninformed or unintentional preference that hinders impartial judgment. Most biases are subtle and difficult to recognize because they are based on value-loaded ideas or beliefs. (p. 1)

In the language testing literature, however, conceptions of test bias differ somewhat from the definition offered by Dixon. Bachman (1990) notes as follows:

> The process of validation is addressed to specific test uses and specific groups of test takers. But within these groups there may be subgroups that differ in ways other than the language ability of interest. These differences may affect their test performance, and hence the validity of inferences we make on the basis of test scores.... When this happens, we speak of *test bias*. (p. 271)

The central point Bachman makes is that differences in group performance, in themselves, do not necessarily indicate the presence of bias, since such differences may reflect genuine differences between groups on the ability in question. It all depends, crucially, on the purpose of the test, the construct being measured, and the decisions that need to be made on the basis of the test. This idea led to comprehensive discussions on the purpose of the CLBA.

As we discussed with stakeholders, the very purpose of the CLBA, consistent with the CLB Working Document, is to determine to what extent a learner can function communicatively within the Canadian context.

> The CLB is learner-centred. It is reflective of the needs of newcomers and the challenges they face in the process of their settlement and integration into Canadian society. (Citizenship and Immigration Canada, 1996, p. 10)

If the CLBA were to strip that context of its cultural component and assess linguistic competence only (to the extent that this is possible), the CLBA would be invalid. This would do a disservice to learners and other stakeholders who expect the instruments to identify the extent to which newcomers can access the symbolic and material resources of Canadian society. The extent to which the tasks used in the CLBA are representative of the tasks newcomers regularly encounter in the Canadian context is another issue. In this regard, the tasks that were developed were based on a Language Benchmarks document that was comprehensively field tested across the country with approximately 3000 participants (Citizenship and Immigration Canada, 1996, p. ii). Further, in the CLBA, which is a low-stakes placement assessment, clients are not penalized for having had no previous access to the symbolic and material resources of Canadian society. Rather, the assessment instruments are used to determine the amount of access achieved, so that learners can be placed in programs where gaps in access can be appropriately addressed.

Conclusion

Accountable test development is an ongoing effort to address contradictions among conceptual framework, technical elements, and political and ethical issues. With any large-scale assessment, this challenge is compounded because of the variety of interests that have to be served. Alderson, Clapham, and Wall (1995) refer to the give-and-take that is an inevitable component of the test development process. While incompatibilities between task-based and separate-skills assessment were problematic, the tensions between authenticity and cultural fairness proved to be the most challenging in this particular project. While stakeholders wanted the tasks to be authentic and realistic, they did not want learners to be penalized if they lacked knowledge about authentic cultural practices in Canadian society. This tension, we have suggested, is associated with three issues: stake-

holders' conceptions of what it means to know a language, their understanding of test bias, and their expectations of the purpose and use of the test. We have argued that the CLBA was developed within the framework of communicative language learning and teaching, in which linguistic competence is considered necessary but not sufficient for full participation in Canadian society. It is a low-stakes assessment instrument whose primary purpose is to place learners in programs where limitations in communicative language ability can be addressed. Future research is necessary to determine whether the CLBA is being used for the purposes for which it was developed and whether it does, indeed, facilitate the integration of newcomers into Canadian society.

Bonny Norton is Assistant Professor in the Department of Language Education at the University of British Columbia. She has worked on test development projects internationally, including the TOEFL, the University of Toronto's COPE admission test, and the CLBA. Her research on language assessment has been published in *TESOL Quarterly, Harvard Educational Review, Language Testing,* and *TESL Canada Journal.*

Gail Stewart is an educational consultant and teacher trainer whose work in the area of second language assessment includes the University of Toronto's COPE admission test, the Midwives' Language Proficiency Test (MLPT), the CLBA, and the Canadian Language Benchmarks Literacy Assessment (CLBLA). She is currently working on the design and development of a streamlined version of the CLBA.

Notes

1 The authors acknowledge the insightful contributions of stakeholders to the development of the CLBA. We thank reviewers of the *Canadian Modern Language Review* for their invaluable comments and Monique Bournot-Trites for her careful translation of the abstract. This article was written when the first author held a Spencer Postdoctoral Fellowship from the National Academy of Education, USA. This support is gratefully acknowledged.

References

Alderson, J.C., & Buck, G. (1993). Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing. *Language Testing, 10,* 1–26.

Alderson, J.C., Clapham, C., & Wall, D. (1995). Standards in language testing: The state of the art. In *Language test construction and evaluation* (pp. 235–260). Cambridge: Cambridge University Press.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Balliro, L. (1993). What kind of alternative? Examining alternative assessment. *TESOL Quarterly, 27,* 558–561.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice, 14*(4), 21–24.

Brindley, G. (1998). Assessment and reporting in language learning programs: Purposes, problems, and pitfalls. *Language Testing, 15,* 45–85.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2–27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1–47.

Citizenship and Immigration Canada. (n.d.). *Language benchmarks: English as a second language for adults.* Ottawa: Citizenship and Integration Policy Division and Communications Branch, Citizenship and Immigration Canada.

Citizenship and Immigration Canada (1995, February). Call for proposals. Ottawa: Citizenship and Immigration Canada.

Citizenship and Immigration Canada. (1996). *Canadian language benchmarks working document.* Ottawa: Minister of Supply and Services.

Cole, N., & Moss, P. (1989). Bias in test use. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 201–219). New York: American Council on Education/Macmillan Publishing.

Cumming, A. (1994). Does language assessment facilitate recent immigrants' participation in Canadian society? *TESL Canada Journal, 2*(2), 117–133.

Cumming, A. (1995). Changing definitions of language proficiency: Functions of language assessment in educational programmes for recent immigrant learners of English in Canada. *Revue de l'ACLA/Journal of the Canadian Association of Applied Linguistics, 17,* 35–48.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society.* Ontario, CA: California Association for Bilingual Education.

Cummins, J. (1999). *The construct of general language proficiency: Theoretical foundations, assessment, and articulation to the Canadian Language Benchmarks.* Unpublished manuscript.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review, 64,* 5–30.

Des Brisay, M. (1994). Problems in developing an alternative to the TOEFL. *TESL Canada Journal, 12*(1), 47–57.

Dixon, B. (1995). *Language Benchmarks bias review report.* Unpublished manuscript.

Duran, R. (1989). Testing of linguistic minorities. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 573–587). New York: American Council on Education/Macmillan Publishing.

Earl, L.M. (1995). Assessment and accountability in education in Ontario. *Canadian Journal of Education, 20,* 45–55.

Gearhart, M., & Herman, J. (1995). *Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability.* Los Angeles: UCLA Center for the Study of Evaluation.

Gipps, C.V. (1994) *Beyond testing: Towards a theory of educational assessment.* Washington, DC: Falmer Press.

Goldstein, T. (1996). *Two languages at work: Bilingual life on the production floor.* Berlin: Mouton de Gruyter.

Hill, C., & Parry, K. (1994) *From testing to assessment: English as an international language.* London: Longman.

Immigration Canada. (1991). *Annual report to Parliament, immigration plan for 1991–1995, year two.* Ottawa: Employment and Immigration Canada.

Kress, G. (1989). *Linguistic processes in sociocultural practice.* Oxford: Oxford University Press.

Lacelle-Peterson, M., & 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.

Lapkin, S.D., Hart, D., & Swain, M. (1991). 'Early' and 'middle' French immersion programs: French language outcomes. *The Canadian Modern Language Review, 48,* 11–40.

Larter, S., & Donnelly, J. (1993). Demystifying the goals of education: Toronto's benchmark program. *Orbit,* 24(2), 22–28

Lewkowicz, J.A. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 121–130). Dordrecht: Kluwer Academic Publishers.

Lowenberg, P. (1993). Issues of validity in tests of English. *World Englishes, 12,* 95–106.

Matthews, M. (1990) The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal, 44,* 117–121.

McNamara, T.F. (1995). Modelling performance: Opening Pandora's Box. *Applied Linguistics, 16,* 159–179.

McNamara, T.F. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 131–135). Dordrecht: Kluwer Academic Publishers.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Morgan, B. (1998). *The ESL classroom: Teaching, critical practice, and community development.* Toronto: University of Toronto Press.

Nagy, P. (1996). *A report on technical aspects of test development for the Canadian Language Benchmarks Assessment.* Unpublished manuscript.

Norton, B. (1997). Accountability in language assessment. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment* (pp. 313–322). Dordrecht: Kluwer Academic Publishers.

Norton Peirce, B. (1992). Demystifying the TOEFL Reading Test. *TESOL Quarterly, 26,* 665–689.

Norton Peirce, B. & Stein, P. (1995). Why the monkeys passage bombed: Tests, genres and teaching. *Harvard Educational Review, 65,* 50–65.

Norton Peirce, B. & Stewart, G. (1997). The development of the Canadian Language Benchmarks Assessment. *TESL Canada Journal, 14*(2), 17–31.

Oller, J.W., Jr. (1976). Evidence of a general language proficiency factor: an expectancy grammar. *Die Neueren Sprachen, 76,* 165–74.

Ontario Welcome House. (1994, November). TESL Canada Fourth Learners' Conference Report, Toronto, ON.

Raimes, A. (1990). The TOEFL Test of Written English: Causes for concern. *TESOL Quarterly, 24,* 427–442.

Reardon, S., Scott, K., & Verre, J. (1994). Symposium: Equity in Educational Assessment. *Harvard Educational Review, 64,* 1–4.

Rogers, E. (1994.) Canadian federal language policy and the Benchmarks project. *TESOL Matters, 3*(6), 1–5.

Shohamy, E. (1993) The exercise of power and control in the rhetorics of testing. In A. Huhta, K. Sajavaara, & S. Takalo (Eds.), *Language testing: New openings* (pp. 23–29). Jyvaskyla, Finland: oy Sisasuomi.

Shohamy, E. (1997). Second language assessment. In G.R. Tucker & D. Corson (Eds.), *Encyclopedia of language and education, volume 4: Second language education* (pp. 121–130). Dordrecht: Kluwer Academic Publishers.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2,* 31–40.

Swain, M. (1985). Large-scale communicative language testing: A case study. In S. Savignon & M. Burns (Eds.), *Initiatives in communicative language teaching.* Reading, MA: Addison-Wesley.

Taborek, E. (1993). The national working group on language benchmarks. *TESL Toronto Newsletter,* fall-winter 93/94, 10–11.

Wesche, M. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4,* 28–47.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing, 3,* 80–98.