

Research Study on Potential Approaches to Second Language Assessment

Philip Nagy
Gail Stewart

Final Report
March 2009

Executive Summary

This report points the way forward to development of an outcomes assessment system for the LINC program, with the possible participation of other agencies and governmental departments. It begins with a review of the development and use of the Canadian Language Benchmarks (CLB) and the original widely-used CLB-referenced test, The Canadian Language Benchmarks Assessment (CLBA). Next is a review of research showing the increasing use of ESL training in recent years as a way to access professions and trades, pointing to the need for a high-stakes, secure test to address the higher range of benchmarks. It is of some concern that tests not designed for high stakes, such as the CLBPT, are being used inappropriately as outcomes measures.

LINC assessment consists of three elements: placement, progress, and outcomes. We view existing procedures for placement and progress, involving use of the CLBA, as essentially sound. We recommend that new forms of the CLBA be developed, and that modifications to the CLBA to produce separate scores for listening and speaking be done, in order to bring the CLBA into line with the 2000 version of the benchmarks.

The focus of our attention is on outcomes assessment procedures. Because LINC program goals include more than just language, we note a danger that a high stakes language test may result in narrowing the focus of LINC, so that the most important goal, language, becomes the only goal. To guard against this, we recommend development of curriculum materials that emphasize non-language goals of LINC, inclusion of Canadian context materials in the proposed testing procedures, and implementation of a broad program evaluation model that includes the non-language goals of LINC.

As background, we review some of the difficulties of using the CLB 2000 as a framework for test development. At the same time, with revisions to the benchmarks document pending, we point out the mutual benefits of developing an assessment and revising the framework concurrently. We review the distinction between a placement test and an outcomes test, arguing against any attempt to use *gain* scores (outcome scores minus placement scores) as measures of either student success or program quality.

An outcomes test focuses on accuracy at particular levels of achievement. The first important decision in designing an outcomes test is selection of these levels. This must be done in broad consultation with a range of stakeholders, including any that might be identified as having a possible future interest in the assessment procedure (e.g., Canadian Experience Class, Enhanced Language Training).

We review the possible methods that might be used for a formal high-stakes outcomes test, concluding that reading and listening should be assessed with objective item types while speaking and writing should be assessed by the direct method. In terms of an informal assessment component, the portfolio approach is examined and recommended as a means of classroom evaluation, but not as a high-stakes option.

Our recommendation is broken down into three components, each accompanied by estimates of time and cost: (a) an informal assessment that can be used for student progress and promotion and for program evaluation; (b) an initial development of a high stakes standardized assessment, starting either *de novo* or from an existing test (possibilities are reviewed); and (c) additional steps required to put in place an ongoing assessment procedure, apart from the initial start-up.

1. Introduction

In 1995, in response to a national call for standardization across English as a Second Language (ESL) programs in Canada, a set of benchmarks was posited for use as a framework to inform curriculum and assessment. The resulting document was an initial draft of the Canadian Language Benchmarks (CLB, Citizenship and Immigration Canada, 1995), describing 12 levels of ESL ability across three skill areas – Reading, Writing, and Listening/Speaking. Subsequently, the CLB was revised and refined over a five-year period through a process of national consultation that included field testing and focus groups. The framework was also informed by research undertaken by the developers of the first CLB-based assessment, the Canadian Language Benchmarks Assessment (CLBA) (Peirce & Stewart, 1997).

In 1996, the Canadian Language Benchmarks Working Document (Citizenship and Immigration Canada, 1996) was produced. This version of the document introduced a theoretical model that had emerged in response to task-based assessment challenges encountered in the development of the original forms of the CLBA. In essence, this model sought to clarify the relationship and distinction between two important concepts – learner proficiency and task difficulty – by reinforcing the notion of the CLB as a continuum of learner proficiency and not of task difficulty. An emphatic feature of this model was the contention that a single language task can never be pegged to one specific benchmark (Stewart, 2005).

The CLBA was well received by the ESL field and has been used for learner placement in Language Instruction for Newcomers to Canada (LINC) programs for more than 12 years. In some LINC and ESL programs, a shorter and less reliable assessment, the Canadian Language Benchmarks Placement Test (CLBPT), is also used. This streamlined assessment was originally designed to meet the needs of administrators who felt that they simply did not have the time or resources to administer the longer and more diagnostic CLBA test. Because of its design and scoring procedures, the CLBPT can be considered an adequate instrument for placement purposes only in programs where there are no stakes attached to the test results and where students can easily be moved from one class level to another if they are misplaced. Unfortunately, it appears that an increasing number of test users, who perhaps lack a complete understanding of the limitations that apply to CLBPT, are adopting this test due to its convenience. It is, in fact, somewhat alarming to note that the CLBPT has been used in some contexts, such as colleges, where the stakes for learners are relatively high.

The colleges are not the only contexts where high stakes are associated with language testing. Even in LINC, the stakes are increasing, largely due to an evolution in the profile of Canadian immigration. Prior to 1990, an estimated 13 percent of newcomers to Canada were destined for highly-skilled occupations, whereas in the period between 1996 and 2000, that average may have

risen as high as 24 percent (Conference Board of Canada, 2004). More recent data indicate that this trend remains strong:

The Canadian labour market has steadily become more knowledge-intensive, with the fastest growth occurring in jobs that require a higher level of postsecondary education. The labour force has responded to this demand for skills by enhancing its educational outcomes. (HRSDC, 2007)

Therefore, a market-driven need exists for newcomers to have more advanced language skills to accompany the levels of education that they have brought from their home countries.

This ongoing trend has resulted in ESL delivery that is increasingly geared toward addressing issues of access to professions and trades. Where the main area of concentration used to be the lower benchmarks we are now finding a much greater emphasis on the higher benchmark levels. This can be seen through an examination of the types of curriculum guidelines published since 1997. The first guidelines (1997) were created for LINC 1-3, while LINC 4-5 guidelines were published in 1999 and LINC 5-7 Guidelines in 2007. Enhanced Language Training (ELT) emerged in 2003, followed by Occupation Specific Language Training (OSLT) in 2008, to address the needs of high-intermediate to advanced learners, while programs that provide a bridge to employment are becoming more common. At present, many Enhanced Language Training programs exist for various sectors (including financial services, technology and business sectors, engineering, health sector).

Within this context, program outcomes take on a greater significance. The very real possibility that these outcomes might be associated with considerations such as workplace eligibility raises the stakes and creates a demand for standardized, consistent and appropriately validated assessment practices. Research on the connection between language skills and labour market integration of newcomers tends to be dated but still relevant. For example, Aydemir and Skuterud (2005) have documented, over the period 1996 to 2000, the difficulties that educated immigrants, especially from non-traditional source countries, have had in converting their education into income (for further examples, see Boyd, 1992; Chiswick & Miller, 1992, 1995, 2003; Pendakur & Pendakur, 1997; Lochead & Mackenzie 2005; Metropolis Secretariat, 2006). In this context, work on developing comparative frameworks for CLB and Essential Skills (Stewart, Geraci & Nagy, 2004) continues.

A recent study of current practices (Makosky, 2008) indicates that LINC outcomes assessment procedures have not kept pace with present and emerging demands for accountability. In most programs, outcomes are determined by classroom instructors based on a variety of ad-hoc methods which are neither standardized nor mandated. The result is a lack of clear information about progress within and across LINC programs. Makosky concludes with a series of recommendations for developing a new and standardized LINC exit procedure.

This paper builds on the information that Makosky has gathered by discussing in detail the ways in which Citizenship and Immigration Canada (CIC) can best respond to the identified need for new assessment practices. It describes the considerations, issues and steps that would be necessary to create a coherent entry/exit assessment system that is CLB-based and suitable for the present and future needs of the LINC programs. The paper is in four parts. The first section describes the LINC assessment context, the second section explores issues that need to be considered, and the third section describes potential approaches to developing both an informal placement and promotion model that will support all LINC goals and a more formal LINC outcomes assessment for exit certification. The fourth section recommends the necessary maintenance and follow-up to ensure successful implementation.

Note that, while the informal assessment component can be developed into a program evaluation model, we have not pursued the question of program evaluation in detail, judging this to be beyond the scope of this document.

2. The Language Program Context

The target assessment context comprises three main components - placement, progress, and outcomes. In terms of placement, it would appear that efficient, effective and sufficiently reliable procedures have been well established over many years of development, beginning with the A-LINC test and culminating with the CLBA. Given that the current LINC approach to placement appears to be fairly stable and quite well accepted, there is no need to consider development of an entirely new placement assessment. Instead, the necessary resources should be allocated to ongoing maintenance of the established placement procedures, including the development of additional forms of the CLBA and implementation of support for all the goals of LINC, including non-language goals. Additional forms are needed to ensure the security of the test, but given the low stakes attached to program placement, this is perhaps not an urgent necessity. The more urgent issue is the need to align the existing assessment with the current separate-skills CLB 2000 model for speaking and listening. When the CLBA was originally developed, its specifications were based on the 1996 CLB Working Document, which presented an integrated speaking/listening skills approach.

As indicated in the Makosky report, the most pressing need in LINC programming is for a procedure that can reliably assess the CLB levels of learners as they graduate or leave the program for some other environment (e.g., further education or employment). Because of the large LINC enrolment, it is important to focus on an outcomes procedure that can be efficiently and flexibly administered in a variety of situations and locations. It is also important to consider the stakes that might be associated with outcomes assessment across the various CLB levels. LINC programming appears to be inclining toward increasingly higher levels, moving from the original focus on Stage I of the benchmarks into the early and mid ranges of Stage II, and these higher levels often imply higher stakes for learners. Between thirty and forty percent of the current LINC enrolment is situated at benchmark 4 or higher where distinctions between program outcomes and real-world eligibility begin to blur somewhat.

We suggest that there exist on the CLB continuum certain key points at which these high-stakes outcomes are most critical and where clear identification and agreement is an urgently-needed high-level decision that concerns agencies outside the language training professions – stakeholders who need some clear and accurate certification of an individual’s language level. For this reason, it may not be unreasonable to contemplate a LINC outcomes assessment approach that would be compatible with other high-stakes applications such as certification and/or eligibility. Such an approach would not have to lead to a single assessment that could be used for all purposes, but it might render a series of related assessments, all sharing the same fundamental underpinnings. This would be advantageous because it would greatly reduce costs and would provide some assurance of

compatibility of results. It would also assist in enhancing the standardization of curricula for specific instructional purposes.

We suggest various professional accreditation bodies who might be interested in using a formalized standard assessment for ESL (perhaps with occupation-specific subtests).

Among these groups we include:

- Selection:
 - CEC Class
 - Foreign Skilled Workers
- Integration
 - LINC (Language Instruction for Newcomers to Canada)
 - CLIC (Cours de langue pour les immigrants au Canada)
 - ELT (Enhanced Language Training)
 - CLNA (Cours de langue niveau avancée)
 - OSLT (Occupation Specific Language Training)
 - FLAP (Formation linguistique axe sur les Professions)
- Citizenship
- Professional Certification Boards
- Colleges or Universities

At the same time, we note that the needs of the LINC program alone justify the test development we are recommending. The participation of other branches of government and other agencies, while highly desirable, is not essential.

3. Issues and Considerations

This section examines salient issues surrounding the development of an assessment system for CLB-based progress and outcomes. Care must be taken that scores from such an assessment are not misused (AERA, APA & NCME, 1999; Joint Advisory Committee, 1993), and existing confusion about the meaning of a benchmark must be addressed. Language milestones deemed important enough to justify the cost of an expensive test process must be identified and agreed-upon, and efforts made to steer and control the inevitable washback effects of such a test. Finally, high-level decisions are required about the manner in which each of the four language genres will be tested.

3.1 The Canadian Language Benchmarks:

An outcomes assessment for LINC and ELT programs must be based on the Canadian Language Benchmarks because both programs operate on curricula articulated to the levels and descriptors in the CLB document. It is essential that an outcomes assessment system be compatible with what is

taught in the classroom, and with what is specified in the CIC Selection criteria. Moreover, a made-in-Canada test would have greater face validity and garner more support than a test designed outside the country for purposes not related to LINC or ELT.

Development of the outcomes assessment would need to be undertaken by a team of experts who are familiar with the intricacies and challenges of working with the CLB document. Although the benchmarks concept has existed for some 14 years, even the most recent document, the CLB 2000, is somewhat inconsistent in its presentation of competencies for test development purposes. For example, gaps exist across performance criteria for the productive skills, which can create a challenge for anyone designing evaluation criteria for speaking or writing. In terms of task design and development, extrapolations and inferences need to be made in cases where the benchmarks do not fully account for certain key features of discourse that affect difficulty, such as, in the case of listening for example, volume, pitch, speed and background noise. A series of tables have been developed to identify and smooth such gaps with the objective of assisting users in applying the benchmarks to curriculum and test development (Nagy & Stewart, 2005), but a great deal of experience and expertise are still required to work with the CLB contents in a responsible way.

A test developer has to be prepared to deal in a defensible way with the challenges that the CLB document presents. Among the key challenges are identifying distinctions between benchmark levels, consistently addressing performance conditions, and defining the role of the task as it relates to the overall assessment. The CLB document is a continuum of learner proficiency, and as such, it describes the abilities that a learner possesses at each benchmark and the conditions under which these abilities can be successfully demonstrated. In other words, its main area of concentration is the content of the various cells at each CLB level. An assessment must necessarily focus on the boundaries between the cells in order to clearly differentiate one benchmark from the next. A test developer has to work on establishing these cell boundaries and reflecting them in the test results.

We note with great interest the recent announcement of a round of consultations by the Centre for Canadian Language Benchmarks (CCLB) intended to lead to possible revisions of the CLB 2000 by spring 2010. This initiative is well timed to coincide with the development of a LINC outcomes system, as the two undertakings can work together in a complementary manner. During the first round of national consultations, which led to the 1996 Working Document revisions, the CLBA test development team was able to provide feedback that became instrumental in shaping the content and theoretical design of the CLB 2000 document.

One example of this iterative process was the ongoing discussion surrounding the role of tasks in the assessment. In the early stages of conceptualizing the CLB continuum, it was not uncommon for practitioners to make unrealistic assumptions about the relationship between tasks and learner levels. At one of the National Working Group meetings, it was suggested that the test developers could determine whether or not a learner was at CLB Reading benchmark 5 by simply

administering a task that appeared in the CLB document at benchmark 5 and evaluating whether or not the person could complete that single task. The test developers had to explain that the tasks in the CLB document are illustrative in nature, meaning that they have not been empirically validated as uniquely representative of a particular benchmark, and that no one task could be considered as an indicator of benchmark level. In fact, if a panel of experts is asked to place a test item at a benchmark, their views vary, typically plus or minus one benchmark, sometimes more, so that an average of their placements needs to be taken. To create a fair assessment, it is necessary to sample adequately from the domain of behaviour, which usually means that a range of tasks should be presented.

As a result of discussions such as these, the conceptualization of tasks in relation to benchmarks was altered in the CLB document to match the model that was used for designing and scoring the CLBA test. Tasks came to be associated with stages of the benchmarks, and proficiency was defined as the relative ability to successfully complete a range of tasks at that stage (see page XII of the CLB 2000 for an illustration of this model). This is just one example of the ways in which a test development project can inform the revision of its underlying scale to improve accuracy and consistency. It would therefore be very beneficial if the LINC assessment development and the CLB revisions were to take place concurrently.

3.2 Respecting the Test Purpose:

In a language program such as LINC, three main assessment objectives – placement, progress, and outcomes - can be identified. The main purpose of a placement assessment is to allow administrators to move students into appropriate classes as smoothly and efficiently as possible. A placement instrument should also provide as much diagnostic information as is required by an instructor to make an initial determination of student needs. As previously indicated in this paper, the role of placement assessment is currently filled effectively by the CLBA.

A progress assessment serves the purpose of informing instructors about incremental learner gains on specific classroom goals and objectives. This type of assessment is very closely linked to course content and may include several different components, most of them informal, that together provide a profile of the learner's ability in various areas of interest. An outcomes assessment is usually more formal and standardized than either of the other two. It is administered at key points throughout a program to ensure that a student has achieved certain milestones that are common across all courses.

These three types of tests have very different purposes and specifications, and for this reason, they are not interchangeable. A placement test is too general for progress purposes and not rigorous enough for outcomes. A progress test is too cumbersome for placement purposes and not reliable

enough for outcomes. An outcomes instrument would serve well for placement, but it would not be appropriate for progress because its results ignore non-language related goals and fail to capture specific classroom learning objectives and incremental degrees of progress on specific tasks that might be of interest to instructors.

Every test is developed with its specified purpose in mind, and the guardians and administrators of these tests are obliged to state what uses of the resulting scores are appropriate. Despite this, different audiences for test scores will use them as they choose, often giving them extremely high importance and assigning to them more accuracy than can be justified. For this reason, it will be important to clearly define the parameters of the LINC outcomes assessment and to ensure that it is used responsibly for the purpose that it is intended. A test that is developed for LINC outcomes purposes should possess the necessary degree of reliability and validity and be kept as secure as is needed given the stakes associated with the outcomes.

We would like to caution against any temptation to measure learner gains by comparing placement and outcomes scores. A placement test differs from an outcomes test in intent and consequences. In terms of intent, the purpose of a placement test is to ensure that students can be directed, as efficiently and as conveniently as possible, to classes that meet their language needs and are consistent with their current abilities. If a program has a narrow range of classes to offer, then often all that is needed is a very quick and simple assessment that ranks students reliably. In programs where more classes exist and instruction is tailored to more specific needs, a longer and more diagnostic placement assessment may be preferred. The intent of an outcomes test is to determine whether or not a learner has mastered the language requirements of a particular class level. For this reason, an outcomes assessment must be linked in a very direct way to the curriculum and the teaching objectives. The result on a single outcomes assessment is not as fair and reliable an indicator of progress as a compilation of results on different kinds of activities, including classroom-based tasks.

A placement test is narrow and deep. It can focus on a very small number of outcomes, those that are pertinent to the context. The goal is to determine whether a student would benefit more from continued instruction on the goals of the present class, or from instruction in a different class, working on different goals, perhaps at a different level. Placement test results are often interpreted in light of other relevant information, including teacher judgment, non-test information such as writing samples, class sizes (and perhaps location), and even class schedule (to accommodate employment and daycare needs). They are usually much shorter, and scored on the spot by the instructor. In contrast, an outcomes test is broad and shallow. It must sample across the entire range of outcomes, in order to provide fair evidence of the student's ability across the spectrum of program goals. This results in too few items on any one goal to assess success on that particular goal. To produce accurate scores, outcome tests must include items that span a large range of

abilities, including some items that are quite easy and others that are quite difficult for a particular candidate.

In terms of consequences, the stakes are much higher with an outcomes test than with a placement test. Placement decisions affect only the class level in which a student is located, a consequence that is easily reversed if an instructor notices that the learner is too high or low for the level. With outcomes testing, the result often determines life-altering events, such as whether or not a student graduates or is considered to be job eligible.

In comparing placement and outcomes results, difference scores or gain scores are very unreliable, even if each score alone is reasonably reliable. As well, a system that rewards individuals for improvement should be approached with caution. Requiring both accurate pre-test and post-test scores will increase the cost of the system, and is prone to abuse (e.g., deliberately doing poorly on the pre-test). It is not unreasonable to use *group* average gains as *one* measure of program success. A pre-post difference in group average scores is far more reliable than an individual difference. However, some caution is still required, somewhat because of the potential for abuse, but more importantly, because of the unintended signals sent to those responsible for instruction.

As mentioned earlier, the focus of the LINC assessment should be on outcomes only, as placement (and internal promotion) procedures have already been established locally for the purpose of situating learners into the most suitable learning environment for their needs. Given the range of LINC program sizes, from small groups with limited placement options to very large groups with much more flexibility, and the relative ease of reversing a placement decision, national resources should focus on exit only.

In order to be effective, post-testing has to be mandated. At the present time, there are several instruments available for use as CLB-based outcomes indicators, but these are used sporadically and without consistency because there is no mandated requirement. If an outcomes assessment system is to be put in place, it will be necessary to also establish a procedure that ensures all LINC programs implement it.

3.3 Interpreting the Meaning of a Benchmark:

There appears to be a fundamental problem concerning differing interpretations of what it means to be “at” a benchmark. The CLB is a continuum of language ability based on a model of communicative competence (Canale & Swain, 1980). It is primarily a set of descriptors that describe what an ESL learner can do at each point along the language-learning continuum from beginner to advanced proficiency. Representative tasks appear throughout the CLB 2000 document as non-validated but illustrative samples of the kinds of things that a learner is able to do at each

benchmark. In using the CLB as a framework for test development, we interpret the sample tasks with caution, preferring to associate a particular task with a stage of the CLB (Stage I benchmarks 1-4, Stage II benchmarks 5-8, or Stage III benchmarks 9-12) or with an otherwise specified range of benchmarks rather than trying to peg specific tasks to dedicated benchmarks.

In setting up scoring procedures, it is important to pay particular attention to the global descriptors that indicate what a learner is able to do at each benchmark, and base calibrations on the assumption that a learner who places “at” a particular benchmark is able to successfully demonstrate the competencies described at that benchmark at least 75 to 80 percent of the time. What this means, in our estimation, is that a learner who places at benchmark 3 has met the requirements of that benchmark and therefore should enter a LINC class in which the competencies for benchmark 4 are introduced and taught.

This is clearly laid out in the points criteria for the Foreign Skilled Worker program (FSW), but CLB-based curriculum frameworks do not necessarily interpret the CLB levels in the same way. Learners who demonstrate on the CLBA that they have achieved the requirements of benchmark 3 are often placed into LINC levels where benchmark 3 competencies are taught and practiced. We have always maintained that this is a misinterpretation of the CLB continuum.

3.4 Identifying Milestones

In developing an assessment system for LINC outcomes, it would be beneficial to identify the thresholds or milestones at which outcomes are most important and meaningful. Test developers could then focus the assessment content and scoring procedures to ensure that these distinction points are reliable. It is beyond the scope of this paper to identify with any certainty what these thresholds or milestones might be. The developers of the LINC outcomes assessment system would need to include, as part of their background research and needs analysis, a methodology for capturing the benchmark levels around which the greatest emphasis currently exists. In addition, it would be advisable for them to also identify those points on the CLB continuum where a growing interest is likely to develop in the future. Data for this research might be gathered from a number of stakeholder sources, including the CIC Selection Branch points system. For example, on their 100-point scale for immigration, the Selection Branch awards 1, 2 or 4 points (per genre, first official language) for having completed Benchmarks 3, 5 and 7 (working on 4, 6 and 8) respectively. The CEC (Canadian Experience Class) system uses a variety of criteria depending on the occupation category. Using these data along with information gathered from LINC stakeholders, those responsible for developing the assessment would need to determine at which benchmarks accurate outcome information would be most useful.

3.5 Approaches to Skill Testing

This section examines the most prevalent approaches to assessment of the four language skills – Speaking, Listening, Reading, and Writing. A distinction is drawn between more formal standardized approaches and the less formal, anecdotal methods of gathering information on learner progress.

3.5.1 Formal Assessment

The most widely used formal CLB-based assessments employ objective measures for the evaluation of reading and listening. This approach enhances reliability and ensures that large-scale scoring can be done quickly and efficiently. Objective measures include the use of multiple-choice and other dichotomously scored items that have clear right and wrong responses.

- Reading tests typically consist of several passages of various genres at increasing degrees of length and complexity. Each task or article is followed by a series of multiple-choice items that tap understanding of the passage. It should be mentioned here that multiple-choice questioning does impose some limits on what kinds of outcomes can be assessed. Recognizing a correct answer is not the same as generating it in one's own words. However, the consensus seems to be that this limitation is a small price to pay for the huge cost advantages and the high reliability achieved by using this format. Equally important to consider is the fact that other more open-ended methods run the risk of confounding reading proficiency with either speaking or writing. This has implications for validity and ultimately impacts the perceived fairness of an assessment.
- Listening tests typically involve several stimuli or passages of different genres drawn from CLB 2000 descriptors. Passages aimed at the Stage 1 benchmarks tend to be video-mediated so that learners can have the benefit of facial expressions, gestures, contextual clues, and other paralinguistic features that enhance comprehension. Passages for Stages II and III may be audio-mediated because learners at these levels are expected to comprehend information with fewer contextual cues. The passages are typically longer and more complex, including a variety of genres that are selected to match the needs of the intended audience. Genres for Stages II and III may include news bulletins, lectures, presentations, formal and informal dialogues and conversations. Passages are typically followed by the objective items similar in type and format to those appearing in the reading tests. The same limitations of the multiple-choice format just mentioned apply equally to listening, but again, other methods of data capture would confound the target proficiency with speaking or writing. The benefits of multiple-choice seem to outweigh the disadvantages.

Existing CLB-based assessments of speaking and writing are usually performance oriented. Learners are expected to write various types of discourse to accomplish real-life tasks and

objectives. In speaking, they are expected to interact with a live interlocutor who both facilitates the exchange of information and evaluates the learner's spoken discourse.

- Writing tests are typically performance-based, involving the collection and evaluation of direct learner samples. All scoring tools and procedures have to be very carefully constructed and monitored to minimize the error associated with the subjectivity of the scoring. Less direct (and much less expensive) methods of administration and scoring exist, such as multiple-choice tests that tap a learner's knowledge of language rather than his or her actual usage. These indirect methods include items that require a candidate to select the best word or phrase to fit into a sentence, find the error in a sentence, choose the sentence that does not belong in a paragraph, identify the incorrectly spelled word, and so on. However, these all have serious validity flaws, and the consensus is that direct sampling, even with the associated costs and challenges, is preferable. We note that scoring of writing (and speaking, below) can be made as reliable as multiple-choice scoring.
- Speaking tests that involve a direct interview seem to be the best approach for CLB-related assessment. A live assessor can first put the learner at ease by initiating some small talk to get to know the person and can then facilitate the assessment so that it unfolds as naturally as is possible in a testing situation. At Stage I of the benchmarks, this is particularly important, as lower-level learners need to be able to ask for clarification and repetition. It is possible that a "live" assessment could be delivered by computer, with the assessor and learner interacting through video technology, and this option should perhaps be explored for some LINC administration situations. In any case, Speaking assessments should be conducted in accordance with a strict written protocol to ensure consistency of administration and accuracy of scoring. Rigorous and comprehensive training methods are essential to ensuring that assessors have the qualifications, experience, and confidence to conduct a live speaking assessment. A well trained and experienced assessor knows how to set up a prompt to elicit a speaking sample in the form of a short or a longer turn and how to then direct the conversation, according to the protocol, to ensure that the upper limit (or threshold) of speaking ability is challenged and observed.

3.5.2 Informal Assessment

The above are the most commonly used CLB-based assessment approaches, but these are not the only available options. It is perhaps worth giving some consideration to the less formal methods, such as portfolio assessment. Portfolio assessment, whereby candidates collect and present evidence of their proficiency or accomplishments, has recently gained currency in educational circles, and in language assessment (for example, Moya & O'Malley, 1994). However, this approach is intended primarily as an instructional tool, and the difficulties in making it precise and

reliable enough to produce accurate scores are well-documented (Delandshere and Petrosky, 1994; Schutz and Moss, 2004).

Nevertheless, portfolios can still be very useful and motivating as tools for informal outcomes assessment in the classroom. The advantages of a portfolio system have to do largely with student empowerment, self-assessment, internal motivation and detailed diagnostic record keeping. Two examples, Manitoba and Ireland, show their utility.

The province of Manitoba uses portfolio assessment in their language training system. They have developed a detailed rubric system and recording forms for instructors to assess reading, using the CLB and the CLB 2000 Companion Tables. They have also made a beginning on developing guidelines for teachers to use in documenting student proficiency in all four genres. Their work is linked to the European Language Portfolio (ELP). The Manitoba portfolio system leads to some degree of standardization across teaching environments, but not enough for high-stakes assessment, for which it was not designed.

Ireland offers a portfolio program, built on the ELP, as part of its adult ESL program. The ELP itself, developed in the European multilingual context, has three obligatory components: a language passport, a language biography, and a dossier. The first two are the individual's language history, and the third a place to put evidence of progress in language and intercultural experience.

The Irish in-service examples are can-do statements copied from the CEFR (Common European Framework of Reference). Language schools are accredited to use the portfolio, and then student performance is assessed locally. The primary use seems to be motivational and curricular. It is a learner-centred, individualized program. There is no information on how difficult it is to get into higher education based on these certificates. Any mention of testing as such seems to be for placement. Note that the Irish program, offered in only nine locations, is approximately 1.5% the size of the Canadian program. We will have more to say below about the *informal* use of portfolios in the section titled Informal Assessment Component.

While portfolios are intended primarily for local use, it is not inconceivable that they could be adapted to an external accreditation system by use of auditing procedures. However, this would require a level of intervention that would not be worth the cost, and it may not result in a system judged credible by all target audiences. We remain convinced that the basic model of multiple-choice for reading and listening, and direct sampling for writing and speaking, are best in the Canadian context.

However, there is a price to pay for this system. Barbara Lazenby Simpson, one of the principal actors in the Irish system expresses the following concerns about external assessment. "The obvious success of the pedagogical approach should not (1) be overturned by the stress of formal

assessment, (2) impose unrealistic types of assessment on individuals, (3) influence the delivery and content of teaching, and (4) ignore the crucial importance of access to the host society through emphasis on language proficiency only.” The next section addresses these concerns.

3.6 Wash-back

Any outcome assessment will have an impact on classroom instruction, and this impact needs to be considered in the development of an assessment system. LINC, as an integration program, has a number of goals that go beyond language *per se*. It is often not possible, and indeed not desirable, to include such goals in a high-stakes assessment. Consider, for example, the inappropriateness and impossibility of designing a test item to judge whether an immigrant from a different culture accepts the values that most Canadians accept.

Any language test will exclude important LINC goals. Thus, there is a danger that an externally imposed test will turn the *most important* goals (language-related outcomes) into the *only* goals of a classroom program. Steps need to be taken to ameliorate this problem.

- Centrally developed curriculum materials should provide support to instructors working to achieve such goals.
- Testing materials can be set in the context of, for want of a better term, Canadian *civics*.
- Every effort should be made to include achievement of such “soft” goals in program assessment, but not in individual assessment.
- We recommend the Manitoba materials as an excellent starting point for improving local promotion procedures to make them more thorough and consistent.

3.7 Test Administration

Administration procedures for the outcomes test should be as efficient, secure, and reliable as possible. This means that the assessment cannot be administered by teachers who know the students they are testing, or the teachers of such students.

Our working assumptions are that reading and listening tests will be multiple-choice and machine-scorable, while writing tests will require samples scored by judges, and speaking tests will require a face-to-face interview. High-speed internet access will allow secure administration of reading and listening tests anywhere in Canada. Local assistance will be needed only for identity checks, and, for a very few candidates, over-the-shoulder support for computer use. Writing samples, whether entered by keyboard or hand-written, can be captured electronically and sent anywhere in the country for scoring. Similarly, face-to-face interviews can be captured on web-cams, or some such device, and transmitted elsewhere for auditing. (There is a well-known problem that third parties

viewing recorded interviews give candidates different speaking scores than the original interviewer does. This is because the camera cannot capture all the nuances of eye contact and body language. Thus, the original interviewer must remain the prime score giver, with a supporting audit process for quality control).

There is no reason why LINC teachers across the country cannot form the backbone of the scoring system under the right circumstances. In fact, such an undertaking would have positive professional development value.

4. Developing the Assessment Procedure

Respecting considerations of cost and efficiency, we posit a two-pronged approach to improving the LINC progress and outcomes assessment system. The first prong would involve a more informal component that would support ongoing diagnosis and promotion, and possibly program evaluation, while the second prong would comprise a more formal standardized assessment procedure designed to ensure the appropriate levels of validity, reliability and quality required for high-stakes purposes.

4.1 The Informal Assessment Component

The informal assessment component should combine the best features of classroom outcomes assessment. The procedure should be administered and scored by instructors within the context of delivering their courses. It should be comprehensive yet flexible, allowing multiple opportunities for students to demonstrate various abilities on different types of tasks. The overall result would be a compilation of materials and assignments that demonstrate the students' progress as they internalize concepts and complete their course work.

Within the LINC outcomes system, the informal assessment is the component that provides the strongest association between what is taught in the classroom and what is ultimately evaluated and therefore valued. It is in this assessment that non-language goals can be emphasized, not with a view to grading individuals, but with a view to signalling their importance as part of the process of learning how to function successfully in the Canadian social and occupational context. The assessment can be much more diagnostic than a formal outcomes assessment and can indicate much finer distinctions, so that incremental degrees of progress can be demonstrated and evaluated. Moreover, an informal assessment often reflects a much higher degree of face validity than its formal counterpart. This is because a formal assessment must necessarily include tasks and items that lend themselves to the most efficient, reliable, and often objective scoring procedures. In a formal assessment, technical validity must take precedence over face validity. Informal assessments can embrace the full range of authentic task performance, thereby enhancing face

validity. An assessment with a high degree of face validity is usually well received and accepted as an authentic and meaningful measure of the kinds of tasks that students are expected to perform in their LINC classes and in the real world. An informal assessment system that includes a portfolio component can be particularly beneficial both for students and for instructors. It is possible for students to make a contribution to the design of such an assessment by suggesting meaningful tasks and activities that could be included.

The design of the informal assessment should take into account the concept of “bias for best” (Swain, 1985), which emphasizes the importance of providing the right tools and circumstances to ensure that students are able to demonstrate the full extent of their capabilities. Adhering to a “bias for best” approach might involve such test development strategies as finding ways to enhance the relevance and meaning of the assessment tasks while allowing students to work in ways that suit their individual learning styles and creative strengths.

The developer of the portfolio assessment should begin by conducting a thorough review of the work that has been done in the field in the area of CLB-based classroom outcomes. The review should include, but not be limited to, the following resources:

- Centre for Canadian Language Benchmarks
 - Summative Assessment Manual
 - CLB 5-10 Exit Tasks
 - Integrating CLB Assessment into Your ESL Classroom
- Manitoba Labour and Immigration Adult Language Training Branch
 - Collaborative Language Portfolio Assessment: Manitoba Best Practices Guide
 - Writing Rubrics for Outcome Assessment
 - Reading Task Outcome Assessment

These resources provide excellent background information for the development of an informal assessment procedure. The CCLB materials include sample tasks and scoring procedures, along with suggestions for development and implementation. The Manitoba materials emphasize the key principles that apply to successful portfolio assessment. They suggest beginning with a needs assessment, a language assessment statement, a goal statement, a checklist of intended CLB outcomes, and samples of learner performance on entry. The portfolio is then built by assembling a range of materials, which are added at a set time each week or month. These materials might include audio tapes, a vocabulary log, completed reading and listening tasks, sample of writing of different types and genres, a daily journal, and classroom tests or dictations. Throughout the course, the portfolio is used to document even the most incremental progress, and the contents are discussed by the teacher and learner at set intervals as the student progresses. At the end of the course, the student is given the portfolio, or it is passed along to the next teacher (Manitoba Labour and Immigration, 2004).

The estimated time and cost for creating an informal classroom assessment approach are:

Project Activity	Description
Preparation	Recruit team Develop workplan and allocate resources
Background research	Conduct research on current practices Conduct needs analysis to determine assessment fit
Assessment design	Design the assessment components Draft specifications for the types of materials to be included in the assessment based on the CLB 2000
Develop procedures	Determine how to gather data Prepare a draft procedure that indicates how progress and outcomes will be measured and/or reported
Field testing	Select representative programs and field test the procedures with instructors and learners Elicit feedback and ideas
Revisions	Revise procedures based on feedback
Prepare materials	Prepare materials Have components produced and duplicated
ESTIMATED TIME	52 - 60 weeks
ESTIMATED COST	\$350,000 - \$475,000

4.2 The Formal Standardized Assessment Component

The formal outcomes procedure should be based on the Canadian Language Benchmarks and should build on previous work that has been carried out in the area of CLB-referenced task-based assessment. It would be ideal if the procedure could be tailor-made for its intended purpose and also related in principle to other high-quality, high-stakes CLB-referenced tools. It should be administered according to a standardized procedure by an assessor who is not also the student's instructor.

The design and development of the formal assessment could be approached as a completely new initiative or it could be organized so that it builds on CLB-related work that has already been completed and validated. In either case, the resulting product needs to meet a number of requirements, which include:

- Demonstrated relationship to the CLB 2000 and its successor
- Criterion referencing that shows more than a high correlation between test scores and the CLB but clearly demonstrates that test scores place candidates at a particular benchmark
- Reliability evidence to show that scores are accurate and repeatable
- Documented due diligence and adherence to accepted practices for test development
- Clear evidence of the technical quality and validity of the assessment system
- Items and tasks that lend a high degree of face validity to the test
- Evidence that the team developing the assessment, along with their consultants, advisors, reviewers, proctors and assessors, are highly qualified in the areas of assessment and CLB-related research and development
- Item and test calibration procedures should follow accepted methods, ensuring high accuracy and low standard errors at all important score junctures.
- Test administration procedures should ensure that all candidates have a full opportunity to demonstrate their skills, including accommodation to special needs where appropriate.
- Test security procedures should be spelled out, including ongoing plans for development of multiple-forms, to ensure that scores are valid reflections of candidates' levels of skill.

A new initiative would necessitate creating an assessment from scratch, using CLB-based specifications and working in consultation with LINC experts, key stakeholders, and members of the ESL field. The second option would involve adapting an existing CLB-referenced test, an approach that is presented here because it would be cost-effective, efficient, and reliable. In many ways, it may be considered prudent to build on existing assessment infrastructure rather than continually trying to invent new procedures.

There are very few existing assessments that could lend themselves to adaptation for LINC purposes. In our research, we have considered the Canadian Test of English for Scholars and Trainees (CanTEST), the CLBA, the CLBPT, and the Canadian Language Assessment International (CLBA-I) in light of the criteria that would need to be met to ensure a high-quality outcome. Our reference for this analysis was the set of criteria used by the Selection Branch to judge whether tests are suitable for use in allocation of points for immigration. These criteria are currently under revision, so our use has been confined to broad brushstrokes only. See the Appendix for details of our analysis, which indicates that of these four existing instruments, only the CLBA-I fulfils the necessary criteria for adaptation.

The estimated time and cost associated with a full test development project are as follows:

Project Activity	Description
Preparation	Recruit and hire project team Develop a detailed workplan and allocate resources
Needs analysis	Conduct Analysis of LINC curriculum Survey LINC stakeholders Determine milestones/thresholds and LINC outcomes
Design test model	Design assessment model Create CLB-based specifications Outline task and item requirements
Item and task writing	Create a battery of tasks and items
Field testing and revision	Field test tasks and analyze results Revise test procedures and content accordingly
Prepare for pilot testing	Design pilot study Prepare pilot test forms Train assessors as required
Pilot testing	Conduct pilot testing Gather all data and code if necessary
Pilot data analysis	Analyze data and interpret results
Prepare operational test	Prepare training materials and technical reports Prepare test forms Prepare promotional and information packages
ESTIMATED TIME	70 - 94 weeks
ESTIMATED COST	\$1,200,000 - \$1,400,000

The estimated time and cost associated with adapting an existing assessment are as follows:

Project Activity	Description
Preparation	Recruit and hire project team Develop workplan and allocate resources
Content review	Assemble a panel of LINC experts Gather feedback on content of existing test
Needs analysis	Analyze LINC curriculum and survey stakeholders Determine milestones/thresholds/ outcome needs
Create item writing specifications	Determine content to be retained and replaced Create specifications for tasks and items
Item and task writing	Create tasks and items for field testing
Field testing and revision	Field test tasks and analyze results Revise procedures and content accordingly
Prepare for pilot testing	Design pilot study and construct pilot test forms Coordinate pilot sites and train assessors as required
Pilot testing	Conduct pilot testing Gather all data and code if necessary
Pilot data analysis	Analyze data and interpret results
Prepare operational test	Prepare training materials and technical manuals Prepare test forms Prepare promotional and information packages
ESTIMATED TIME	48 – 52 weeks
ESTIMATED COST	\$650,000 - \$750,000

All time and cost estimates included in this paper are necessarily broad and would of course vary depending on a test development team's approach and methodology. The steps shown may appear to be very simple and straightforward to a reader not familiar with test development, but in fact, each cell in the above charts represents a great deal of complicated and time-consuming work. For example, the item and task writing step alone involves recruiting and training a team of writers, orienting them to the test model and specifications, calibrating them on the CLB benchmarks, overseeing the writing process, revising, refining, and assembling the resulting work. This level of detail has not been provided for each step in the chart, as the way in which these steps are managed may vary based on an individual test developer's approach. Suffice to say that a test development project is a relatively costly and time-consuming undertaking because it is real research. It begins with a theoretical model and a draft design, and from that point, it is informed by a process of gathering and analyzing data, both qualitatively and quantitatively. From the outset, the test development team needs to have a strong plan in place but also must be sensitive and responsive to the research process, prepared to move in new and unexpected directions to achieve the best and most defensible final results.

Given the high profile of LINC programming, we are assuming that a large number of stakeholders across the country will need to be involved in the consultation and review process. The pilot design should also take into account the national scope of LINC to ensure an accurate representation of learner demographics. These considerations add time and cost to a project.

The steps shown in the above charts are those that represent generally accepted standard practice for responsible test development. Further elaboration on the procedures and on the time required to complete each of the steps would be a consideration for an individual test developer, based on their intended design and approach. For further details on standard test development and validation procedures, see Weir (2005) or Bachman and Palmer (1996).

5. Implementing the Assessment Procedure

A test development undertaking of this scope requires a sustainable management infrastructure which can only be ensured by means of ongoing financial support. Since Canadian language testing needs are not large enough for such an infrastructure to be entirely market-supported, there would need to be financial assistance from government sources. Specific details pertaining to the location(s) and structure of the organization would need to be discussed and negotiated during the test development timeframe.

The main staff of the organization would include:

- A CEO
- One or more full-time project managers
- Clerical support;
- A chief test developer
- A chief statistician
- A chief IT specialist
- Part-time contractual writers, experts, and AV specialists

The following chart outlines some activities that would be necessary in the first year of operation.

Project Activity	Description
Establish the Organization	Endorse an existing organization or create a new entity (which could have a name such as the CCLA - Canadian Centre for Language Assessment) Establish the operational model and administrative infrastructure Recruit and hire experienced personnel
Create a Business Plan	Develop a detailed business plan Estimate ongoing funding requirements
Begin Operations	Design and commence test operation and administration, including ongoing development, validation, management, maintenance, quality control
ESTIMATED START-UP TIME	48 weeks
ESTIMATED ANNUAL SALARIES and BENEFITS	\$1,200,000
OTHER COSTS IN THE FIRST YEAR	\$1,300,000
TOTAL COSTS IN THE FIRST YEAR	\$2,500,000

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aydemir, A. & Skuterud, M. (2005). Explaining the Deteriorating Entry Earnings of Canada's Immigrant Cohorts: 1996-2000. Statistics Canada.

Bachman, L.F. & Palmer, A.S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

Boyd, M. (1992) Gender issues in immigration. In B.R. Chiswick (Ed.), *Immigration, language and ethnicity: Canada and the United States*. (pp. 305-372). Washington, D.C.: The AEI Press.

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

Centre for Canadian Language Benchmarks. (2007) *CLB 5-10 Exit Tasks*. Ottawa.

Chiswick, B.R. and Miller, P.W. (1992). Language in the immigrant labour market. In BR. Chiswick (Ed.), *Immigration, language and ethnicity: Canada and the United States*. (pp. 229-296). Washington, D.C.: The AEI Press.

Chiswick, B.R., & Miller, P.W. (2003). The complementarity of language and other human capital: immigrant earnings in Canada. *Economics of Education Review* 22 (2003), 469-480.

Chiswick, B.R., & Miller, P.W. (1995). The endogeneity between language and earnings: International analyses. *Journal of Labor Economics* (13), no. 2

Conference Board of Canada (2004). *Performance Potential: How can Canada Prosper in Tomorrow's World?*

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge. *Educational Researcher*, 23 (5), 11-18.

Holmes, T. (2005). *Integrating Assessment into Your ESL Classroom*. Centre for Canadian Language Benchmarks, Ottawa.

Human Resources and Skill Development Canada (2007). *Looking-Ahead: A 10-Year Outlook for the Canadian Labour Market*.

[http://www.hrsdc.gc.ca/eng/publications_resources/research/categories/labour_market_e/sp_615_10_06/page04.shtm!](http://www.hrsdc.gc.ca/eng/publications_resources/research/categories/labour_market_e/sp_615_10_06/page04.shtm)

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. (http://www.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf)

Locchhead, C., & Mackenzie, P. (2005). Integrating newcomers into the Canadian labour market. *Canadian Issues* (Spring 2005), 103-106.

Makosky, L. (2008). *The Feasibility and Way Forward for a Standardized Exit Assessment and Test for Newcomers in LINC Training*. Citizenship and Immigration Canada.

Manitoba Labour and Immigration Adult Language Training Branch (2004). *Collaborative Language Assessment: Manitoba Best Practices Guide*. Winnipeg, Manitoba.

Manitoba Labour and Immigration Adult Language Training Branch (2008). *Reading Task Outcome Assessment*. Winnipeg, Manitoba.

Manitoba Labour and Immigration Adult Language Training Branch (2006). *Writing Rubrics for Outcome Assessment*. Winnipeg, Manitoba.

Metropolis Secretariat. Citizenship and Immigration Canada. (2006). *Final Report: G8 Experts Roundtable on Diversity and Integration* (October 4, 2006. Lisbon, Portugal).

Moya, S.S. and O'Malley, J.M. (1994). A portfolio assessment model for ESL. *The Journal of Educational Issues of Language Minority Students*, 13, 13-36.

Nagy, P. & Stewart, G. (2005). *Canadian Language Benchmarks 2000 Companion Tables*. Centre for Canadian Language Benchmarks, www.language.ca.

Pawlikowska-Smith, G. (2000). *Canadian Language Benchmarks 2000: English as a second language for adults*. Centre for Canadian Language Benchmarks, Ottawa.

Pawlikowska-Smith, G. (2008). *Summative Assessment Tasks, volumes 1 & 2*. Centre for Canadian Language Benchmarks, Ottawa.

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

Pendakur, K., & Pendakur, (Nov. 1997). Speak and ye shall receive: Language knowledge as human capital. RIIM (Research on Immigration and Integration in the Metropolis).

Schutz, A. M. and Moss, P. A. (2004). "Reasonable" decisions in portfolio assessment: Evaluating complex evidence of teaching. *Educational Policy Analysis Archives*, 12 (33).
[<http://epaa.asu.edu/epaa/v12n33>]

Stewart, G. (2005). Issues in Task-based Assessment of Communicative Competence. *TESL Special Research Symposium Issue*, 31, 2, 27-42.

Stewart, G, Geraci, K, Nagy, P. (2004). *Essential Skills and Canadian Language Benchmarks: Considerations in comparing conceptual frameworks*. Ottawa: Centre for Canadian Language Benchmarks.

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.

Weir, C.J. (2005). *Language Testing and Validation*. MacMillan: Palgrave.

Appendix: Analysis of Assessments for Adaptation Purposes

CanTEST:

The purpose of the CanTEST is to indicate whether a person has enough English proficiency to engage in university studies or professional activities. As such, although it does adhere to the high-stakes requirements that may be associated with LINC outcomes, its content is focussed on the higher degrees of proficiency associated with post-secondary and professional pursuits. This range of focus would be too narrow to fully address LINC outcomes. An even greater concern with the CanTEST is the fact that it was not developed according to specifications drawn from the CLB. In other words, it is not a CLB-based instrument. Though its results have been aligned to the CLB, this was done by means of retrofit. A test that has been retrofitted to the CLB is not the same thing as a CLB-based instrument. In addition, the development procedures used for the CanTEST do not conform to the requirements in the Selection Branch criteria. For these reasons, the CanTEST would not be a suitable starting point for adaptation.

CLBA:

In considering the CLBA, we first note that it is a CLB-based assessment, but its specifications have been drawn from a previous version of the benchmarks document, not from the CLB 2000. This is not a problem for placement purposes as programs tend to place students on the basis of one or two language skills and most administrators have found satisfactory ways to work with this test, but it would pose challenges for test adaptation in the present context. The CLBA model requires that an assessor score Stage 1 of the Reading and Writing components before the learner can go on to Stage 2. This would create administrative difficulties in an outcomes context. In addition, the CLBA only reports results to a maximum of benchmark 8. While this range might be sufficient for current LINC outcomes, it would be preferable to have an assessment that would reach into the advanced levels of CLB Stage 3. In light of these observations, and given the fact that the CLBA has not been validated for high-stakes purposes, including those of the Selection Branch, we are led to conclude that it would not be feasible to adapt this instrument for LINC outcomes.

CLBPT:

The CLBPT exhibits many of the same disadvantages as the CLBA. Its purpose is low-stakes classroom placement, and its reported scores do not extend beyond benchmark 8. While the CLBPT does render separate scores for Speaking and Listening, its administration integrates the two skills. The Listening results for this assessment are particularly problematic, as there are very few items on which to base an evaluation. Given its short length and lack of rigour, this test simply does not have the validity or reliability to serve as a model for outcomes.

CLBA-I:

Finally, we come to the CLBA-I, which was originally developed for the high-stakes purpose of CIC selection. This test is based on the CLB 2000 and includes a separate instrument to measure each of Speaking, Listening, Reading and Writing. Of the assessments considered, this is the only one that fulfils the criteria for adaptation. One of its best features is the fact that it is a derivative of the CLBA. This means that it is linked in theory and content to the test that is currently used for LINC placement purposes. CLBA-I results are reported in the range of benchmarks 3 to 9, which would be a sufficient range for LINC outcomes purposes, and the test procedures have been validated for high-stakes purposes on a large sample of ESL learners in Canada and off-shore in India and Korea. Though the test has been endorsed for use with the Foreign Skilled Workers (FSW) group for the Canadian immigration points system, it has not yet been used for its intended purpose. Its relatively generic content makes it a solid foundation for multi-purpose adaptation. In fact, if the tests for FSW and LINC were related through an adaptation procedure, it would be greatly advantageous. The FSW scale program differentiates at CLB benchmarks 3/4, 5/6 and 7/8, that is for having “completed” benchmarks 3, 5 or 7. These distinctions could relate to the thresholds or milestones identified for LINC progress and outcomes.