Sustaining an Occupation-Specific Language Assessment for the Canadian Healthcare Field

Gail Stewart & Andrea Strachan

Since its implementation in 2004, the Canadian English Language Benchmark Assessment for Nurses (CELBAN) has been accepted as evidence of language ability for licensure of internationally educated nurses (IENs) in Canada. This article focuses on the complexities of sustaining an occupation-specific assessment over time. The authors reference the seminal work of Epp and Lewis, who developed the original CELBAN test forms and aligned the test results with the Canadian Language Benchmarks (CLB), and then go on to describe a research and development project that was carried out under the direction of Touchstone Institute and overseen by the Centre for Canadian Language Benchmarks (CCLB) to renew the test model and develop additional content. This is followed by a discussion of the maintenance strategies required to sustain a secure assessment within the evolving Canadian context.

Depuis sa mise en place en 2004, le Canadian English Language Benchmark Assessment for Nurses (CELBAN) a été accepté comme preuve de compétence linguistique pour l'obtention du permis d'exercer au Canada pour le personnel infirmier formé à l'étranger. Cet article porte sur les complexités liées au maintien d'une évaluation propre à une profession au fil du temps. Les auteurs font référence au travail précurseur d'Epp et Lewis qui ont mis au point les formulaires du test CELBAN original et aligné les résultats du test avec les niveaux de compétences linguistiques canadiens, ensuite ont décrit un projet de recherche et de développement qui s'est effectué sous la direction du Touchstone Institute et a été supervisé par le Centre des niveaux de compétence linguistique canadiens pour renouveler le modèle de test et mettre au point des contenus supplémentaires. Cet article est suivi d'une discussion des stratégies d'entretien nécessaires pour maintenir une évaluation sûre dans le contexte évolutif canadien.

The Canadian English Language Benchmark Assessment for Nurses (CELBAN) is an occupation-specific measure of communicative competence in English as a second language (ESL). It was developed by Epp and Lewis (2004b) of the Red River College Learning Centre in a research project

overseen by the original sole owner of the test, the Centre for Canadian Language Benchmarks (CCLB). CELBAN was administered and managed by the Canadian English Language Assessment Services (CELAS) Centre at Red River College from 2004 to 2014, with the test results accepted by nursing regulators across Canada as evidence of language proficiency for internationally educated nurses (IENs). In 2014, Touchstone Institute assumed responsibility for CELBAN administration, and in 2020 also assumed partial ownership of the test, along with the CCLB.

When a test is implemented, a complex infrastructure is required to sustain its ongoing usage and maintain its security and validity. As the assessment context shifts over time, continuous attention must be paid to influences that can impact sustainability. For CELBAN, these influences have included changes in nursing regulatory policy, fluctuations in administration volume, revisions to the Canadian Language Benchmarks (CLB) document and, most recently, restrictions imposed by the COVID-19 pandemic.

This article traces influences that have prevailed across the stages of CELBAN development, implementation, renewal, and maintenance, with the aim of shedding light on the responsive research and development activity required to maintain the viability and integrity of an occupation-specific assessment.

Foundational Research and Development

The concept of a Canadian nursing-specific language assessment arose in the early 2000s, largely due to labour market growth in Canada attributable to increased immigration and federal and provincial government initiatives aimed at expediting the integration of immigrants into the Canadian economy (Alboim, 2002; Blythe et al., 2006; Johnson & Baumall, 2011; Russell et al., 2009). A central critique of professional credentialing systems in general has been the use of language proficiency tests that do not specifically measure the communication competencies most relevant to professional practice (Austin et al., 2003; Jeans et al., 2005; Strachan, 2007). Within this context, a project was initiated to investigate the possibility of a Canadian nursing-specific language assessment that would address concerns raised by nursing regulators, IENs, immigrant settlement officers and language training professionals about the critical shortage of nurses in Canada and the limitations of large-scale general proficiency tests to effectively measure communication for the healthcare professions. (Epp & Stawychny, 2002).

Identifying Proficiency Levels for the Nursing Profession

The first phase of the project was devoted to identifying the proficiency levels required to successfully negotiate nursing communication tasks in the four language skills—speaking, listening, writing, and reading. To

carry out this research, Epp and Stawychny used the Canadian Language Benchmarks (CLB) to analyze nursing tasks and relate them to ESL ability levels in a process that the researchers referred to as "benchmarking" (Epp & Stawychny, 2002). The CLB framework was particularly suited to this type of occupational language analysis, as it represents an underlying scale that spans 12 levels of ESL ability in each of the four language skills. These ability levels, or benchmarks, are fleshed out in a document (Centre for Canadian Language Benchmarks, 2000, 2012) that provides descriptors of ability at each CLB level, along with examples of tasks that can typically be performed by ESL users at that level, or benchmark. Presenting "a clear hierarchy, or a progressive continuum of language knowledge and skills" (Centre for Canadian Language Benchmarks, 2000, p. viii), along with illustrative communication tasks within community, academic and workplace contexts, the CLB 2000 document proved to be a flexible resource for the identification of ability levels for the nursing profession.

The project leads conducted clinical observations and worked with a team that included academics, registered nurses (RNs), registered practical nurses (RPNs), registered nursing assistants (RNAs) and licensed practical nurses (LPNs) to identify the specific communication skills required for nurses to function successfully in practice within Canada's healthcare system. The tasks that were identified during the clinical observations were charted in reference to CLB levels based on the language skills required, the conditions of performance, and specific features of the interactions. By comparing the characteristics of the tasks to descriptors in the CLB document, the researchers were able to assign a CLB level to each of the nursing tasks.

A key finding that emerged from this CLB-referenced research study was a reinforcement of the recommendation for a language assessment specific to the nursing profession. As previously mentioned, nursing regulators had some doubts about the suitability of general language proficiency assessments for identifying communicative ability specific to the healthcare professions. They were also aware of the existence of occupation-specific tests, such as the Midwives' Language Proficiency Test (Mendelsohn & Stewart, 1999) and the Occupational English Test for Nurses (McNamara, 1990), which were being used successfully to measure communication for healthcare contexts. They believed that a nursing-specific test referenced to the CLB would provide a valid indicator of the language proficiency required to function in the Canadian nursing profession.

Development of CELBAN

Building on the findings from the CLB-referenced research study, test developers Epp and Lewis (2004b) undertook a project to design an assessment that would measure language ability for the nursing profession in each of the four separate communication skills. The resulting CELBAN test

instruments included authentic task types that had been identified during the clinical observations. These tasks reflected descriptors in the CLB document (CCLB, 2000) and were scored according to the levels on the CLB scale.

Extensive validation research was carried out by the test developers, including broad consultations with a wide range of assessment and subject-matter experts, pilot testing of tasks and items, reliability analysis, and comparisons of CELBAN with other relevant language assessments (Epp & Lewis, 2004c). Based on this research, the developers were confident that CELBAN results accurately represented CLB levels, and the test was implemented with the following required scores based on recommendations from the CLB-referenced research study:

- Speaking CLB 8

- Listening CLB 9

- Writing CLB 7

- Reading CLB 8

CELBAN was adopted by nursing regulators as proof of language ability for licensure due to the clear linkage of its content to nursing communication requirements (Jeans et al., 2005), and information was shared with the ESL field in a series of documents and articles that chronicled the initial research, test development and follow-up activities (Epp & Lewis, 2004a, 2004b; Lewis & Kingdon, 2016). The remainder of this article is intended to build on that narrative by informing the field about further research and development aimed at sustaining the test and responding to changing circumstances.

Post-Implementation Influences on CELBAN

Shifts in the Nursing Regulatory Context

After CELBAN's introduction, stakeholders continued to call for improvements in the ways in which IENs were being integrated into the Canadian healthcare system. The use of paper-based credentials began to be questioned, as a new focus on performance and competency-based evaluation emerged, along with significant changes in IEN credential recognition and within the nursing profession itself. The Canadian Nurses Association Position Statement in Interprofessional Collaboration (2010) stressed the importance of professional collaboration, client-centred care, evidence-informed decision making, ethics and communication. This position translated into changes in the ways in which professional competencies were observed and measured. In 2013, there was a switch from the Canadian Registered Nurse Examination (CRNE),

administered nationally by the Canadian Nurses Association (CNA), to the American-developed National Council Licensure Examination-Registered Nurse (NCLEX-RN).

At the same time, new registration requirements came into effect requiring a baccalaureate degree, such as a Bachelor of Science in Nursing (BSN), Bachelor of Nursing (BN) or equivalent. For IENs, this meant a new competency evaluation system in addition to the nursing qualifying exam. The College of Nurses of Ontario (CNO) engaged the Centre for the Evaluation of Health Professionals Educated Abroad (CEHPEA), now Touchstone Institute, to develop the Internationally Educated Nurse Competency Assessment Program (IENCAP), which was implemented in 2014. The IENCAP is a competency-based Objective Structured Clinical Examination (OSCE), which assesses nursing practices through interactions with standardized patients. The test content reflects Canadian nursing competencies, which include concepts such as client-centredness, interdisciplinary collaboration, and problem-based communication.

Changes to Language Proficiency Standards

In addition to these regulatory and professional changes, there was a new impetus towards a pan-Canadian approach to IEN licensure and registration. Previously, each province had made decisions and set standards for nursing regulation, but in 2011, the Canadian Council of Registered Nurse Regulators (CCRNR) was convened to act as a national forum and voice regarding interprovincial/territorial, national, and global nursing regulatory matters (Shaffer et al., 2016). One of these matters was to confirm the currency of the test scores that were required as proof of language proficiency for the nursing profession. It is common practice to conduct standard-setting procedures to ensure that cut scores, or passing levels, are fair to examinees (Cizek & Bunch, 2007) and indicative of the language proficiency that is needed to function effectively and safely in the profession (Hull, 2015). Although the test development process had ensured that CELBAN results accurately reflected CLB levels, it was equally important to ensure that the passing score, or cut score, on each skill test was considered appropriate and fair by nursing regulators and stakeholders.

To this end, the CNO engaged a team of researchers in 2009 to carry out a standard-setting project. The project involved a panel of stakeholders who looked at the CELBAN test items and determined the passing level for each skill test (Office of the Fairness Commissioner, 2010). As a result of this standard-setting activity, the following CELBAN cut scores were recommended:

- Speaking CLB 8

- Listening CLB 10

- Writing CLB 7

- Reading CLB 8

This outcome reflected a change to the original listening cut score of CLB 9, which had been established during the CLB-referenced research. Implications of this change and the impact on CELBAN are discussed further along in this article.

Increasing Pressures on Test Usage

After the establishment of the CCRNR, additional changes to entry-to-practice standards impacted IEN language proficiency requirements. For example, in 2012 in Ontario, five language tests were accepted as evidence of English language proficiency. These were TOEFL (Test of English as a Foreign Language), TOEIC (Test of English for International Communication), IELTS (International English Language Testing System), MELAB (Michigan English Language Assessment Battery) and CELBAN (CNO, 2012). By 2013, only two approved tests, IELTS and CELBAN, were listed (CNO, 2013). Similar changes occurred across Canada.

As might be expected, with a change in regulatory policy limiting the number of acceptable English language tests to two (CELBAN and IELTS), the number of IENs registering for CELBAN increased. Between 2011 and 2013, the number of CELBAN examinees grew from 574 to 1,104 (Lewis & Kingdon, 2016). During this period, there was pressure on the national test administrators to address a waiting list of IENs who wanted to attempt the test.

As the size of the test-taking population continued to grow, the existing CELBAN content was being exposed to a greater number of examinees, a matter that could potentially impact the security of test tasks and items (Wendler & Walker, 2009). For the first 3 years of administration, there had been two versions of CELBAN in usage for each language skill, with a third version introduced in 2007 (Lewis & Kingdon, 2016). Over time, with administration pressures increasing and more than 1,000 IENs attempting CELBAN annually, the need for new forms of the test became critical, so that in 2014 when the national test administration was transferred to the CELBAN Centre at Touchstone Institute, strategies were sought to support research and development aimed at creating additional forms.

CLB Revision 2012

Another change that followed the development and implementation of CELBAN was a revision to the CLB document. This work was informed by broad consultations across the ESL field and a thorough review by curriculum and test developers. A team of Canadian practitioners and academics collaborated on the development of the revised 2012 edition, which remains aligned with the original CLB scale and its underlying theory of communicative language ability (Bachman, 1990; Celce-Murcia et al., 1995; Bachman & Palmer, 1996). Retaining the original interpretation of levels was important for the CLB document revision in order to ensure that existing assessment instruments would not have to be re-aligned to a new scale. However, with the revision process, the CLB document was altered somewhat in its format and wording for greater user friendliness, clarity, and specificity based on feedback from the field. From this perspective, it would be important to ensure that any newly developed CELBAN test content and scoring criteria would reflect the language used for descriptors and competency statements in the revised CLB 2012 document.

CELBAN Renewal Project

Factors referenced in the previous section of this article created the impetus for a project aimed at reviewing the CELBAN test model and developing additional content for each of the skill tests. To summarize, these factors were:

- shifts in the nursing regulatory environment

- changes in nursing practice towards a competency-based model

- launch of the CLB 2012 and its accompanying theoretical framework

- increase in CELBAN administration numbers

- transfer of test administration responsibilities from CELAS at Red River College to the CELBAN Centre at Touchstone Institute

The resulting project was termed a "renewal" because, in addition to the development of content to support additional forms of the test, the workplan also involved revisions to the test model and modifications to test specifications, procedures, and scoring. The following main objectives guided the project:

- Retain the intent and integrity of the original work carried out by the CELBAN test developers (Epp & Lewis, 2004a, 2004b, 2004c; Lewis & Kingdon, 2016).

- Consider feedback and observations gathered over a decade of test usage to determine what adjustments might be indicated for the test model, content and/or procedures (CELAS, 2014).

- Consider the implications, if any, of the CELBAN standard-setting procedure results.

- Bring together a synthesis of nursing and CLB expertise to create new content.

- Ensure compatibility of the new content and scoring procedures with the format and wording used in the CLB 2012 document (Centre for Canadian Language Benchmarks, 2013a).

Revisiting the CLB Difficulty Range

One of the first steps in the renewal process was to consider the range of CLB levels represented in each of the skill tests. As previously indicated, results of the national standard-setting procedure had established cut scores at speaking CLB 8, listening CLB 10, writing CLB 7 and reading CLB 8. For the skills of speaking, writing, and reading, these cut scores reflected the same ranges established in the original benchmarking activity, which meant that the renewed test forms for those skills could retain the same upper limits of difficulty as the original forms. However, for the skill of listening, the standard setting had established a cut score that was one benchmark higher than the range posited during the original benchmarking activity (Epp & Stawychny, 2002). Assuming that the original test forms contained a large proportion of items representing the original cut-score level of CLB 9, it would be important in the renewal project to ensure sufficient coverage at the newly established cut point of CLB 10.

A perennial question for test development is the range of difficulty that should be represented in a test's content. In a test of eligibility, one approach is to include only content that is calibrated at the passing-score level. In other words, if a reading test requires a passing score of CLB 8, all of the items in that test could be calibrated at a difficulty level of CLB 8. There might be no reason to include content across a broader range of levels unless there is some benefit to doing so. In the case of CELBAN, providing content across a range of ability has certain advantages, as it serves the following purposes:

1. It facilitates authentic representation of the range of tasks that nurses need to accomplish. A variety of task types can be included in the test reflecting a progression of competency and complexity.

2. It supports the assignment of criterion-referenced test scores that accurately reflect CLB levels. This, in turn, assists examinees in

understanding where they fit on the continuum of language ability. When they are not successful in meeting the CELBAN cut score, they are provided with their achieved CLB level for each skill. With this knowledge they can consult the CLB document and other related resources to understand what their levels mean and map out a language training plan to improve their skills and eventually meet the standard required by the regulators.

3. It allows for direct connection with language training providers, such as the many IEN bridging programs at colleges and universities where CELBAN results are accepted for entry and placement purposes. It also benefits community colleges across Canada where the language demands of healthcare programs have been aligned to the CLB (Hammond & Holmes, 2011), and it eliminates the need for IENs to take additional exams for entry into these programs.

4. It fits with the learner-centred "can do" (Centre for Canadian Language Benchmarks, 2013b) approach of the CLB, which devotes attention to recognizing communicative strengths.

It is in the spirit of supporting IENs as they progress towards their language goals that CELBAN has always offered more than a pass/fail result and has always included content across a range of complexity for each of the skill tests. If the test is to provide a valid result at each reported benchmark, the range of content must sufficiently accommodate each of those levels.

CELBAN test scores locate examinees on the CLB scale to inform them about their proficiency levels at the time of the test and to let them know how close they are to meeting the required passing scores. In order to provide this added value, it is necessary for the score range to reflect a reasonable span of benchmarks—narrow enough to produce reliable results yet broad enough to adequately inform examinees of their relative position on the scale of language ability. In the renewal project, it was determined that the results range for each CELBAN skill test would span four benchmarks. Speaking and writing results would be reported across the range of CLB 6 to 9, while results for listening and reading would be reported across the range of CLB 7 to 10.

For the productive skills, this decision would demand prompts and tasks accessible to the target range, along with a set of scoring criteria that could accurately distinguish the target benchmarks. For the receptive language skills, both concrete and abstract content would have to be included in the test, with items to tap some of the more basic elements of comprehension, such as getting the main idea and identifying factual details, along with more complex abilities, such as recognizing nuance and drawing inferences.

Selecting Renewal Task Types

When the renewal project began, the original CELBAN test content was reviewed by a test development expert and a panel of nursing professionals. In addition, feedback was gathered from experienced assessors who had administered the test over a number of years. The feedback was analyzed to inform adjustments to protocols, task types, and scoring procedures.

When making adjustments to task types, it was critical to keep in mind that an occupation-specific assessment is distinguished from a test of general language proficiency by "authenticity of task" (Douglas, 2000). Douglas defines authenticity as the "interaction between language knowledge and specific-purpose content knowledge. Authenticity of task means that the test tasks should share critical features of tasks in the target language use situation of interest to test takers" (2000, p. 2).

While authenticity is always important in an occupation-specific test, it is equally important to consider potential issues that can arise when authentic tasks are transferred to the assessment context. The fact that a task is realistic does not automatically make it suitable for testing purposes (Norton & Stewart, 1999; Stewart, 2008). To ensure a balance between the need for a task to be authentic to the nursing context and the need for it to be suitable for testing purposes, the following guiding criteria were used to inform the task selection process.

- Relevance to the healthcare context

- Compatibility with CLB descriptors

- Suitability for testing language rather than nursing knowledge

- Efficiency of administration

- Efficiency and reliability of scoring

These task selection criteria were helpful when it came to evaluating the suitability of content suggested by subject matter experts (SMEs). As practicing nurses and nurse educators, the SMEs were often focused on professional competencies, so that in some cases, their suggested tasks were not suitable as measures of communicative ability. For example, a task that requires an examinee to describe or explain a specific medical procedure would be unfair in a language test, as it would require the application of nursing knowledge. When examinees take CELBAN, they are told that the test is not measuring their nursing knowledge but only their language ability, and this premise has to be reflected in all of the tasks. For this reason, prompts such as "Explain how you would insert a catheter" or "Describe the symptoms of multiple sclerosis" could not be accepted for CELBAN tasks. With the list of guiding

criteria for referral, the content team could identify any suggested tasks that might be highly relevant to nursing but not appropriate for a CLB-based language assessment. SMEs could be referred to specific pages in the CLB document for information on the features of communicative language tasks and clarification on the reasons why some of their suggested nursing tasks were unsuitable for the purpose of language assessment.

The guiding criteria also informed decisions about whether to retain certain task types from the original CELBAN test forms. For example, the writing test included a video-mediated task that required examinees to listen to an interaction and fill out a chart with accurate information while listening. This task was authentic. It replicated many professional situations in which nurses listen to information from clients and make notes about symptoms, medical history, lifestyle choices and medications. Furthermore, the specifications were compatible with CLB 7–10 writing task types under the Reproducing Information competency, which include taking notes while listening to information (Centre for Canadian Language Benchmarks, 2012). However, as a writing assessment task, this activity relied quite heavily on listening comprehension, making it likely that examinees with higher listening proficiency levels would perform better on the writing task than other examinees at the same level of writing ability whose listening proficiency was not as advanced. Although many authentic tasks combine listening and writing, performance on those two skills is actually very different. Listening is more closely related to reading as both are receptive skills, and to speaking as both are oral/aural skills, but the relationship of listening to writing is not as strong (Hossein, 2012). Since there is a separate CELBAN instrument to evaluate listening ability, and because listening and writing abilities are not strongly related, it was determined that the listening load should be eliminated from the writing test. This meant that, despite its authenticity and CLB compatibility, the task in question had to be replaced during the renewal project.

For different reasons, an adjustment had to be made to one of the original listening task types that had been based on a series of relatively lengthy video-mediated scenarios. The length of the scenarios was compatible with CLB descriptors, and the task required examinees to concentrate intently on each situation just as they would do in an actual nursing context. However, the length of these scenarios had to be reduced because the longer passages required large numbers of test items to be associated with a single video clip, which limited the number of scenarios that could be included in a single test form and made it challenging to mix and match items when creating alternate forms of the test. When the clips were shortened, it became possible to increase the number of scenarios and the variety of contexts and speakers represented in the test, thereby expanding domain coverage without extending administration time.

Referencing Test Results to the CLB 2012

In making adjustments to task types and content, it was important to retain the CLB compatibility established by the CELBAN benchmarking team (Epp & Stawychny, 2002) and the test developers (Epp & Lewis, 2004b, Lewis & Kingdon, 2016). The original scoring procedures were referenced to CLB 2000 descriptors, and as the underlying scale had not been altered in the 2012 revisions to the CLB document, there was no requirement to adjust the interpretation of CELBAN test results. In other words, the original alignment of test results to CLB levels could be presumed accurate. However, it was necessary to refine the scoring criteria to reflect new task types and to ensure consistency with the wording of CLB 2012 descriptors.

Accordingly, the scoring grids for productive skills underwent a revision, with the CLB 2012 document providing direct support for this undertaking. The CLB Profiles of Ability were particularly helpful in this regard, as they provided summaries of performance at each benchmark. A Profile of Ability indicates, in general terms, what a person can do, and also hones in on specific features of communication, such as vocabulary, structure, and mechanics. This approach is highly compatible with the development of holistic and analytic scoring grids, as it provides a snapshot of the progression of language ability across levels.

The first round of revision resulted in draft versions of new scoring grids for the productive skills, and in order to test out the suitability and utility of these grids, a group of trained CELBAN examiners reviewed the scoring criteria and provided feedback to inform a round of revision. They then used the revised grids to evaluate samples of speaking and writing performance that had been gathered for this purpose across the range of CLB levels. Further feedback was provided on the clarity, specificity, and usability of the scoring criteria, and the grids were further refined based on the examiners' experiences working with the grids. The final grids were then used for scoring productive skill performance during pilot testing (Touchstone Institute, 2018a, 2019b).

For the receptive skills, newly developed items had to be referenced to the CLB, and one way to do this would have been to pilot new items and compare their calculated difficulties with the difficulties established for the original test items in the 2004 development project. This strictly data-based approach was rejected because CELBAN data analysis relies primarily on Classical Test Theory (CTT), a method in which the item statistics are sample dependent. Using CTT, it would not have been possible to accurately compare item difficulties derived at different times from two different samples of the population (Magno, 2009; Schumacker, 2010).

Another possible approach would have been to select a suitable range of anchor items from the original test and embed them, along with newly developed items, into pilot test forms, so that all of the items could be tested

on the same sample. The challenge in this regard was the possibility that the original test items might have been known to pilot participants. CELBAN pilot samples are drawn from the population of IENs, and the three original CELBAN test forms had been in usage for a lengthy period. It was therefore possible that some volunteer pilot participants might have had knowledge of the original test content, either directly as a result of having taken CELBAN themselves, or indirectly through information passed along by other examinees. If the original items had been compromised in any way, the relative difficulties established in the pilot data analysis would not be accurate.

For these reasons, an alternative approach was taken for relating the new receptive-skill items to the CLB scale. This involved applying a judgement-based approach (Tannenbaum & Wylie, 2004), with elements of the bookmark method (Mitzel et al., 2001) often used for standard-setting activities. The specific methodology was devised and refined through a series of previous CLB-based test development projects (Stewart & Nagy, 2004, 2015). It involved a facilitated interaction among a small panel of CLB experts, who relied on CLB descriptors and empirical difficulty indices to assign a CLB level to each item in the reading and listening tests. The objective was to achieve consensus on the CLB level of each item, as consensus methods generally render more accurate results than an average of independent judgements (Hambleton & Jirka, 2009). It was considered important to employ experts with a high degree of CLB knowledge and experience, as consensus among a small panel of true experts was preferable to averaging the opinions of a larger group with less expertise.

The panel comprised four individuals, with every expert meeting the following requirements:

- TESL certification and ESL teaching experience

- at least 20 years working in the ESL field

- CLB-based test item writing experience

- experience in developing CLB-based curricula

All but one panelist also met these additional criteria:

- CLB-based test development experience

- previous experience in health-related assessment

- member of CLB 2012 document revision team

- experience in test administration and management

The CLB experts worked together over a series of three meetings to reach consensus on the assignment of a CLB level to each receptive-skill test item, after which they determined how many correct responses would be required to achieve each CLB level on the overall test. In this endeavour, the group relied on data analysis results from the reading and listening pilot tests (Touchstone Institute, 2018b, 2019a), which provided a difficulty index for each item. These difficulty rankings were used, along with criteria from the CLB Profiles of Ability and Competency Indicators, to reach consensus on the CLB level of each test item. Throughout the work on scoring procedures, the consensus of CLB experts was limited to linking test scores with the Canadian Language Benchmarks. Consideration of cut scores for acceptance to nursing practice was beyond the scope of this process, as standard setting is the responsibility of nursing regulatory bodies.

Maintenance and Continuous Improvement

Updating Examinee Resources

Following the renewal of the test model and configuration of additional test forms, it was important to ensure that the CELBAN information available to examinees remained current and relevant, as stipulated in the standards that apply to development and maintenance of assessments (AERA, APA, NCME, 2014). To this end, the CELBAN Test Information Manual (CELBAN Centre, 2019) was updated with details about the test format, task features, administration procedures and evaluation criteria. Refinements were also made to the CELBAN diagnostic feedback process. Examiners regularly provide "actionable feedback" (Cannon & Witherspoon, 2005) for each person who attempts CELBAN so that examinees know what their strengths and weaknesses are. Following the renewal project, a new set of CLB-based standardized descriptors was framed to reference key elements of communication that are fundamental to performance in the general nursing environment (Hull, 2015). Based on the feedback that CELBAN provides, examinees who have not met the language requirement can map out plans to improve their communicative ability in the areas of weakness that have been identified.

In terms of test preparation, the CELBAN Readiness Self-assessment (CRSA), which was developed in 2005 by CELAS at Red River College (Lewis & Kingdon, 2016), has always been a popular resource for prospective test takers (CELBAN Centre, 2019), and following the test renewal, the CELBAN Centre and the CCLB worked on a series of updated test practice materials that reflect features of the renewed model.

Migrating CELBAN to Online Delivery

The CELBAN development and renewal process resulted in a paper-based assessment, with the speaking component administered individually in a face-to-face context and the other skill tests administered to large groups in proctored sessions. Under this model, CELBAN test administrations were conducted until early 2020, at which time the COVID-19 pandemic introduced restrictions that impacted test operations. Because the assessment was not designed to function amid pandemic restrictions or to allow for physical distance, administrations had to be cancelled and other delivery approaches immediately considered.

To this end, a research project was initiated to investigate the feasibility of migrating CELBAN to an online environment. This undertaking, which involved collaboration between Touchstone Institute and the Centre for Canadian Language Benchmarks, led to the launch in January, 2021 of a delivery system that allows candidates to access CELBAN through approved Canadian examination centres or from their homes. The two-part administration consists of the speaking interview, conducted via a virtual platform with a trained CELBAN examiner who interacts with the examinee in real time, and a computer-based assessment of reading, writing, and listening.

With the shift to online delivery, CELBAN constructs and format have not been altered. The assessment retains its original task-based approach and linkage to the CLB scale. Psychometric validation has been conducted to ensure inter-rater reliability of online speaking and writing test results, as well as equivalence with results derived from the former paper-based test model for the four language skills. Security of online delivery is ensured by means of an environment scan, strict examinee identification protocols and live monitoring of test sessions, with every assessment also video recorded.

Ongoing Content Development

Ongoing development of CELBAN content is essential to ensure that the items and prompts in each skill test do not become overly exposed to the examinee population and therefore widely known. This is an important component of regular test maintenance, but it is one that is often overlooked. With the renewal process completed and regular administration activity in place, attention has turned to the consideration of regular content development cycles to ensure that CELBAN prompts and items remain fresh and new for examinees. The planned approach to content development will be different for each language skill and will be based on the procedures that were applied during the renewal process for preparing, pilot testing, and configuring additional content for new test forms.

Conclusion

A language assessment is not a finite commodity to be developed in a single endeavour and then administered indefinitely. Instead, it is a dynamic procedure that requires ongoing consideration and maintenance of its content, protocols, and procedures. The maintenance of an occupation-specific language assessment has many facets. In addition to the predictable day-to-day activities, such as examinee registration, test administration, delivery of results, and dissemination of information, there are evolving circumstances and unforeseen events that impact the flow of operations.

CELBAN has weathered shifts in the nursing regulatory context, a revision of the underlying CLB framework, adjustments to the accepted cut scores, and most recently, the COVID-19 pandemic. The response to these influences has resulted in a renewal of the CELBAN test model and an adaptation of the test content and procedures for online administration, along with corresponding updates to supporting resources. As the future of occupation-specific language assessment unfolds, and as circumstances in the Canadian healthcare context continue to evolve over time, it is hoped that CELBAN will have the resilience to go on serving the needs of IENs and nursing regulators for years to come.

The Authors

Gail Stewart has been a language assessment consultant and advisor for more than 25 years. Her consulting activities include test development, framework design, practitioner training, validation studies and standard setting. Gail has served as lead on numerous assessment projects and has developed instruments for placement, diagnostic, outcomes and eligibility purposes.

Andrea Strachan is the director, learning & simulation and the lead for the Centre for Communicative Competence at Touchstone Institute, an organization that develops and administers competency assessments and education programs for internationally educated health professionals in Canada, including the Canadian English Language Benchmark Assessment for Nurses (CELBAN).

References

Alboim, N. (2002). *Fulfilling the promise: Integrating immigrant skills into the Canadian economy*. Ottawa: Caledon Institute of Social Policy. www.maytree.com/publications/fulfiling-the-promise-integrating-immigrant-skills-into-publications/the-canadian-economy

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). Standards for educational and psychological testing. Washington: American Educational Research Association.

Austin, Z., Galli, M., & Diamantouros, A. (2003). Development of a prior learning assessment for pharmacists seeking licensure in Canada. *Pharmacy Education 3*(2), 87–96.

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

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

Blythe, J., Baumann, A., Rheaume, A., & McIntosh, K. (2006). *Internationally educated nurses in Ontario: Maximizing the brain gain*. Nursing Health Services Research Unit: Human Resource, 3, 2, 1–43. www.fhs.mcmaster.ca/nhsru/documents/SeriesReport3InternationallyEducatedNursesinOntarioMaximizingtheBrainGain.pdf

Canadian English Language Assessment Services (CELAS) Centre. (2014). *Report on the final seven months of operations at CELAS in the ninth year of official CELBAN administration in Canada, January 1, 2014–July 31, 2014*. Winnipeg: Red River College. https://cpb-ca-c1.wpmucdn.com/www.rrc.ca/dist/6/47/files/2016/06/Ninth-Year-Official-CELBAN-November-2014.pdf

Canadian Nurses Association. (2010). *Interprofessional collaboration: Position statement*. Ottawa: CNA. www.cna-aiic.ca/~/media/cna/page-content/pdf-en/Interproffessional-Collaboration_position-statement.pdf

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Journal*, *19*(2), 120–134.

CELBAN Centre. (2019). *Test information manual*. Toronto: Touchstone Institute.

Celce-Murcia, M. Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, *6*(2), 5–35.

Centre for Canadian Language Benchmarks. (2000). *Canadian Language Benchmarks 2000: English as a second language for adults*. Ottawa: Citizenship and Immigration Canada.

Centre for Canadian Language Benchmarks. (2012). *Canadian Language Benchmarks: English as a second language for adults*. Ottawa: Citizenship and Immigration Canada.

Centre for Canadian Language Benchmarks. (2013a). *Theoretical framework for the Canadian Language Benchmarks*. Ottawa: Centre for Canadian Language Benchmarks.

Centre for Canadian Language Benchmarks. (2013b). *Canadian Language Benchmarks can do statements*. Ottawa: Centre for Canadian Language Benchmarks.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

College of Nurses of Ontario (CNO). (2012). *Fair registration practices report to the Office of the Fairness Commissioner (OFC).* Ontario: CNO.

College of Nurses of Ontario (CNO). (2013). *Fair registration practices report to the Office of the Fairness Commissioner (OFC).* Ontario: CNO.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Epp. L., & Lewis, C. (2004a). *Developing an occupation-specific language assessment tool using the Canadian Language Benchmarks*. Ottawa: Centre for Canadian Language Benchmarks.

Epp. L., & Lewis, C. (2004b). *The Development of CELBAN (Canadian English Language Benchmark Assessment for Nurses): A Nursing-specific language assessment tool*. Ottawa: Centre for Canadian Language Benchmarks.

Epp. L., & Lewis, C. (2004c). *Position paper: An analysis comparing CELBAN with other English language tests including TOEFL*. Winnipeg: Red River College Language Training Centre. www. https://cpb-ca-c1.wpmucdn.com/www.rrc.ca/dist/6/47/files/2016/06/Position-Paper-An-Analysis-Comparing-CELBAN-with-other-English-Language-Tests-including-TOEFL.pdf

Epp, L., & Stawychny, M. (2002). *Benchmarking the English language demands of the nursing profession across Canada*. Ottawa: Centre for Canadian Language Benchmarks.

Hambleton, R. K., & Jirka, S. J. (2009). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladayna (Eds.), *Handbook of test development*. New Jersey: Lawrence Erlbaum Associates.

Hammond, K., & Holmes, T. (2011). *Comparative analysis of CLB benchmarking methods*. Edmonton: Government of Alberta.

Hossein, B. (2012). The relationship between listening and other language skills in international English language testing system. *Theory and Practice in Language Studies, 2*(4), 657–663.

Hull, M. (2015). Medical language proficiency: A discussion of interprofessional language competencies and potential for patient risk. *International Journal of Nursing Studies*, *54*, 158–172.

Jeans, M. E., Hadley, F., Green, J., & Da Pratt, C. (2005). *Navigating to become a nurse in Canada: Assessment of international nurse applicants*. Ottawa: Canadian Nurses Association. www.cna-aiic.ca/~/media/cna/page-content/pdf-fr/ien_technical_report_e.pdf?la=en

Johnson, K., & Baumal, B. (2011). *Assessing the workforce integration of internationally educated professionals*. Hamilton: Canadian Society for Medical Laboratory Sciences (CSMLS). www.csmls.org/csmls/media/documents/publications/CSMLS-KS-0001_WFI_Report_ENG.pdf

Lewis, C., & Kingdon, B. (2016). CELBAN: A 10-year retrospective. *TESL Canada Journal*, *33*(2), 69–82.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

McNamara, T. F. (1990). *The Occupational English Test*. Melbourne: University of Melbourne.

Mendelsohn, D., & Stewart, G. (1999). Developing the Midwives' Language Proficiency Test: A field-specific performance test. *ESP Malsysia: A Journal on English for Specific Purposes, 6*, 11–27.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. New Jersey: Lawrence Erlbaum Associates.

Norton, B., & Stewart, G. (1999). Accountability in language assessment of adult immigrants in Canada. *Canadian Modern Language Review, 56*(2), 223–244.

Office of the Fairness Commissioner. (2010). *Annual report 2009–2010*. Toronto. www.fairnesscommissioner.ca/files_docs/content/pdf/en/OFC_Annual_Report_2009-2010_English.pdf

Russell, B., Strachan, A., & Shek, P. (2009). Developing CLB-referenced assessments for internationally trained health professionals. *TESL Contact Special Research Symposium Edition*, *32*(2), 86–103.

Schumacker, R. E. (2010). *Classical test analysis*. University of Alabama: Applied Measurement Associates.

Shaffer, F., Robinson, M., Dutka, J., & Tuttas, C. (2016). A new model for assessing entry-level education of internationally educated nurses: A retrospective perspective. *Journal of Nursing Regulation*, *6*(4), 51–57.

Stewart, G. (2008). Issues in task-based assessment of communicative competence. *TESL Contact Special Research Symposium Issue*, *31*(2), 27–42.

Stewart, G., & Nagy, P. (2004). *Development of the CLBA International* [unpublished report]. Mississauga: The Centre for Education and Training.

Stewart, G., & Nagy, P. (2015). *Milestones Canadian Language Benchmarks listening test development report* [unpublished report]. Ottawa: Centre for Canadian Language Benchmarks.

Strachan, A. (2007). *English language proficiency and internationally educated medical laboratory technologists (IEMLT): An investigation of language benchmarks and assessment tools for IEMLT success*. Hamilton: CSMLS. Retrieved from https://csmls.org/csmls/media/documents/publications/reports/csmls_final_report_-_investigation_of_language_tools.pdf

Tannenbaum, R. J., & Wylie, E. C. (2004). *Mapping test scores onto the Canadian Language Benchmarks: Setting standards of English language proficiency on the Test of English for International Communication (TOEIC), the Test of Spoken English (TSE) and the Test of Written English (TWE)*. Princeton: Educational Testing Services.

Touchstone Institute. (2018a). Speaking test renewal. *CELBAN Facts and Figures, Issue 3*.

Touchstone Institute. (2018b). Reading test renewal. *CELBAN Facts and Figures, Issue 4*.

Touchstone Institute. (2019a). Listening test renewal. *CELBAN Facts and Figures, Issue 5*.

Touchstone Institute. (2019b). Writing test renewal. *CELBAN Facts and Figures, Issue 6*.

Wendler, C. L. W., & Walker, M. E. (2009). Practical Issues in Designing and Maintaining Multiple Test Forms for Large-Scale Programs. In S. M. Downing & T. M. Haladayna (Eds.), *Handbook of test development*. New Jersey: Lawrence Erlbaum Associates.