Investigating the Alignment Between the CELPIP-General Reading Test and the Canadian Language Benchmarks: A Content Validation Study

Michelle Y. Chen
Paragon Testing Enterprises

Jennifer J. Flasko
Paragon Testing Enterprises

Abstract

Seeking evidence to support content validity is essential to test validation. This is especially the case in contexts where test scores are interpreted in relation to external proficiency standards and where new test content is constantly being produced to meet test administration and security demands. In this paper, we describe a modified scale-anchoring approach to assessing the alignment between the Canadian English Language Proficiency Index Program (CELPIP) test and the Canadian Language Benchmarks (CLB), the proficiency framework to which the test scores are linked. We discuss how proficiency frameworks such as the CLB can be used to support the content validation of large-scale standardized tests through an evaluation of the alignment between the test content and the performance standards. By sharing both the positive implications and challenges of working with the CLB in high-stakes language test validation, we hope to help raise the profile of this national language framework among scholars and practitioners.

Résumé

La recherche sur la validité du contenu est essentielle à la validation des tests. Cette recherche est encore plus importante dans les contextes où les résultats des tests sont interprétés par rapport à des normes de compétence externes et dont le contenu du test est constamment révisé pour répondre aux exigences de l'administration et de la sécurité du test. Dans cet article, nous décrivons une approche d'ancrage d'échelle modifiée pour évaluer l'alignement entre le test du *Canadian English Language Proficiency Index Program* (CELPIP) et les *Canadian Lagunage Benchmarks* (CLB), le cadre de compétence linguistique auquel les résultats du test sont liés. Nous discutons comment les cadres de compétence tels que le CLB peuvent être utilisés pour soutenir la validation du contenu des tests standardisés à grandes échelles grâce à l'évaluation de l'alignement entre le contenu du test et les normes de performance. Nous évaluons les forces et les défis de l'utilisation du CLB comme outil de validation des tests linguistiques à enjeux élevés et ce faisant nous espérons contribuer à relever le profil de ce cadre linguistique national auprès des universitaires et des praticiens.

Investigating the Alignment Between the CELIPIP-General Reading Test and the Canadian Language Benchmarks: A Content Validation Study

Test scores alone are insufficient in supporting test users to make meaningful decisions about test takers. They can also leave test takers insufficiently informed of their own proficiency levels and abilities. Aligning a test to an external proficiency framework links the test scores to a set of language criteria, lending greater meaning to the scores (Kane, 2012) and allowing scores from different tests to be indirectly compared. As a result, the past decade has seen an emerging interest in test alignment (Brunfaut & Harding, 2014; Papageorgiou et al., 2015; Tannenbaum & Wylie, 2004, 2008). Importantly, the relationship between the test and the proficiency framework is not an observable fact, but an assertion for which we, as test developers and researchers, must continuously provide evidence.

The present study uses a variation of the scale anchoring method to evaluate the content validity of the high-stakes, large-scale test, the Canadian English Language Proficiency Index Program (CELPIP)-General, the scores of which are linked to the Canadian Language Benchmarks (CLB). This approach uses a combination of quantitative and qualitative methods in which the former selects *anchor* items that are most discriminating between adjacent score bands and the latter draws in expert judgements to map the selected items to the levels of the external proficiency framework to which the test is linked. This method is particularly helpful in contexts where new test items are continuously being adding to the item bank or the number of items is too large to be individually reviewed by an expert panel in a validation study. We start by introducing the CELPIP-General test as well as the CLB and their use in Canada. Next, we discuss test linking, content validity, and the use of the scale anchoring method in large-scale test settings. We then present a modified scale-anchoring approach for validating test content in relation to external performance standards using data from the CELPIP-General reading test. To better prepare researchers and practitioners to use the CLB in a similar context, we end the paper by discussing the challenges associated with using the CLB in projects that rely on expert judgement.

The CELPIP-General Test

The CELPIP-General test is designed to measure the communicative competence or functional English proficiency required for successful participation in Canadian communities where English is used as a medium for communication in various social, educational, or workplace contexts. Following Bachman and Palmer's model, communicative competence refers to an individual's ability to integrate language knowledge and skills in order to understand and produce language to achieve communicative goals (Bachman & Palmer, 1996, 2010). This implies the comprehension and production of not only the forms and structures of the language but also its objectives and rhetorical conventions. Communicative competence is also described as functional language proficiency or "the expression, interpretation, and negotiation of meaning involving interaction between two or more persons belonging to the same (or different) speech community" (Savignon, 1997, p.272). The communicative approach to language teaching and assessment views language as a vehicle for meaning-making and focuses on the development and measurement of learners' functional proficiency in authentic contexts (Savignon, 1991, 1997). Consistent with the underlying theory and construct of the test, CELPIP-General test

tasks assess the skills needed for the interpretation and production of language as it is used in a variety of general or day-to-day interactions in common social and workplace contexts.

The interpretations of CELPIP scores are criterion-referenced to the 12 benchmarks of the CLB, and these scores are used for Canadian immigration and citizenship purposes. The CELPIP-General test scores have been linked to the CLB through standard-setting studies (Chen, 2016; Paragon Testing Enterprises, 2013a, 2013b). Multiple methods were used in these standard-setting studies to establish the correspondence between CELPIP scores and CLB levels. For the listening and reading tests, both of which consist of multiple-choice questions, Paragon Testing Enterprises (hereafter, Paragon) used a modified Angoff method (Angoff, 1971) to link the CELPIP scores to the CLB levels and consolidated the results using the Direct Consensus method (Sireci et al., 2004). For the speaking and writing tests, which are based on raters' evaluations of test taker performances, Paragon used a modified Judgmental Policy Capturing procedure (Hambleton & Pitoniak, 2006) in the initial standard-setting studies and triangulated the results using the Body of Work method (Kingston et al., 2001). These standard-setting procedures allowed Paragon to establish a correspondence between CELPIP test scores and CLB levels, providing initial evidence of the alignment between the two.

The CLB and Their Use in Canada

Language proficiency frameworks are an established set of criteria that describe the language ability of learners at various levels. These language standards are developed by experts in the field to help bring scholars and practitioners together to share a common understanding of language abilities across the proficiency spectrum. Several language proficiency frameworks are currently used in Canada, including the Common European Framework of Reference for Languages (CEFR), the *Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes* (EQ), and the Canadian Language Benchmarks (CLB)/*Niveaux de compétence linguistique canadiens* (NCLC; the French-language counterpart of the CLB, Centre for Canadian Language Benchmarks [CCLB], 2012; *Centre des niveaux de compétence linguistique canadiens*, 2012). Among them, the CEFR has been most widely adopted and used for multiple languages and contexts worldwide, providing an international standard for the description of second language proficiency. In Canada, the EQ provides a common framework of reference for describing the French language competence of immigrants to Quebec, and the CLB/NCLC provide the national language standards for adult users of English/French as a second language (ESL/FSL) in work, study, and social contexts.

Like the CEFR, the CLB have been used in the development of a wide range of language curriculum and assessment tools. In contrast to the CEFR, which was designed to be a generic language reference document, the CLB, by comparison, are designed specifically for the English language and contextualized within work, study, and social contexts in Canadian society. Consequently, while the CEFR has been criticized for failing to account for the influence of context on language proficiency (termed "context validity" by Weir, 2005), the CLB embed the demand of the context within the proficiency descriptors. The CLB describe language progression not only in terms of increasingly precise, complex, lengthy, and flexible language use but also in terms of the increasing demand associated with the context of the communicative task.

Validity evidence for the CLB (and NCLC) has been reported in multiple sources (Bournot-Trites et al., 2015; Bournot-Trites & Barbour, 2012 and Elson, 2012a, 2012b as cited in Bournot-Trites, 2017; North & Piccardo, 2018). The initial construct and content validity of the CLB (and NCLC) was established through a three-stage validation process undertaken by a Canadian team of experts in 2010. The first stage involved the development of a common theoretical framework for the CLB and NCLC, which was subsequently reviewed and validated against the relevant literature, as well as against the CLB and NCLC descriptors, by teams of independent experts. The second stage involved a comparison of the common theoretical framework against other common proficiency frameworks, namely, the CEFR, the EQ, and the American Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (Bournot-Trites et al., 2015). In the third stage, content experts developed sets of exemplars for each of the 12 CLB/NCLC benchmarks including reading and listening texts and tasks as well as speaking and writing prompts and corresponding samples of learner performances. These exemplars were then trialled with over 100 practitioners across Canada to confirm the appropriateness of the exemplars representing each benchmark with respect to language instructors' firsthand experience with learners at these levels. According to Bournot-Trites et al., the revised and validated CLB/NCLC conform to the standards for reliability and validity imposed by the *Standards for Educational and Psychological Testing* (hereafter *Standards*; American Educational Research Association [AERA] et al., 1999). They also report that the results of the validation process support the use of the CLB/NCLC as national language standards in Canada as well as for other purposes including use in high-stakes contexts. For example, the CLB/NCLC may serve as a reference for the desired indicators of ability at different assessment levels of a high-stakes language test.

As a set of national language standards, one of the primary users of the CLB is the Language Instruction for Newcomers to Canada (LINC) program for adult newcomers (permanent resident or Convention Refugee), funded by the Government of Canada (Immigration, Refugees and Citizenship Canada, Government of Canada, 2018). LINC curriculum guidelines are based on the CLB and developed in consultation with CLB experts. The guidelines instruct program coordinators and teachers to develop and plan course content consistent with the criteria of the CLB (Hajer et al., 2002).

Additionally, the CLB have been used to develop various assessment tools. These assessments take many forms, including learner self-assessment, portfolio assessment, and instructor-based assessment. For example, the Canadian Language Benchmarks Placement Test (CLBPT), designed by the CCLB, is a low-stakes placement tool used for entry into language programs such as LINC (Bruni & Irwin, 2007).

The CLB are also heavily involved in government-mandated assessment provisions (e.g., for immigration and professional certification purposes). For example, to support decisions regarding immigration and citizenship applications, the Government of Canada sets English (and French) language proficiency standards in reference to the CLB (and its French counterpart, the NCLC) levels (Government of Canada, 2020). Scores of standardized tests, including the CELPIP and the International English Language Testing System (IELTS), are accepted as proof of English language proficiency. Although neither of these tests is an assessment of the CLB per se, through the process of score linking, their scores have been aligned with the CLB and can be interpreted in relation to the CLB levels (Chen, 2016; Paragon Testing Enterprises, 2013a, 2013b). In these government-mandated testing contexts, test scores are used to

support high-stakes decisions. A misclassification of a test taker's language proficiency, as expressed in CLB levels, could result in the delay or rejection of the individual's application. Thus, the validity of the test scores and, accordingly, the evidence that supports the alignment between the test scores and the CLB is of the utmost importance.

Linking Test Scores to External Proficiency Standards

Test scores are summative indicators of an individual's proficiency levels; however, they are abstract and may not always convey a clear meaning. Even with the labels and brief descriptions that are typically provided in a score report, it may still be difficult for stakeholders (e.g., test takers and score users) to interpret the scores clearly and consistently. On the other hand, language proficiency frameworks and standards often detail the criteria for achieving each performance level as well as the strengths and limitations of learners at each level. They also provide some indication to learners as to the skills and abilities needed in order to progress to higher levels. Linking test scores to such frameworks and standards facilitates the interpretability of the scores and enables indirect comparisons across tests of similar constructs. Although it is possible to compare scores from different tests indirectly when they are linked to a common scale, such indirect comparisons must be interpreted with caution. Test linking, like many other measurement procedures, is rarely able to precisely translate every score from one scale to the other (i.e., there exists some degree of inaccuracy or measurement error) and indirect comparisons may amplify such discrepancies. This is particularly concerning when the tests, linked by indirect comparisons, differ in their constructs, formats, and/or reporting scales. Despite the caution against indirect comparisons, the linking of test scores to external standards is considered one method of test validation.

It is widely accepted that the interpretation and use of test scores are an essential consideration in the validity argument (Bachman & Palmer, 2010; Kane, 1992, 2002b, 2006 as cited in Kane, 2013). For example, according to Bachman and Palmer's *Assessment Use Argument (AUA)* framework, interpretations about the ability to be assessed must be sufficient for the decision to be made. The warrant for this claim is that the interpretation of the scores provides sufficient information for score users to make the required decisions concerning test takers. Linking test scores to external language standards is, therefore, "relevant to the sufficiency of interpretations" (Papageorgiou & Tannenbaum, 2016, p. 117) and facilitates proper use of test scores as the link enhances score interpretation (Kane, 2012).

While the act of linking test scores to external standards can be considered a form of test validation in and of itself, the results of the linking study must also be validated. According to the manuals for relating language tests to the CEFR (Figueras et al., 2009; North & Jones, 2009), validation of the linking study results should be a regular part of the linking process. Broadly speaking, one could validate the results of a linking study in one of two ways, replication or independent validation. The replication approach involves repeating the linking study but varying some of its features (e.g., recruiting another group of experts, selecting different sets of items and/or responses, or adopting a different linking method). The second approach involves conducting an independent validation study (e.g., evaluating the consequences of applying the linking results to a different population; Kane, 1994).

The alignment of test scores and external standards is not a relationship that exists objectively and statically, rather, it is a value-driven claim that must be continually supported by evidence. For tests that have new content and items

continuously being developed and administered to test takers, a one-time validity check of the test alignment at the end of a linking study is *not* sufficient. Ideally, all items created according to the same test specifications are expected to be interchangeable and remain so over time. Thus, the results of a linking study based on a specific subset of items should be generalizable to all items, and the relationship between the test scores and the external proficiency framework should be stable over time. In reality, this may not always be the case. Although many test organizations have rigorous procedures to ensure the high quality of test content, including the evaluation of content compliance with specifications, it is still possible that, over time, the features of items (i.e. their target knowledge, skills, and performance levels) may slightly drift away from those used initially to establish the linkage between the test scores and the proficiency standards. Although linking test scores to external standards primarily concerns the comparability of the performance levels, a shift in the features of the test content could still threaten the appropriateness of the previously established alignment (Dorans, 2018; Liu & Walker, 2007). Therefore, to support the validity of the linking results, the test content and items must continue to reflect the relevant domains and criteria described in the chosen proficiency framework.

Content Validity Evidence to Support Linking Results

Many researchers (Brown, 1996; Kane, 2013; Lissitz & Samuelsen, 2007) and the *Standards* (AERA et al., 2014) recognize content validity as a major source of support for test score validity. Brown defines content validity as the extent to which the test content is representative of what the test intends to measure, and he describes it as one of the "three main strategies" for validating test scores (p.232). Within Kane's argument-based framework for test validation, the extrapolation inference claims that the test is representative of the construct such that test scores can be taken to represent language ability in the target domain. This claim assumes that test performance reflects the criteria for language proficiency (Kane, 2013). Content-based validity evidence, therefore, supports the extrapolation inference of the validity argument by confirming the relationship between the test content and the language requirements of the target domain. According to the *Standards*, content-based validity evidence concerns the adequacy with which the test covers the content domain and the appropriateness of the content difficulty with respect to the target domain (AERA et al., 2014).

Content validity evidence may be particularly relevant for tests that are linked to external proficiency frameworks as the interpretation and use of test scores rely on the alignment to the performance standards. Considering that the majority of test alignment occurs in retrospect when a previously developed test is linked retroactively to an external framework, some discrepancy is likely to exist between their constructs or target domains. Test linking, especially the methods based on statistical analyses of scores (e.g., linking through equal percentile equating), focuses on aligning scores between different reporting scales and does *not* fully evaluate or account for the qualitative differences underlying these scales (e.g., small differences in their purposes, target populations, and domain coverage). A mismatch between the test at hand and the content domain as described by the external proficiency framework could limit the interpretability and usefulness of the linking results. In other words, the validity of the link relies on the correspondence between the content of the test and the proficiency descriptions/criteria detailed by the framework to which the test is linked. As such, to establish a stronger basis to support the correspondence, test developers and researchers

should collect additional evidence to investigate the adequacy and appropriateness of test content coverage in relation to the chosen proficiency framework.

To demonstrate content validity, typically, researchers and test developers enlist well-trained colleagues (e.g., subject matter experts) to make judgments about the degree to which the test items match the test objectives or target construct. Compared with validation studies that focus on evaluating final test scores using correlational based approaches, well-designed content validation studies often draw on multiple sources of data (e.g., expert judgement, test taker performance, and item statistics) and allow for a more detailed analysis of test items. By providing more fine-grained information, this type of analysis can help test developers better understand, and, if necessary, improve the extent to which the test covers its intended scope.

However, it is challenging to directly apply this approach to the content validation of large-scale tests where qualitatively reviewing all test content is impossible or inefficient. Large-scale tests often have a great number of test items and they constantly add new items to the pool. Without an explicit and well-laid-out strategy, investigating the content validity of such tests is daunting and may result in weak evidence to support or refute the validity of the test scores.

To this end, this study proposes an approach to the content validation of a large-scale language proficiency test based on the scale anchoring method. Scale anchoring is used internationally in educational testing contexts including language and subject matter assessment (Gomez et al., 2007; Jaeger, 2003; Liao, 2010; Philips et al., 1993). Conventionally, it is a process that attributes meaning to test scores by identifying test items representative of particular score points along a score scale (Beaton & Allen, 1992; Kelly, 1999). The typical scale anchoring methodology involves two main steps. Performance data is first analyzed to identify items associated with particular score points, or anchor points. These are items that are likely to be answered correctly by test takers scoring at each anchor point, but not by test takers at the anchor point below; in other words, items that discriminate between performance levels. Next, a panel of experts examines the selected items to identify the language knowledge or skills demonstrated by these items. These are the language abilities that are said to *anchor* at each performance level. These abilities are then used to develop statements describing the language competence that would be expected of test takers at each level. Instead of applying the scale anchoring approach to deriving descriptors for each score level, we slightly modified the original methodology to focus on assessing the alignment between the test content and the external proficiency framework to which the test is linked (see the *Method* section for more details on our modified approach).

Method

The present study assesses the alignment between the content of the CELPIP-General reading test and the reading proficiency descriptors of the CLB. The CELPIP-General reading test was designed to assess test takers' ability to comprehend a variety of written English texts. In order to support continuous test administration, new test content must be constantly developed, and a large number of test items are used on a rotating basis. It would therefore be impossible to qualitatively assess each of these items in one validation study. Instead, we identify the *anchor* items for critical score levels using a modified scale-anchoring approach (see Beaton & Allen, 1992 and Kelly, 1999 for examples of standard scale-anchoring methods) and map these *anchor* items to the CLB levels through expert judgement. To seek evidence upon which to evaluate the content validity of the reading test, we focused on the comparison of two elements: (1)

the items' anchor levels as determined by test taker response data and the scoring model (i.e., a two-parameter item response [2PL IRT] model for the CELPIP-General reading test), and (2) the assessment levels of the items as judged by experts using CLB descriptors (CCLB, 2012).

Data

A pool of 341 reading items was analyzed. Each item was answered by an average of 3,172 test takers (minimum 198, maximum 6,611). The analysis focused on seven performance levels from CELPIP 4 to 10, which correspond to CLB 4 to 10. These seven levels were selected because they cover the range of proficiency levels that are often used to support high-stakes decisions, such as those related to Canadian immigration and citizenship applications.

A total of 35 items (5 items × 7 score levels) were selected to be reviewed by a panel of four CLB experts. The number of panellists was largely constrained by operational limitations, including the budget, time, and the availability of the experts. As the first reported study using the modified scale-anchoring method for content validation, when recruiting the panellists, we prioritized their experience and expertise with the CLB and the target population. All the panellists had extensive experience (minimally five years) working with the CLB in the context of teaching, curriculum development, and assessment design. Additionally, they had intimate knowledge of the target test taker population (i.e., new Canadian immigrants) through their work as English language teachers and LINC program coordinators. All the panellists reviewed the 35 items and judged the target skills, knowledge, and contexts of each item, as well as the item's correspondence to the CLB descriptors.

Materials

The CELPIP-General Reading Items

The reading component of the CELPIP-General test is presented in testlet format. Each testlet consists of a passage and a corresponding set of items. The passages represent a variety of text types including correspondence, brochures, articles, and opinion pieces. The items are written to evaluate test takers' reading comprehension in terms of their ability to understand the main ideas, identify details, and make inferences about the content. A total of 35 items were selected for the panel to review.

The CLB Reading Component

The CLB document is organized by language component (listening, speaking, reading, and writing). Each component is divided into 12 benchmarks, which are grouped into three stages of Basic (CLB 1-4), Intermediate (CLB 5-8), and Advanced (CLB 9-12) language ability. For each component, the CLB document is composed of three main sections: Profiles of Ability, Knowledge and Strategies, and the Canadian Language Benchmark pages.

Figure 1
An illustration of the content covered by the Profiles of Ability in the CLB

CLB 4	CLB 5	CLB 6
Overview of reader abilities		
• Understands short, simple texts related to everyday topics that are personally relevant	• Understands simple and some moderately complex texts related to predictable situations	• Understands an adequate range of moderately complex texts related to predictable situations
Features of the text		
• Factual • Limited to common and concrete vocabulary • Short and clearly organized	• Mostly factual and descriptive • Mostly concrete and some abstract vocabulary	• Mostly factual and descriptive • Mostly concrete and some abstract vocabulary • Relatively short
Strengths and limitations		
• Understands overall meaning in simply connected discourse • Occasionally guesses unknown words • Comprehension is based on knowledge of basic grammar	• Identifies purpose, main ideas and important details • Occasionally guesses unknown words • Comprehension is based on some developing understanding of complex structures	• Identifies specific factual details and implied meaning • Sometimes guesses unknown words • Comprehension is based on a developing understanding of complex structures

Note. This is not a complete or verbatim representation of the descriptors provided in the CLB Profiles of Ability nor is it the version used by the panellists; only some key features within each proficiency level are paraphrased here to demonstrate the range of features covered in the CLB document. This table is presented to acquaint readers who are not familiar with the CLB with the kind of information provided in the CLB document.

While the entire CLB document was available for reference during the study, the panellists started by mapping items according to the descriptions in the Profiles (see Figure 1 for an illustrative example and see CCLB, 2012, p.86 for an example of Profiles in the CLB document). During the independent review stage, panellists were able to consult the Benchmark pages to make a judgement if they could not directly map an item to the descriptors provided in the summary tables of the Profiles. During the discussion stage, panellists were also able to refer to either the Profiles or the Benchmark pages to justify their evaluations of items.

Procedures

To evaluate the alignment of items to the performance standards of the CLB, we adopted a two-stage procedure similar to a typical scale anchoring study. Stage 1 focused on identifying anchor items. Anchor items are defined as items that demonstrate strong discrimination power at a given score level. For example, an anchor item for score level 7 is an item that a typical level 7 test taker would answer correctly while a typical level 6 test taker would get wrong. The CELPIP-General reading test uses a 2PL IRT model to predict test takers' proficiency and then transfer the predicted continuous score to the reporting scale by applying the cut scores established in previous standard-setting studies (Paragon Testing Enterprises, 2013a, 2013b). As such, in this study, we operationalize a "typical" test taker at a given score level as one whose proficiency level is at the mid-point of the adjacent cut scores. In the above example, a

typical level 7 test taker is represented as someone whose theta score (i.e., proficiency on the IRT theta scale) equals the median of the cut scores for level 7 and level 8.

For each item, we first computed the probability of a correct response by test takers at a given proficiency level (i.e., conditional probability) along the proficiency continuum—at each of the mid-points of the adjacent cut scores. The conditional probabilities were calculated based on the 2PL IRT model for dichotomous items,

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

where $P(X_i = 1|\theta)$ represents the conditional probability of a correct response to item *i* at proficiency level θ or theta, and a_i and b_i are the item discrimination and difficulty parameters of item *i*.

Then, we grouped items based on their anchor levels. An item (*i*) is deemed to anchor level *K*, if $P(X_i = 1|\theta_k) > 0.50$ and $P(X_i = 1|\theta_{k-1}) < 0.50$, where *K* is a score level from the reporting scale and θ_k represents a "typical" test taker at band level *K* (i.e., the test taker's theta score is at the mid-point of the adjacent cut scores). This implies that typical test takers at level *K* have a higher chance of answering this item correctly than getting it wrong (i.e., $P(X_i = 1|\theta_k) > 0.50$), while typical test takers at one level lower (*K-1*) are more likely to answer incorrectly than correctly (i.e., $P(X_i = 1|\theta_{k-1}) < 0.50$).

After grouping items by their anchor levels, we selected five items that showed the highest discrimination power for each of the CELPIP levels. The difference between the conditional probabilities of adjacent levels (i.e., the level of focus and one level below), $\Delta P\big(X_i = 1|\theta_{k,k-1}\big) = P(X_i = 1|\theta_k) - P\big(X_i = 1|\theta_{k-1}\big)$, indicates an item's discrimination power at that particular proficiency range. For levels 4 through 10, a total of 35 items were selected for expert review (see Table 1).

Stage 2 involved a qualitative analysis of the selected anchor items by a panel of CLB experts. This panel analyzed each item and members offered their opinions as to which CLB competency statements and linguistic functions were assessed by each item and to which CLB level they corresponded. The items were reviewed in random order. Neither the panellists nor the facilitator was aware of the CELPIP levels that each item was selected to represent. The review was done in two steps. First, each panel member made their judgements individually. Then, they discussed their responses at an in-person meeting during which the panellists could edit their responses. The panellists were requested to justify their evaluations but were not required to reach a consensus, and their final judgements were submitted individually after the meeting. These individual judgements were aggregated and then compared with the model-suggested anchor levels (i.e., the CELPIP levels that the items were selected to represent).

Results

Table 1 lists the five anchor items selected for each proficiency level along with the probability of a correct response to each item at two points along the proficiency scale. The difference between the probabilities of a correct response to each anchor item at the anchoring level and at the level beneath it represents the discrimination power of the selected item. Compared to the anchor items at levels 4 to 6, the anchor items for the higher levels had relatively low discrimination power (an average of 0.24 vs. 0.15). Lower discrimination at some proficiency levels indicates that less information could be obtained there. Currently, no clear guidelines exist for interpreting and evaluating the

discrimination power of anchor items. Also, the total number of levels selected along the proficiency continuum may affect the observed discrimination power. The more levels there are, the closer the adjacent levels are to each other on the theta scale and the more difficult it is to achieve strong discrimination for all levels. Despite some variability in the discrimination power for CELPIP levels 4 through 10, anchor items were identified for every level. Item 20 (score level 9) had the lowest discrimination power among the selected items. The probability of answering this item correctly by typical test takers at level 9 was 0.58, which was 0.11 higher than the probability of a correct answer by typical level 8 test takers.

Table 1

Selected Anchor Items and Their Model-Estimated Difficulties Across Adjacent Levels on the Reporting Scale

Score Level	Item ID (Testlet Type)	N	Mid 3	Mid 4	Mid 5	Mid 6	Mid 7	Mid 8	Mid 9	Mid 10	Mid-levels diff.	Mid-levels diff. (mean)
4	Item 1 (R1)	2257	0.32	0.60							0.28	0.24
	Item 2 (R1)	4194	0.36	0.57							0.21	
	Item 3 (R1)	4194	0.32	0.55							0.23	
	Item 4 (R1)	2257	0.29	0.51							0.22	
	Item 5 (R1)	4194	0.28	0.52							0.24	
5	Item 6 (R1)	3875		0.45	0.67						0.22	0.24
	Item 7 (R1)	3875		0.42	0.68						0.26	
	Item 8 (R1)	3875		0.39	0.66						0.27	
	Item 9 (R1)	3875		0.40	0.63						0.23	
	Item 10 (R1)	3875		0.32	0.55						0.23	
6	Item 11 (R1)	1225			0.43	0.68					0.25	0.23
	Item 12 (R3)	1225			0.43	0.65					0.22	
	Item 13 (R2)	2257			0.42	0.64					0.22	
	Item 14 (R2)	2257			0.37	0.60					0.23	
	Item 15 (R3)	1225			0.36	0.59					0.23	
7	Item 16 (R3)	2257				0.49	0.70				0.21	0.18
	Item 17 (R4)	3714				0.44	0.60				0.16	
	Item 18 (R1)	3875				0.43	0.59				0.16	
	Item 19 (R4)	2455				0.43	0.59				0.16	
	Item 20 (R3)	1225				0.39	0.58				0.19	
8	Item 21 (R4)	1599					0.47	0.59			0.12	0.14
	Item 22 (R3)	4049					0.46	0.60			0.14	
	Item 23 (R2)	2382					0.44	0.60			0.16	
	Item 24 (R3)	4094					0.44	0.58			0.14	
	Item 25 (R4)	2455					0.42	0.55			0.13	
9	Item 26 (R4)	3714						0.49	0.68		0.19	0.15
	Item 27 (R4)	3714						0.48	0.63		0.15	
	Item 28 (R3)	4123						0.46	0.60		0.14	
	Item 29 (R4)	1742						0.47	0.58		0.11	
	Item 30 (R4)	1599						0.42	0.58		0.16	
10	Item 31 (R3)	2257							0.49	0.66	0.17	0.14
	Item 32 (R4)	1742							0.47	0.60	0.13	
	Item 33 (R4)	5856							0.43	0.56	0.13	

Item 34 (R4)	1742	0.40	0.54	0.14
Item 35 (R4)	3714	0.45	0.58	0.13

Note. CELPIP levels 4 to 10 correspond to CLB 4 to 10. N=number of test taker responses to that item, mid=middle point of the score level, and diff.=difference.

As shown in Table 1, the anchor items for the lower proficiency levels (CLB 4 and 5) were more likely to be items from CELPIP reading testlet types 1 and 2, and the anchor items for the higher proficiency levels (CLB 9 and 10) mostly belonged to CELPIP reading testlet types 3 and 4. This pattern is consistent with the specifications for the CELPIP reading testlets, which dictate the target contexts, language functions, and intended difficulty of the four testlet types. Table 2 provides a brief description of each of the four testlet types in the CELPIP reading test.

Table 2

Structure of the CELPIP-General Reading Test

Testlet Type	Name	Description
R1	Reading Correspondence	First, read a letter or email and answer a set of multiple-choice questions; then, read a response and complete the blanks with the provided options.
R2	Reading to Apply a Diagram	First, read a general text with accompanying graphics and complete an email with the provided options; then, answer a set of multiple-choice questions.
R3	Reading for Information	First, read a general text and a set of statements; then, decide which paragraph (if any) supports each statement.
R4	Reading for Viewpoints	First, read an opinion article and answer a set of multiple-choice questions; then read a response and complete the blanks with the provided options.

Note. To view a sample of the test content, follow the link below: https://secure.paragontesting.ca/InstructionalProducts/FreeOnlineSampleTest/FOST/View/1ba67a01-763a-487c-9efe-5000023fe7b4

To aggregate panellists' judgements, we computed both the mean and the median of their ratings. While the mean accounts for all panellists' judgements, it can be distorted by outliers. In contrast, the median is more robust in the presence of outliers but does not account for everyone's opinion. As shown in Table 3, the mean and median ratings are very similar to each other for each level. That said, using means, rather than medians, makes it easier to interpret the variability of the ratings (i.e., standard deviations in Table 3) and the mean differences (i.e., the last two columns of Table 3). Thus, we used the mean ratings to represent the panellists' collective judgments for each item, and for consistency, we also used the mean to summarize the panellists' judgments for all the items within a level. The difference between the panellists' aggregate judgement and the model-suggested anchoring level was calculated in both overall and absolute terms.

Table 3

Comparison Between Model-Suggested Anchoring Levels and Panellists' Judgements

Model-Suggested CELPIP Level	Panel-Estimated CLB Level			Mean of Difference between Panel Estimated CLB Level and CELPIP Level	Mean of Absolute Difference between Panel Estimated CLB Level and CELPIP Level
	Median	Mean	Standard Deviation		
4	4.75	5.00	0.61	1.00	1.00
5	5.00	4.90	0.45	-0.10	0.30
6	6.00	6.40	1.01	0.40	0.70
7	8.00	7.95	1.79	0.95	1.55
8	8.25	8.00	1.77	0.00	1.30
9	8.50	8.45	0.62	-0.55	0.55
10	9.25	9.15	0.66	-0.85	0.85
Average difference for levels 4 through 10				0.12	0.89

Note. The Panel Estimated CLB Level for each CELPIP level was calculated by averaging the Panel Estimated CLB Levels for the five anchor items. The Panel Estimated CLB Level for each item (not presented in this table) was the mean of the final ratings of the item submitted by the four panellists.

Overall, the panel-estimated CLB levels increase along with the increase in the items' anchoring levels suggested by the model. However, there were some minor discrepancies. The panellists believed that the five anchor items at Level 4 tapped higher-level knowledge and skills. For the items anchoring at Level 10, the panellists considered them slightly easier than that described by the CLB 10 descriptors. For levels 5, 6, 8, and 9, the average difference within each level was small, indicating that the item content and their corresponding performance levels (as judged by experts) were in line with the anchor levels identified by the statistical analysis of test taker responses. Compared to their judgements at other levels, the panellists showed higher variability in their views at levels 7 and 8, suggesting a potentially inconsistent interpretation of the criteria for CLB 7 and 8 and/or the correspondence between the item content and those criteria.

Discussion

Using a modified scale-anchoring method, this study evaluates the content validity of the CELPIP-General reading test by mapping a selected set of items to their corresponding CLB levels. The CLB provide a nationally recognized set of language proficiency indicators that essentially describe the CELPIP target language use (TLU) domain, namely English as a second language use in work, study, and social contexts in Canada. A close alignment between the CELPIP test items and the criteria of the CLB is

one critical piece of evidence in support of the claim that the test reflects its target language use domain and measures it intended construct.

As discussed in the introduction, the link between test scores and external proficiency frameworks should *not* be taken for granted; rather, it is a claim that must be supported by empirical evidence. One way to assert such a claim is to conduct independent validation studies to collect evidence from various sources, including evidence related to test content. In this study, we presented a method that may be useful in the planning and implementation of content validation studies for large-scale tests. In the context of large-scale tests, content validation studies often rely on experts evaluating a portion of the test content. Having a strategy that systematically selects items for expert review enhances the transparency and replicability of such validation studies. Future studies could apply our method to other tests and content areas.

Future research could also examine other criteria for selecting anchor items. For example, instead of focusing on *typical* or *average* test takers, it is possible to conceptualize a *minimally competent* test taker for each level and identify anchoring items accordingly. In the literature concerning scale anchoring methods, cut-offs other than 0.50 have also been suggested for the conditional probability, $P(X_i = 1|\theta_k)$, when deciding the anchor level of an item (e.g., 0.65 and 0.80; Beaton & Allen, 1992). Over time, with evidence accumulated from a wider application of this method, researchers and practitioners will develop a better understanding of how to optimize these parameters for content validation studies.

Admittedly, as a relatively small-scale study (i.e., 35 items and four panellists), the present study alone serves as just one piece of validity evidence in support of the alignment between the CELPIP-General reading test content and the CLB proficiency indicators. Ongoing efforts to continue this line of research will further strengthen the link between CELPIP test scores and the criteria of the CLB. Working with a small group of experts with rich experience allowed us to create a proof of concept to test the modified scale anchoring method for collecting content-based validity evidence. According to the general principles of qualitative studies, an adequate sample is achieved when researchers observe sufficient variability and an indication of convergence (i.e., saturation) in the study (Fusch & Ness, 2015; Guest et al., 2006). The degree of variability and agreement among panellists' evaluations of the items (as shown in Table 3) lend some support to the credibility of the results. That said, when resources allow, it could be beneficial to repeat the study and involve more experts in the review panel to represent a broader range of views.

Both the linking of test scores to proficiency frameworks such as the CLB and the validation of such alignment help strengthen the validity of the scores. In doing so, researchers and test developers often rely on expert judgements. In the present study, we recruited experts who had been working with the CLB in teaching and assessment settings for many years; however, we observed some differences in their interpretations of the CLB when reviewing the anchor items. From our perspective, these differences may be attributable to four main issues concerning the CLB descriptors: (1) variation in word choice or synonymy across the benchmarks, (2) under-defined terminology, (3) limitations in the operationalization of key features (e.g., reading text length as specified by the CLB – "moderate length" at CLB 8 means "up to about 5 pages" (CCLB, 2012, p.96) – may be unattainable in some assessment contexts), and (4) concerns about the cultural context.

Some of these issues mirror those reported by Alderson and colleagues (2006) during a project to develop a CEFR-based reading and listening assessment tool. Future content validation studies could consider adding a training session at the beginning of the panel meeting to address these issues and ensure a shared understanding among the panellists.

The challenges involved in applying a general language proficiency framework to the development and/or validation of a standardized test are at least partly due to the conflicting nature of the two. While language proficiency frameworks are often designed to account for a wide range of contexts and uses, a test often serves more specific purposes within more limited contexts. In our case, the commonalities between the target domains of the CELPIP test and the CLB permit the use of the CLB for the purpose of test score interpretation and content validation, and even to inform some aspects of on-going test development; however, we must not ignore the differences between the two. The CLB are neither a test nor a test blueprint; they are a set of general proficiency indicators that describe the progression in the features of language and contexts of use across the proficiency spectrum. It is expected that teachers, test developers, and researchers will make the necessary judgements, modifications, and adaptations when applying the CLB to more specific instructional and/or assessment contexts.

Conclusion

The relationship between the test scores and the proficiency frameworks or standards to which they are aligned is *not* directly observable, nor is it a constant connection. It is an indirect relationship that test developers must continuously provide evidence to support. In addition, it is important to consider that high-stakes language proficiency tests can affect the lives of many individuals, such as students and immigrants, as critical decisions are made based on the results of these tests. Therefore, it is crucial for test organizations to actively evaluate and maintain their tests in order to ensure consistent alignment with the chosen proficiency standards over time.

In this paper, we describe one approach to test validation focusing on content validation and score interpretation using a process of scale anchoring and item mapping. In doing so, we share some of the benefits and challenges of working with the CLB in a standardized testing context. Although we have primarily worked with the descriptors of ability to help infuse detail and substance into the interpretation of the CELPIP score levels, the CLB offer much more to language practitioners. We encourage those engaged in the instruction or assessment of English as a second language in Canada to consider how the CLB might support their work.

Correspondence should be addressed to Michelle Chen.
Email: mchen@paragontesting.ca

Acknowledgements

The authors would like to thank Tingfeng Fu for her assistance in collecting data during the study and Ben Unterman for his help in composing the French abstract.

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly: An International Journal, 3*(1), 3-30.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R.L. Thomdike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). American Council on Education.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191–204.

Bournot-Trites, M. (2017). Design-based research methodology for establishing the common theoretical framework and the CLB/NCLC scales. In Monika Jezak (Ed.), *Language is the key. The Canadian Language Benchmarks model* (pp. 31-53). University of Ottawa Press.

Bournot-Trites, M., Barbour, R., Jezak, M., Stewart, G., & Blouin Carbonneau, D. (2015). *The Theoretical Framework for the Canadian Language Benchmarks and Niveaux de compétence linguistique canadiens*. Centre des niveaux de compétence linguistique canadiens.

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. Language Training and Testing Centre.

Bruni, L., & Irwin, P. (2007). The Canadian Language Benchmarks Placement Test. *The Canadian Modern Language Review/La revue canadienne des langues vivantes, 64*(1), 220-226.

Brown, J. D. (1996). *Testing in language programs*. Prentice Hall Regents.

Centre for Canadian Language Benchmarks (CCLB). (2012). *Canadian Language Benchmarks: English as a second language for adults.* https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf

Centre des niveaux de compétence linguistique canadiens. (2012). *Niveaux de compétence linguistique canadiens: Français langue seconde pour adultes (3e Éd.).* https://www.canada.ca/content/dam/ircc/migration/ircc/francais/pdf/pub/competence-linguistique.pdf

Chen, M. Y. (2016). *Unpublished report on the standard setting for the revised CELPIP-General listening test*. Paragon Testing Enterprises.

Dorans, N. J. (2018). Scores, Scales, and Score Linking. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary*

reference on survey, scale and test development (Vol II, pp. 573-606). Wiley. https://doi.org/10.1002/9781118489772.ch19

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual*. Council of Europe. https://biblio.ugent.be/publication/4270320/file/4270333

Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report, 20*(9), 1408-1416.

Gomez, P. G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*(3), 417-444.

Government of Canada (2018). *Language classes funded by the Government of Canada*. https://www.canada.ca/en/immigration-refugees-citizenship/services/new-immigrants/new-life-canada/improve-english-french/classes.html

Government of Canada (2020). *Language testing—skilled immigrants (Express Entry)*. https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/express-entry/documents/language-requirements/language-testing.html

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82.

Hajer, A., Robinson, J., & Witol, P. (2002). *LINC curriculum guidelines: Language instruction for newcomers to Canada*. Toronto Catholic District School Board.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). American Council on Education and Praeger.

Jaeger, R. M. (2003). Reporting the Results of the National Assessment of Educational Progress. NAEP Validity Study. *Working Paper Series*. https://files.eric.ed.gov/fulltext/ED478972.pdf

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461.

Kane, M. T. (2012) Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3–17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. [Unpublished doctoral dissertation, Boston College].

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 219-248). Erlbaum.

Liao, C. W. (2010). *TOEIC listening and reading test scale anchoring study*. http://www.ets.org/Media/Research/pdf/TC-10-05.pdf

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448.

Liu, J., & Walker, M. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109-134). Springer-Verlag.

North, B., & Jones, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling*. Council of Europe. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.3003&rep=rep1&type=pdf

North, B., & Piccardo, E. (2018). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of References (CERF)*. Centre for Canadian Language Benchmarks. https://www.language.ca/wp-content/uploads/2019/01/Aligning-the-CLB-and-CEFR.pdf

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly, 13*(2), 109–123.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between the TOEFL iBT test scores and the Common European Framework of Reference (CEFR) Levels* (Research Memorandum No. RM-15-06). Educational Testing Service.

Paragon Testing Enterprises (2013a). *Unpublished report on assigning CELPIP-General test scores to the Canadian Language Benchmarks (CLB)*. Paragon Testing Enterprises.

Paragon Testing Enterprises (2013b). *Unpublished report on standard-setting procedures used to set cut-scores for the CELPIP-General listening and reading tests*. Paragon Testing Enterprises.

Philips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). *Interpreting NAEP scales*. U.S. Department of Education, National Center for Education Statistics.

Savignon, S. J. (1991). Communicative language teaching: State of the art. *TESOL Quarterly, 25*(2), 261-278.

Savignon, S. J. (1997). *Communicative competence: Theory and classroom practice (2nd Edition)*. McGraw-Hill.

Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review, 15*, 21-25.

Tannenbaum, R. J., & Wylie, E. C. (2004). *Mapping test scores onto the Common European Framework: Setting standards of language proficiency on the Test of English as a Foreign Language (TOEFL), the Test of Spoken English (TSE), the Test of Written English (TWE) and the Test of English for International Communication (TOEIC).* Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Report RR–08–34). Educational Testing Service.

Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan