CELPIP®

CELPIP Test Review Report I :

Test Content, Development, and Validation

Test Research and Development

Paragon Testing Enterprises

April 2021

CELPIP Test Review Report I :

Test Content, Development, and Validation

Test Research and Development

Paragon Testing Enterprises

April 2021

Citation:
Paragon Testing Enterprises. (2021, April). *CELPIP Test Review Report I: Test Content, Development, and Validation* (No. 2021–1). Vancouver, BC: Paragon Testing Enterprises.

Table of Contents

List of Tables

List of Figures

1. About the CELPIP Test

This report provides an overview of the Canadian English Language Proficiency Index Program (CELPIP) Test, including its structure, content, operations, and the ongoing validation work to illustrate how the test continues to produce evidence to evaluate the abilities corresponding to a broad range of proficiency in reading, writing, listening, and speaking in functional English.

The CELPIP Test is a complete English language testing program designed to measure the functional language proficiency required for successful communication in general Canadian social, educational, and workplace contexts. The test is completely computer-delivered and can be taken in one of Paragon's designated test centres at computer stations. The CELPIP-General Test consists of four components: listening, reading, writing, speaking. The CELPIP-General LS Test consists of two of the four components: listening and speaking, with the same structure and specifications of the listening and speaking components of the CELPIP-General Test. The listening and reading components of the test are computer-scored, while the writing and speaking components are recorded and rated via an online rating system by trained raters located across Canada.

Paragon Testing Enterprises has a Service Agreement with the Government of Canada to deliver language testing services for economic immigration. CELPIP has been designated by the Immigration, Refugees and Citizenship Canada (IRCC) as evidence of English language proficiency for applications for permanent residence immigration in Canada and Canadian citizenship.

2. Test Construct and Validity Framework

The CELPIP Test is designed to measure the functional English proficiency required for successful communication in social, educational, and workplace contexts in Canada. Functional English proficiency is defined as the ability to integrate language knowledge and skills in order to perform various social functions. As such, the CELPIP Test construct fits within Bachman and Palmer's (1996, 2010) theoretical model of communicative language ability, the foundation of the Canadian Language Benchmark (CLB) framework (CCLB, 2012).

CELPIP component scores are reported on 11 bands: M, 3–12. CELPIP test scores have been calibrated against the Canadian Language Benchmark (CLB) levels to evaluate abilities corresponding to a broad range of proficiency in listening, reading, writing, and speaking in functional English.

Table 1 shows each CELPIP level and its corresponding description, with the CLB level equivalencies included in the third column.

Table 1. CELPIP Levels and Corresponding CLB Levels

CELPIP LEVEL	CELPIP DESCRIPTOR	CLB LEVEL
M	Minimal proficiency or insufficient information to assess	1 and 2
3	Some proficiency in limited contexts	3
4	Adequate proficiency for daily life activities	4
5	Acquiring proficiency in workplace and community contexts	5
6	Developing proficiency in workplace and community contexts	6
7	Adequate proficiency in workplace and community contexts	7
8	Good proficiency in workplace and community contexts	8
9	Effective proficiency in workplace and community contexts	9
10	Highly effective proficiency in workplace and community contexts	10
11	Advanced proficiency in workplace and community contexts	11
12	Advanced proficiency in workplace and community contexts	12

Paragon Testing Enterprises has adopted the widely used description of test validity by the *Standards for Educational and Psychological Testing* (hereafter, *Test Standards*, AERA, APA, & NCME, 2014).

> *"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (Test Standards, p.11).*

In terms of the CELPIP Test, validity refers to the degree to which evidence supports the interpretation made by test users from the CELPIP Levels (and corresponding CLB Levels) on the listening, reading, writing, and speaking components of the CELPIP regarding CLB levels 3 through 12. In short, test validation is a process that results in a compelling body of evidence that the test does what it was designed to do, that the design was good, and that the resulting scores have their intended meaning.

To that end, Paragon has continued to engage in test development and research in order to stay current to the assessment context and collect evidence to support the intended use and interpretation of CELPIP test scores. After reporting on the test structure, content, and operations, we present the ongoing validation work on the CELPIP-General Test, including conferences, publications, online studies at Paragon website, and ongoing research (Section 6).

While Paragon's approach to validity is informed by the *Test Standards,* we believe that there are multiple methods and approaches to the collection of validity evidence, instead of one model of validation. As such, Paragon does not follow a single theory of validity to the exclusion of others.

3. Test Structure and Test Specifications

The CELPIP-General Test consists of four components: listening, reading, writing, speaking. The CELPIP-General LS Test consists of two of the four components: listening and speaking, with the same structure and specifications of the listening and speaking components of the CELPIP-General Test. Table 2 below describes the format and content of each CELPIP test component.

Table 2. Structure of the CELPIP Test

COMPONENT	NUMBER OF TESTLET TYPES	OPERATIONAL ITEMS/TASKS
LISTENING	6	38
READING	4	38
WRITING	2	2
SPEAKING	8	8

Table 2 includes only operational testlet types and items/tasks. The listening and reading components may include non-scored field test testlets currently undergoing final stage revision prior to deployment in operational forms (see Section 4.2 for the Item review, editing, and field-testing processes). Field test testlets are not treated as part of an operational form and are delivered through a system that is autonomous from the form construction and assignment system. Hence, two test takers who were assigned the same test form may nonetheless encounter different field test testlets.

3.1. Listening Component

The listening component of the test is based on the CLB 2012. As such, it measures the construct of functional listening proficiency in English. Functional listening proficiency is defined as an individual's ability to engage with, understand, and respond to spoken English so as to achieve day-to-day and general workplace communicative functions. Communicative functions refer to the use of language to convey ideas, influence the actions of other people, and participate effectively in diverse social and workplace contexts.

The CELPIP listening component is composed of six types of listening testlets (Table 3). The listening form includes one operational testlet of each type. All six listening testlets require test takers to engage with a spoken text delivered as an audio recording through a set of headphones. The first testlet (L1) is delivered as a series of three short segments and involves images along with audio. The fifth testlet (L5) involves video as well as audio. All six testlets require test takers to respond to either open or closed stem multiple choice questions related to the text. For all six testlets, the audio recording is played first before the questions are available. For the first three testlets (L1, L2, L3), the question stems are played as audio recordings one after another with the corresponding response options for each item appearing concurrently on the screen. For the fourth, fifth, and sixth testlets (L4, L5, L6), the item stems and response options appear together as a set of items on the test screen following the audio or video recording.

Table 3. Structure of the Listening Component

TESTLET TYPE	TESTLET NAME	TESTLET TIME	ITEMS
L1	Listening to Problem Solving	8 minutes	8 items
L2	Conversation One	5 minutes	5 items
L3	Conversation Two	6 minutes	6 items
L4	Listening to a News Item	4 minutes 30 seconds	5 items
L5	Listening to a Discussion	8 minutes 30 seconds	8 items
L6	Listening to Viewpoints	7 minutes 30 seconds	6 items

The CELPIP listening component assesses general listening proficiency from CLB 2012 levels 3 through 12. Each testlet type is designed to represent specific language functions in English.

3.2. Reading Component

The CELPIP reading component is composed of four types of reading testlets (Table 4). The reading form includes one operational testlet of each type. All four reading testlets require test takers to interact with a written text presented on a computer screen. The second testlet (R2) involves graphics as well as text. All four testlets require test takers to respond to either open or closed stem multiple choice questions related to the text. The text and corresponding items are available to test takers concurrently on the same test screen.

Table 4. Structure of the Reading Component

TESTLET TYPE	TESTLET NAME	TESTLET TIME	ITEMS
R1	Reading Personal Correspondence	11 minutes	11 items
R2	Reading to Apply a Diagram	9 minutes	8 items
R3	Reading for Information	10 minutes	9 items
R4	Reading for Viewpoints	13 minutes	10 items

The CELPIP reading component assesses general reading proficiency from CLB 2012 levels 3 through 12. Each testlet type is designed to represent specific language functions in English.

3.3. Writing Component

The CELPIP writing component is composed of two types of writing testlets (Table 5). The writing form includes one operational testlet of each type. Test takers read the instructions and writing prompt on the computer screen, and then write and edit their response to each writing task on the computer within the allotted time frame (a maximum of 27 minutes for W1 and 26 minutes for W2). A word count and basic spell check function (similar to Microsoft Word) is provided.

Table 5. Structure of the Writing Component

TESTLET TYPE	TESTLET NAME	TESTLET TIME	RESPONSE LENGTH
W1	Writing an Email	27 minutes	150-200 words
W2	Responding to Survey Questions	26 minutes	150-200 words

The CELPIP writing component assesses general writing proficiency from CLB 2012 levels 3 through 12. Each testlet type is designed to represent specific language functions in English. All writing tasks are rated by human raters according to a standardized analytic scoring rubric. At minimum, four raters rate a test taker's writing responses. The scoring rubric is proprietary and not available for public access; some interpretations of the writing assessment scale dimensions are presented in Table 6.

Table 6. Interpretations of the Writing Assessment Scale Dimensions

DIMENSION	DESCRIPTION
COHERENCE/MEANING	This dimension evaluates the extent to which the test taker's purpose is clearly and effectively conveyed in the written response. This is described in terms of how well the writer's ideas communicate a coherent purpose, and whether meaning is developed in depth and supported with relevant, precise, and accurate information.
LEXICAL RANGE	This dimension evaluates the range of expressions that the test taker uses to address the task. The writer's lexical range is described in terms of its precision and effectiveness for completing the task. Depending on the context of the task, this may or may not imply the use of abstract, conceptual or highly specialized vocabulary.
READABILITY	This dimension evaluates the extent to which the written response is delivered through clear, intelligible, and fluent writing. This is described in terms of the comprehensibility of the writing as supported by the accuracy and effectiveness of language forms, including grammar, syntax, orthography, formatting, and cohesive and transitional devices.
TASK FULFILLMENT	This dimension evaluates the extent to which the task is completed. A completed task is described as one in which the response sufficiently and efficiently addresses all aspects of the task. This dimension also accounts for the use of appropriate tone and register within the context of the task.

3.4. Speaking Component

The CELPIP speaking component is composed of eight types of speaking testlets (Table 7). The speaking form includes one operational testlet of each type. Each task provides the test takers with a preparation period (ranging from 30 to 60 seconds) in which they prepare an oral response to the question or situation presented. The test takers then provide their responses orally (within the allotted timeframe ranging from 60 to 90 seconds), which are recorded by a microphone on their computer headsets.

Table 7. Structure of the Speaking Component

TESTLET TYPE	TESTLET NAME	PREPARATION TIME	SPEAKING TIME
S01	Giving Advice	30 seconds	90 seconds
S02	Talking about a Personal Experience	30 seconds	60 seconds
S03	Describing a Scene	30 seconds	60 seconds
S04	Making Predictions	30 seconds	60 seconds
S05	Comparing and Persuading	60 seconds	60 seconds
S06	Dealing with a Difficult Situation	60 seconds	60 seconds
S07	Expressing Opinions	30 seconds	90 seconds
S08	Describing an Unusual Situation	30 seconds	60 seconds

The CELPIP speaking component assesses general speaking proficiency from CLB 2012 levels 3 through 12. Each testlet type is designed to represent specific language functions in English. All speaking tasks are rated by human raters according to a standardized analytic scoring rubric. At minimum, three raters rate a test taker's speaking responses. The scoring rubric is proprietary and not available for public access; some interpretations of the speaking assessment scale dimensions are presented in Table 8.

Table 8. Interpretations of the Speaking Assessment Scale Dimensions

DIMENSION	DESCRIPTION
COHERENCE/MEANING	This dimension evaluates the extent to which the test taker's purpose is clearly and effectively communicated in the speech. This is described in terms of the depth and precision of the meaning expressed, and the coherence of the overall response, i.e., how well the speaker communicates a coherent message and develops ideas with relevant, precise, and accurate supporting information.
LEXICAL RANGE	This dimension evaluates the extent to which the test taker successfully employs lexical tools in the response. This is described in terms of the accuracy and range of word choices employed by the test taker as well as the naturalness, suitability, and appropriateness of the chosen lexical units to the context of the task.
LISTENABILITY	This dimension evaluates the clarity and ease with which the provided response is understood by the listener. This is described in terms of the clarity and accuracy of the pronunciation, the naturalness of rhythm and intonation, the fluency and intelligibility of speech, as well as the use of grammar and structures to support the intelligibility of the ideas.
TASK FULFILLMENT	This dimension evaluates the degree to which the test taker's response addresses the task requirements. This involves an evaluation of the extent to which the response directly relates to the question or situation presented, develops the task in depth rather than simply lists relevant points, and uses an appropriate tone/register for the task.

4. Test Item Development

Paragon administers the development of test items through a careful and meticulous process to ensure content accuracy, fairness, and accessibility. This is enabled by a rigorous item writer training, monitoring, and feedback program, and a cycle of item review, editing, and field-testing.

4.1. Item Writer Training

Item writer training occurs in two main stages (Figure 1). The first stage of training involves a program of self-guided reading including procedural guides and test development documents. Procedural guides assist item writers in learning the various systems and procedures involved in remote item writing work.

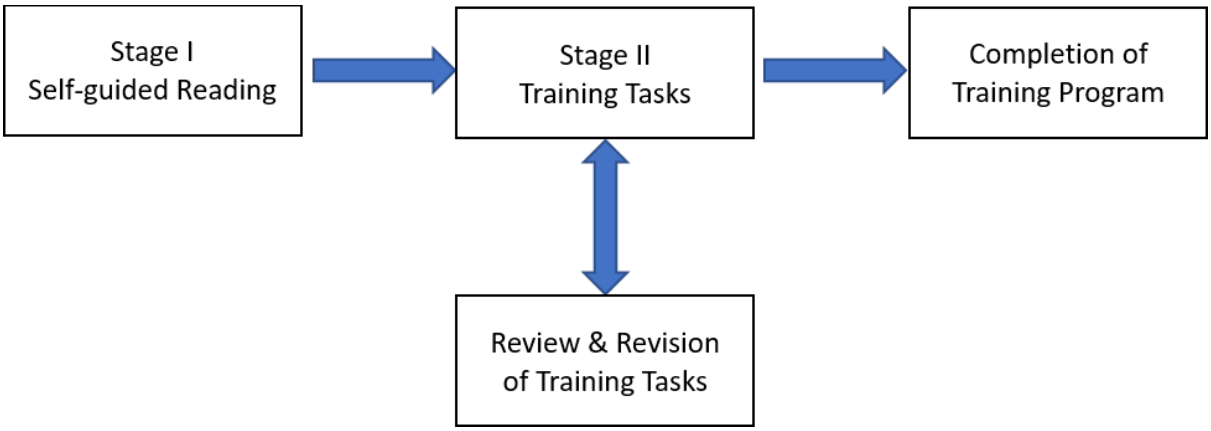

Figure 1. CELPIP Item Writer Training Program

The second stage of training involves the completion of a series of training tasks that simulate actual test tasks. Several rounds of review and revision occur until the tasks are deemed satisfactory by the internal reviewers. The feedback and revision process also provides an opportunity for dialogue between the remote item writers and Paragon's content development team to address any questions or concerns. Item writer training is compensated and a typical training program will last several months. Successful completion of the program is a prerequisite for receiving commissioned item writing work.

4.2. Item Review, Editing, and Field-Testing Processes

The CELPIP Test is designed in a testlet format in which each language component (listening, reading, writing, speaking) includes several different testlet types. Each testlet is composed of either a passage and a corresponding set of items (for listening and reading components) or a prompt (for writing and speaking components). All testlets go through a rigorous process of review and revision before they appear on operational test forms.

Upon submission of a testlet by an item writer, the testlet content is reviewed by a team of editors who make various rounds of revisions to the passages, items, and prompts before the testlet is sent for field testing. These revisions are made for the purpose of ensuring high quality, effective, and fair test content. This process includes revisions to ensure strong alignment of content with test specifications, an appropriate level of difficulty for the target proficiency range, and adherence to item and prompt writing principles.

Upon completion of the review and revision process, CELPIP listening and reading testlets go through several rounds of field testing. Item performance statistics for listening and reading testlets are collected and analyzed after each round of field testing. These statistics indicate the difficulty as well as the discriminatory power (i.e., the ability to distinguish between high and low proficiency test takers) of each item in a testlet.

The difficulty of an item (p-error) is measured by the proportion of test takers who endorsed the key. Therefore, the higher the p-error value, the lower the difficulty. Different target difficulty parameters have been established for each testlet type in accordance with the target proficiency ranges.

Item discrimination (T-Rpbis) is measured by a correlation between test takers' success on the individual item and their ability in the related test component (listening, reading, writing, speaking) overall. Therefore, the higher the T-Rpbis value, the greater the discriminatory power of the item. Standard thresholds of T-Rpbis have been established to determine if an item meets the requirement for discriminatory power.

These item performance statistics collected after each round of field-testing inform further revisions by a team of post field test editors who are trained to interpret the statistics and make the appropriate revisions. Once items have met performance standards in terms of both difficulty and discrimination parameters, they are banked for operational testing. Items that fail to meet these statistical requirements after a certain number of field-testing rounds are decommissioned and not used for operational testing.

4.3. Rater Feedback on Prompt Quality

For CELPIP writing and speaking testlets, a feedback loop has been established that allows rater feedback on writing and speaking prompts to be relayed directly to the content development team for internal review and revision.

Raters are encouraged to provide feedback to Paragon if they have concerns about a writing or speaking prompt that they encounter in rating. In this manner, the raters may act as a quality assurance check. At times, they bring attention to issues that may not have been previously identified. These instances are rare, but they may include:

- errors in spelling or grammar
- elements that may confuse a test taker
- identification of multiple test takers misunderstanding or misinterpreting a prompt

5. Quality Control Procedures for Human Rating

CELPIP listening and reading components are assessed through multiple choice questions and are machine scored. The writing and speaking components are assessed by human raters using scoring rubrics developed by Paragon. Paragon maintains the quality of rating through the recruitment of highly qualified individuals as raters and the implementation of a rigorous program of rater certification, training, feedback, and monitoring program.

5.1. CELPIP Rater Pool

Paragon employs a team of remote writing and speaking raters who rate CELPIP test-taker responses through an online rating system. The raters are recruited from highly qualified individuals who must possess the following qualifications:

- 4-year bachelor's degree (or higher) from an accredited post-secondary institution
- 3 years of work experience in ESL teaching, language education, or a field closely related to linguistics or language assessment
- ESL teaching certification recognized by TESL Canada or completion of graduate training in language education or in linguistics
- native English speaker or evidence of CLB 11/12 English language proficiency
- Canadian citizenship, permanent residency, or a valid work permit in Canada

5.2. Rater Training and Feedback

Paragon implements a rigorous rater training, feedback, and monitoring program to ensure rating accuracy and consistency, as illustrated by Figure 2 and described in Sections 5.2.1 - 5.2.2.

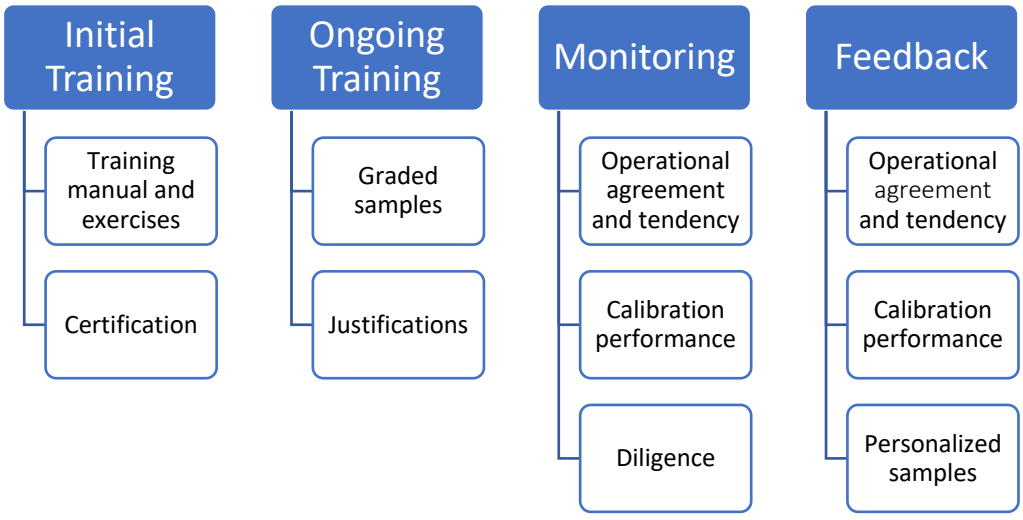

Figure 2. Rater Training and Feedback Procedures

5.2.1. Initial Rater Training and Certification

Raters must complete initial training and pass a certification test before they can rate operationally. All raters rate only one component, either writing or speaking, and have access to the relevant materials for that component only.

The initial training process requires trainees to work through a training guide that introduces Paragon's approach to language assessment, the construct of functional proficiency, the CELPIP theoretical framework, and an explanation of each dimension of the rating scale with real examples and exercises that allow the trainees to actively engage with the rating process. In addition, they provide training on high-stakes standardized testing and instruction on the use of the online rating system.

Rater trainees are required to certify upon completion of training through the online rating system. To certify, trainees must complete and meet the required agreement threshold for operational rating on multiple certification sets, which consist of authentic test-taker responses across a range of proficiency levels.

Newly certified raters are required to complete a 90-day probationary period, during which they undergo monthly performance evaluations, and may be terminated for any performance concerns.

5.2.2. Ongoing Rater Training and Monitoring

Paragon provides regular and ongoing training and feedback to all raters to ensure that they adhere to CELPIP rating guidelines and apply the rating scale consistently and accurately when assessing the speaking and writing responses.

Rater training materials are compiled and reviewed by Paragon's in-house rating specialists with input from experienced benchmark raters to maintain the quality and consistency of rating. Benchmark raters are selected from raters who demonstrate consistently strong performance and a clear understanding of the rating principles. Benchmark raters rate assignments that generate disagreement in the primary round of rating (see Section 5.3) and participate in activities that generate training materials for the rating pool as a whole, including graded rating samples, written justifications to rating, and seminar discussions of rating principles and specific responses.

Training is provided both quantitatively and qualitatively. Graded samples with recommended ratings are provided regularly to ensure that raters stay calibrated to the scale. In addition, written justifications to recommended ratings are provided as a qualitative review to connect the recommended ratings to characteristics of the performance and the descriptors of the rating scale.

Paragon conducts regular rater performance monitoring of all operational rating to ensure that any deficiencies in rater behavior are identified and remediated. Rating data is also collected and analyzed via regular rater calibrations seeded in operational assignments and distributed to all raters. The results from the regular rater monitoring are transformed into ongoing feedback of rater performance, including:

- *Operational performance*: Raters receive regular performance reports documenting their agreement with other raters in operational rating, and whether they demonstrate any rating tendency towards leniency or severity.
- *Calibration performance:* Calibration performance reports are issued comparing the rater's rating and the recommended rating and identifying the rater's tendency relative to the rating pool. Justifications to calibration ratings are also provided periodically as qualitatively feedback.
- *Personalized feedback:* In addition to agreement and tendency feedback, Paragon may also provide additional feedback samples upon individual rater's request. The personalized feedback includes samples in which the rater is highly discrepant with a benchmark rater on the same assignment, curated to target areas identified as problematic in the monthly performance review.

All raters are expected to attend to training materials regularly and address any performance issues identified and communicated through rater performance feedback. Failure to adhere to the performance standards may result in termination of employment.

5.3. Double Rating and Benchmark Rating

Quality control of CELPIP rating is operationalized through double rating and benchmarking procedures. In the primary round of rating, each of the two writing responses is rated by two writing raters, and the speaking responses are rated by three raters, each rating a subset of partially overlapped responses, with the four most discriminated speaking responses double rated. If the ratings generate disagreement in the primary round, the responses will be automatically sent for benchmarking. Benchmark rating procedures ensure that assignments that have generated disagreement in the primary round are reviewed and rated by additional raters who have demonstrated high consistency in their rating.

6. Recent Research Supporting the Validity of the CELPIP Test

As mentioned in Section 2, Paragon Testing Enterprises has adopted the widely used *Test Standards*' (AERA, APA, & NCME, 2014) description of test validity, repeated below:

> *"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (Test Standards, p.11).*

Test validation is not an activity that occurs once the assessments are developed and only one time, but rather is an ongoing process that is initiated at the beginning of assessment design and continues throughout development, implementation, and operation (Messick, 1995, Zumbo, 2007b).

The evidence reported herein supports the validity and reliability of the interpretations made using CELPIP about test-takers' proficiency in listening, reading, writing, and speaking in functional English. Paragon continues to collect additional robust empirical evidence for the reliability and validity that goes over and above what is typically reported in validation studies (see summaries of commonly reported validity evidence in Zumbo & Chan, 2014). In the sections that follow, a summary of this additional research evidence is provided, each contributing to a piece of validity evidence to support the use and interpretation of the CELPIP Test.

6.1. Validation Studies with a Focus on Response Processes

During a test, test takers actively engage with the questions and tasks. By investigating how test takers use different strategies to complete the test, Paragon could better understand their test-taking experiences and, equally importantly, could collect (or identify a lack of) process-based validity evidence that supports appropriate score interpretation and use.

Response-process-based evidence has been viewed as an essential source to score validity (AERA, APA, & NCME, 2014; Messick, 1995; Zumbo & Hubley, 2017). From 2017 to 2020, two peer-reviewed journal articles and a book chapter have been published surrounding this topic. The papers by Wu & Zumbo (2017) and Wu, Chen, & Stone (2018) focused on reading strategies, using data from the CELPIP-General reading test. Chen, Wu, & Liu (2020), on the other hand, studied the test-taking process of two listening testlet types that are used in the CELPIP-General Test.

Using a variety of statistical methods, these studies analyzed self-report response processes and strategies and investigated the relationship between different types of processes/strategies and test-takers' performance. The results show that to complete CELPIP-General reading or listening questions, test takers frequently used processes and strategies that are directly associated with comprehending meanings (e.g., understand/summarize the key information; Chen, Wu & Liu, 2020; Wu, Chen, & Stone, 2018). Also, the findings suggest that using construct relevant strategies (i.e., strategies that help comprehension) or not using test-wiseness strategies (e.g., guessing by clues from other questions) explains the variation in test-takers' scores (Chen, Liu, & Zumbo, 2020; Wu & Zumbo, 2017). Together, these studies lend support to the score interpretation for CELPIP-General reading and listening tests from the perspective of response processes.

6.2. Research on Writing Quality of Repeat Test Takers

In a recent paper in the journal *Language Testing,* Lin and Chen (2020) examined the writing score and writing feature changes of 562 repeat CELPIP–General test takers who took the test at least three times, with a short (30–40 day) interval between the first and second attempts and a longer (90–180 day) interval between the first and third attempts.

Analysis was conducted to uncover whether changes occurred at different testing durations (short vs. long) and whether the observed changes varied across repeater's initial proficiency groups (low, mid, high).

- The writing scores measured by CELPIP bands showed great stability over the 6-month period, but the trends of development differed by proficiency group. Low proficiency test takers were more likely to have faster observable score gains, compared to the medium proficiency group, whereas high proficiency repeaters may not maintain their score levels at later attempts.
- Analysis of the writing features suggested that for all proficiency groups, lexical features were more likely to improve over the 6-month period, with some measures showing improvement at 1 month; features in cohesion and syntactic sophistication, however, did not change significantly.

Taken together, the results showed that proficiency influenced repeaters' writing score and fluency change between attempts, while for other linguistic features (lexical, cohesion, and syntactic), developmental patterns remained largely stable across proficiency groups over the 6-month period.

6.3. Validity Evidence from Mapping Language Use and Communication

Funded by the Paragon Research Grant program, Doe, Douglas, and Cheng (2019) examined the language use and communication challenges related by 23 Canadian new immigrants in entry-level workplace contexts. The key competencies associated with workplace communication were identified and mapped onto the Canadian Language Benchmarks (CLB) and the CELPIP-General LS levels of test performance.

- Communicative events were derived from thematic analysis of interview data where participants responded to questions about their language background and workplace language use related to specific events. These communicative events were assigned CLB levels and coded based on participants' perceived successes or challenges regarding their performance.
- These identified workplace interactions revealed that participants are typically required to perform tasks at CLB levels 4 to 6 in their entry-level positions, which correspond to CELPIP-General LS levels 4 to 6.
- Speaking was the predominant skill for the communication events labelled as successes, while communication events coded as challenges were balanced between speaking and listening.

This mapping provided actual indicators of language use and communication challenges in relation to what new Canadian immigrants working in entry-level-type workplace positions can do, and how well they do in reference against CLB and CELPIP-General LS. The findings can inform test design as the basis for measuring language proficiency within workplace contexts.

6.4. Advances in Psychometric Methods Used in Validation Studies and Operational Testing

Although contemporary validity theory expands the evidential basis beyond the conventionally reported studies of (i) internal structure of the test (dimensionality, reliability, and DIF) and (ii) relations to other variables (evidence of predictive, concurrent and discriminant validity), psychometric and statistical methods continue to play a role in validation practices (Chapelle, 2020; Kane, 2013; Plake & Wise, 2014; Zumbo & Padilla, 2020) as well as routine test-operations practices.

Paragon has an active program of psychometric research fostered, in part, by the Paragon-UBC Research Agreement (the first Agreement was from 2015 to 2020 and the second Agreement is from 2020 to 2024) and the Paragon UBC Professorship in Psychometrics & Measurement held by Professor Bruno Zumbo. Three research themes discussed below are representative of the on-going psychometric research in support of Paragon's test validation and operational testing procedures.

6.5. Recent Research on Test Fairness and Differential Item Functioning

Procedures to identify differential item functioning (DIF), and thus potential item bias, are frequently used in the process of developing and adapting language tests, as well as for the validation of test-score interpretation. In a widely cited article in the journal *Language Assessment Quarterly*, Zumbo introduced the concept of *Third Generation DIF* wherein the analysis of DIF is performed to examine five issues that are foundational for establishing test validity and supporting operational testing practices (Zumbo, 2007a). In Paragon Testing's context the five issues are listed below.

(a) Fairness and equity in testing for test participants from different groups.
(b) Adaptation of measures to different test-delivery methods (e.g., at-home versus test centre), languages, or cultures.
(c) Identifying group differences in item responding that—pending further investigation—arise from group differences that are either criterion-relevant or -irrelevant, such as differences in ability, differences in cognitive processing, and/or differences in contextual or psychosocial factors.
(d) As part of model checking for item response theory and other such latent variable modeling during operational testing.
(e) Ruling out measurement artifact as potential threat to internal validity for studies of language-policy interventions.

Recently, an important step forward in DIF methods for language assessment was made by Chen, Liu, and Zumbo (2020) when they introduced a novel DIF method based on propensity-score matching that allows researchers to investigate DIF for performance tasks such as writing or speaking with greater statistical efficiency. They demonstrated this propensity DIF method using the CELPIP-General writing tasks. Paragon is currently investigating the implementation of this novel method in its operational testing context as well as future validation studies.

A series of studies funded by the Paragon-UBC Research Agreement developed and tested novel statistical methods applying propensity-score matching in DIF analysis to determine the cause of DIF (Liu et al., 2016; Liu et al., 2019).

In 2015 in the journal *Language Assessment Quarterly*, a theme issue of the journal on advances in language assessment in Canada, Zumbo and his colleagues introduced a psychosocial ecological theory of item responding and a novel statistical method for DIF analysis to support this program of research (Zumbo et al., 2015).

6.6. Recent Research on Concurrent Calibration

Paragon uses concurrent calibration for test equating and scoring of the listening and reading components in its operational testing activities to place test takers onto a common IRT scale. Paragon conducted two recent studies to: (i) investigate the robustness of concurrent calibration methods to changes the location (mean) and scale (variance) of the test score distribution of CELPIP test takers, and (ii) determine minimum sample sizes for which concurrent calibration remains consistent. Both studies used a computer-simulation method developed within Paragon to mimic the test responding, equating,

and eventual scoring to result in a CELPIP level for computer-simulated test takers. The simulation method uses the mathematics of item response theory along with real operational item parameter values and operational cut scores to closely reflect the CELPIP test process and experimentally manipulate the study conditions. Early evidence suggests that concurrent calibration is robust and provides consistent CELPIP levels in the presence of a change of nearly half a standard deviation in the mean and a fourfold increase in the variance of the population of test scores over time— those values could be considered as reflecting a substantial change in the test-taker population. Using the same simulation methodology, minimum sample sizes were determined and are in use operationally to select items that warrant concurrent calibration.

7. Concluding Remarks

The Key findings and highlights of the report are summarized below:

- The CELPIP Test is designed to assess a wide ability range, which covers CLB levels 3 to 12.
- Paragon has continued to engage in test development and research in order to stay current to the assessment context, reflect current practices, and improve efficiencies. Paragon researchers continue to conduct assessment and psychometric research conducted in collaboration with external researchers to supports the validity of interpretations made by test users of the CELPIP Test.
- As highlighted by *Test Standards* (AERA, APA, & NCME, 2014), maintaining the highest test standards and practices is not an activity that stops once the assessments are developed, but rather is an ongoing process that is initiated at the beginning of assessment design and continues throughout development and implementation in the life-course of a testing system.
- Procedures are implemented to ensure high quality content and accurate and consistent scores, including: (a) a meticulous item writer training, monitoring, and feedback program, (b) a rigorous cycle of item review, editing, and field-testing, (c) careful rater certification, training, feedback, and monitoring procedures. In addition, regular validation checks have been implemented to ensure the accuracy of the psychometric analyses (results will be presented in CELPIP Test Review Report II).

References

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

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* New York: Oxford University Press.

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

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

Chapelle, C.A. (2020). *Argument-Based Validation in Testing and Assessment*. Thousand Oaks, CA: Sage Publishing.

Chen, M. Y., Liu, Y., & Zumbo, B.D. (2020). A Propensity Score Method for Investigating Differential Item Functioning in Performance Assessment. *Educational and Psychological Measurement, 80(3)*, 476–498.

Chen, M. Y., Wu, A. D., & Liu, Y. (2020). Linking test-taking process to performance through mixed-effects regression models: A response process–based validation study. *Journal of Psychoeducational Assessment, 38(3)*, 389-401.

Doe, C., Douglas, S., & Cheng, L. (2019). *Mapping language use and communication: Challenges to the Canadian Language Benchmarks and CELPIP-General LS within workplace contexts for Canadian new immigrants*. Paragon Testing Enterprises, Vancouver, Canada. https://www.paragontesting.ca/wp-content/uploads/2019/06/Final-Report_June-21.pdf

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

Lin, Y. & Chen, M. Y., (2020). Understanding writing quality change: A longitudinal study of repeaters of a high-stakes standardized English proficiency test. *Language Testing*, *37*(4), 523–549. https://doi.org/10.1177/0265532220925448

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., Kroc, E., & Zumbo, B. D. (2019). Investigating the performance of propensity score approaches for differential item functioning analysis. *Journal of Modern Applied Statistical Methods*. https://doi: 10.1177/0013164419878861

Liu, Y., Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. D. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research & Evaluation, 21*(13). Retrieved from http://pareonline.net/getvn.asp?v=21&n=13

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice, 33(4)*, 4–12.

Wu, A. D., Chen, M. Y., & Stone, J. E. (2018). Investigating how test-takers change their strategies to handle difficulty in taking a reading comprehension test: Implications for score validation. *International Journal of Testing, 18(3)*, 253–275.

Wu, A. D., & Zumbo, B. D. (2017). Understanding test-taking strategies for a reading comprehension test via latent variable regression with Pratt's importance measures. In B. D. Zumbo and A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 305–320). New York: Springer.

Zumbo, B. D. (2007a). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D. (2007b). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science.

Zumbo, B. D., & Chan, E. K. H, (Eds.) (2014). V*alidity and validation in social, behavioral, and health sciences*. New York: Springer.

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. New York: Springer.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B.R., Astivia, O. L. O. & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*, 136-151.

Zumbo, B. D., & Padilla, J. L. (2020). The interplay between survey research and psychometrics, with a focus on validity theory. In P.C. Beatty, D., Collins, L., Kaye, J.L. Padilla, G. Willis, and A. Wilmot, (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 593-612). Hoboken, NJ: Wiley.