

CELP[®]-G Research

Series I, Volume 1

February 2011

REPORT ON THE EVALUATION OF CELP[®]-G TEST SCORES



PARAGON TESTING ENTERPRISES

Contents

- Introduction..... 3**
 - CELPPIP-G Test Overview.....3
 - CELPPIP-G Test and CIC.....3
- Measurement Error..... 3**
 - Reliability for the Four Components of the CELPIP-G Test3
 - Measurement Error for the CELPIP-G Test.....4
- CIC Classification..... 5**
 - CIC Cut Scores.....5
 - Classification Consistency and Classification Accuracy5
- Validity 5**
 - Gender Bias.....5
 - Correlations Among the Four Components of the CELPIP-G Test.....6
 - Criterion Validity of the Writing Component of the CELPIP-G Test.....6
- Conclusion 6**

Introduction

The purpose of this report is to evaluate the appropriateness and usefulness of the CELPIP®-General (CELPIP®-G) Test to assess the English language proficiency of applicants for immigration to Canada. Paragon Testing Enterprises is dedicated to sharing the research on its English language testing program with users of its tests.

CELPIP-G Test Overview

The CELPIP-G Test is a complete English language testing program that assesses general levels of functional competency. The test consists of four components: the Listening component, which measures an individual's comprehension of spoken English in everyday situations; the Speaking component, which measures an individual's ability to communicate orally in English; the Reading component, which assesses an individual's general reading comprehension; and the Writing component, which assesses an individual's general composition skills.

CELPIP-G Test and CIC

The CELPIP-G Test has been approved by Citizenship and Immigration Canada (CIC) as a measure of English language proficiency for the Skilled Workers and Professionals Class for immigration to Canada. CIC awards immigration points based on CELPIP-G Test score levels as part of the application process for immigration.

Measurement Error

Reliability for the Four Components of the CELPIP-G Test

Reliability is an important quality of a trustworthy standardized test. Reliability refers to the extent to which a measurement produces the same results if used again in similar circumstances. A measure of internal consistency is Cronbach's reliability coefficient alpha (between 0 and 1), which measures how closely related the items of a test are. A high Cronbach's alpha provides evidence that the items measure a construct reliably.

As shown in Table 1, there is a good degree of internal consistency among the items of all four components of the CELPIP-G Test. The reliability of the Listening component is 0.93, which is slightly higher than that of TOEIC (0.92) and IELTS-General Training (0.91). The reliability of the Speaking component is 0.95, which would be considered, by all standards, as outstanding for a test that is 15 items in length. The reliability of the Reading component (0.86) is slightly lower than that of TOEIC (0.93) and IELTS-G (0.90) but meets the standards of the industry. The reliability of the Writing component (0.85) is mid-way in the range of 0.80 to 0.89 reported for the IELTS-G Writing Test and is deemed to have met the accepted industry standard. To sum up, the reliability of all four components of the CELPIP-G Test meets the standards of the industry and is comparable to that of the leading tests in the field¹.

¹ Comparisons between the reliability of the CELPIP-G Test and the reliability of other English language testing programs are not conclusive because of differences in test design.

Table 1: Comparison of the Reliability of the CELPIP-G Test with TOEIC and IELTS-G

Test	Listening Component		Speaking Component		Reading Component		Writing Component	
	Number of Items	Reliability (Alpha)	Number of Items	Reliability (Alpha)	Number of Items	Reliability (Alpha)	Number of Tasks	Reliability (Alpha)
CELP-IP-G	45	0.93	15	0.95	35	0.86	2	0.85
TOEIC	100	0.92	11		100	0.93	8	
IELTS-G	40	0.91			40	0.90	2	0.80-0.89

Measurement Error for the CELPIP-G Test

Measurement error estimates the difference between a recorded score and its true value. Measurement error often arises from test design and examinees’ transient states (e.g., mood, health, and stress levels). In practice, no test can be designed that provides a perfect reflection of examinees’ true scores. The greater the reliability of a test, the smaller the measurement error and the more precise the score.

Table 2 shows the relative measurement error of the four components of the CELPIP-G Test in relation to the score variation (technically, this is the ratio of the standard error of measurement to the standard deviation of scores). The relative measurement error of the Listening component is at a level (0.25) that is slightly lower than that of TOEIC (0.26) and IELTS-G (0.30). The relative measurement error of the Speaking component is considered small at 0.22; that is, only about one fifth of the score variation is due to measurement error. This is, in fact, smaller than that of the speaking components of TOEIC (0.25) and IELTS-G (0.26). The relative measurement error of the Reading component (0.38) is slightly higher than that of TOEIC (0.29) and IELTS-G (0.32) but is considered to be at an industry-appropriate level. To sum up, although the four components of the CELPIP-G Test have different levels of measurement error, their performances meet the standards of the industry and are comparable to those of the leading tests in the field².

Table 2: Comparison of the Relative Measurement Error (RME) of the CELPIP-G Test with TOEIC and IELTS-G as a Proportion of Score Variation

Test	Listening Component		Speaking Component		Reading Component		Writing Component	
	Number of Items	RME	Number of Items	RME	Number of Items	RME	Number of Tasks	RME
CELP-IP-G	45	0.25	15	0.22	35	0.38	2	0.39
TOEIC	100	0.26	11	0.25	100	0.29	8	
IELTS-G	40	0.30		0.26	40	0.32	2	

² Comparisons between the relative measurement error of the CELPIP-G Test and the relative measurement error of other English language testing programs are not conclusive because of differences in test design.

CIC Classification

CIC Cut Scores

Citizenship and Immigration Canada cut scores are the required minimum scores that map onto the CIC immigration points of 0, 1, 2, and 4 for English language proficiency. These cut scores correspond to the CELPIP-G Test levels 0 through 4 High as well as to the *minimal*, *developing*, *adequate*, and *effective* proficiency levels of the Canadian Language Benchmarks 2000.

Classification Consistency and Classification Accuracy

Classification consistency refers to the agreement between classifications of CIC immigration points based on the observed scores of two alternate forms having equal difficulty.

Classification accuracy refers to the classification agreement between the observed scores and the true scores expected over repeated administrations of test forms having equal difficulty.

The CELPIP-G Test cut scores have very high accuracy and consistency when used to assign immigration points to candidates. The overall consistency rates are 93.6%, 93.3%, 90.4%, and 88.4% for the Listening, Speaking, Reading, and Writing components, respectively. The overall accuracy rates are 95.4%, 95.2%, 93.2%, and 91.8% for the Listening, Speaking, Reading, and Writing components, respectively. This indicates that all four components of the CELPIP-G Test classify examinees into the same categories of CIC points over repeated test administrations and that the examinees' scores agree with the true classifications.

Validity

Validity refers to how grounded the test score interpretation is in relation to the intended use of a test. Paragon has an extensive research program for test score validation that will be carried out continually. The completed research on test score validation addresses gender bias detection, convergent validity among the four test components, and criterion validity in distinguishing between ESL and English first examinees.

Gender Bias

This research evaluates the extent to which the Listening, Reading, and Speaking components of the CELPIP-G Test exhibit *gender bias*³. For the purposes of this research, an item is considered to be biased if the probability of getting that item correct is different between male and female examinees having the same English language proficiency. The research shows that only 7.18%, 9.01%, and 3.83% of the items of the CELPIP-G Test exhibit moderate-to-large gender bias for the Listening, Reading, and Speaking components, respectively. As Table 3 shows, the items favor male examinees slightly more than female examinees. These figures are much lower than those reported in the literature for standardized language testing. It is seen in practice that standardized tests usually contain up to about 10% to 15% gender-biased items.

³ The research did not address gender bias for the Writing component because the method used to analyze gender bias for the items of these three components could not be used for the tasks of the Writing component.

Table 3: Percentage Gender-Biased Items and Gender Favored

Gender Favored	Listening Component	Reading Component	Speaking Component
Females	3.42%	3.74%	1.69%
Males	3.76%	5.27%	2.54%
All Examinees	7.18%	9.01%	3.83%

Correlations Among the Four Components of the CELPIP-G Test

An important indicator of the quality of a test is *convergent validity*, which refers to the degree to which scores on a test correlate with (or are related to) scores on other tests that measure the same construct. The research provides solid evidence concerning the convergent validity of the Writing component of the CELPIP-G Test.

The examinees’ CELPIP-G Writing scores correlate with the scores on the Listening, Reading, and Speaking components at 0.63, 0.72, and 0.70, respectively. These correlations have sizes deemed at an industry-appropriate level and are consistent with the correlations found in the language testing literature.

Criterion Validity of the Writing Component of the CELPIP-G Test

Another important indicator of the quality of a test is *criterion validity*, which analyzes a test against a known standard. For the Writing component of the CELPIP-G Test, the research analyzes the scores of ESL examinees (Chinese- and Spanish -first language examinees) against the scores of English-first language examinees to assess how well the CELPIP-G Test scores for this component distinguish between these two groups of examinees. The finding that the English-first language examinees (Mean=21.0) score much higher than the ESL examinees (Mean=15.75) provides solid evidence in support of the criterion validity of the CELPIP-G Test.

Conclusion

The research confirms the appropriateness and usefulness of the CELPIP-G Test scores to assess the English proficiency of individuals applying for permanent residence in Canada. The CELPIP-G Test has high quality of item performance and overall reliability, and the score distributions are appropriate for these applicants. To sum up, the Listening and Speaking components have comparable properties that are of great quality. The Reading component has good psychometric characteristics, while the Writing component has acceptable qualities. Although the four components of the CELPIP-G Test have different psychometric characteristics, their performances are comparable to those of the leading tests in the field.

Contact: info@paragontesting.ca



Cover image: Shutterstock

Copyright © 2011 Paragon Testing Enterprises, a subsidiary of The University of British Columbia. All rights reserved. CELPIP-G is registered trademark of Paragon Testing Enterprises. It is illegal to reproduce any portion of this material except by special arrangement with Paragon Testing Enterprises. Reproduction of this material without authorization, by any duplication process whatsoever, is a violation of copyright.

PARAGON TESTING ENTERPRISES
www.paragontesting.ca