CELPIP Annual Report of 2023 Test Takers

research@prometric.com

CELPIP® | by PROMETRIC

1 Description of the Test

1.1 General Description

The Canadian English Language Proficiency Index Program (the CELPIP Test) is a standardized test of general English language proficiency. It is administered by Paragon Testing Enterprises Inc, a Prometric company.

The purpose of the CELPIP Test is to provide a valid and reliable measurement of a test taker's English abilities in a variety of everyday situations, such as communicating with co-workers and superiors in the workplace, interacting with friends, understanding newscasts, and interpreting and responding to written materials. The CELPIP Test is designated by Immigration, Refugees and Citizenship Canada (IRCC) for permanent residency and citizenship of Canada. The CELPIP Test is also accepted by a number of post-secondary institutions and professional associations as proof of English language proficiency for academic and professional purposes.

Paragon is committed to upholding the highest standards in educational measurement. All parts of the CELPIP Test are written following specified guidelines, and results are closely monitored to ensure they are accurate, informative, and defensible. Paragon works closely with test centres to make certain that the CELPIP Test is administered fairly, securely, and is accessible to all individuals who wish to take the exam.

1.2 Test Format

There are two versions of the CELPIP Test: the CELPIP-General Test and the CELPIP-General LS Test. Individuals who take the CELPIP-General Test are assessed on on four components: Listening, Reading, Writing, Speaking. Individuals who take the CELPIP-General LS Test are assessed on the Listening and Speaking components. Table 1 describes the format and content of each test component.

Table 1: Format and Content of the CELPIP Test

Component	Duration (Minutes)	Item Description	Items*
Listening	*47 – 55*	Test takers listen to seven passages and answer comprehension questions. The listening passages cover topics in daily conversation, problem-solving, news items, discussions, and viewpoints.	38
Reading	*55 – 60*	Test takers read several passages and answer comprehension questions. The reading passages engage the test takers in understanding correspondence, interpreting a diagram, and reading for viewpoints.	38
Writing	*53*	Test takers write an email and write a response to survey questions.	2
Speaking	*15*	Test takers speak to give advice, talk about personal experiences, describe scenes, make predictions, compare and persuade, deal with difficult situations, express opinions, and to describe an unusual situation.	8

The Listening and Reading components may contain additional unscored items. These unscored items will have the same format as the scored items.

1.3 Scoring and Reporting of CELPIP Results

The CELPIP Test has been designed to assess the English language ability of test takers in general social, educational, and workplace contexts.

Test takers receive a score report that provides a score for each component. The multiple-choice items in the Listening and Reading components are scored by computer. Each correct answer contributes proportionately to the final score, and no points are deducted for wrong answers. The Writing and Speaking components are each evaluated by at least three Paragon-certified raters according to a scale established by Paragon.

CELPIP scores are reported in bands ranging from 0 to 12 . Test scores have been calibrated against the Canadian Language Benchmark (CLB) levels. Table 2 shows each CELPIP level and its corresponding description, with the CLB level equivalencies included in the third column.

Table 2: Interpretation of CELPIP Test Scores

CELPIP Level	CELPIP Descriptor	CLB Level
12	Advanced proficiency in workplace and community contexts	12
11	Advanced proficiency in workplace and community contexts	11
10	Highly effective proficiency in workplace and community contexts	10
9	Effective proficiency in workplace and community contexts	9
8	Good proficiency in workplace and community contexts	8
7	Adequate proficiency in workplace and community contexts	7
6	Developing proficiency in workplace and community contexts	6
5	Acquiring proficiency in workplace and community contexts	5
4	Adequate proficiency for daily life activities	4
3	Some proficiency in limited contexts of personal relevance	3
2	Limited ability in contexts related to immediate needs	1, 2
1	Insufficient information to assess	/
0	Insufficient information to assess	/
NA	Not Administered: test taker did not receive this test component	/

CELPIP levels 0, 1, and 2 were previously reported as level "M"

When interpreting a CELPIP Test score report, it is important to remember that the CELPIP Test estimates test takers' true proficiency by approximating the kinds of tasks they may encounter during their daily lives, study, or at work. There are, inevitably, small mismatches between the tasks that test takers complete as part of the test and the tasks that they perform in a specific context. Also, temporary factors unrelated to test takers' true proficiency, such as fatigue, anxiety, or illness, may affect their CELPIP test results.

2 Test-Taking Population

This section presents an overview of the test takers who took the CELPIP Test in 2023. Figure 1 presents the distributions of test takers' self-reported purpose for taking the test. The most commonly reported purpose for taking the CELPIP Test is for Canadian immigration and citizenship applications (IRCC = 95.5%).

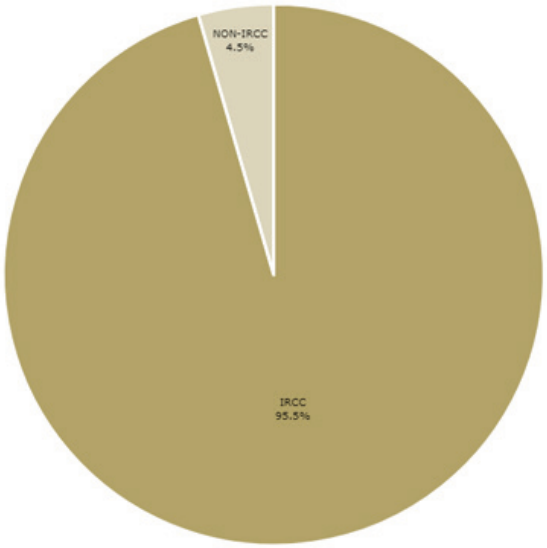

Figure 1: Distribution of CELPIP Test Takers by Test-Taking Purpose

Figure 2 shows the distribution of test takers for the CELPIP-General and the CELPIP-General LS Tests in the total test-taker population.

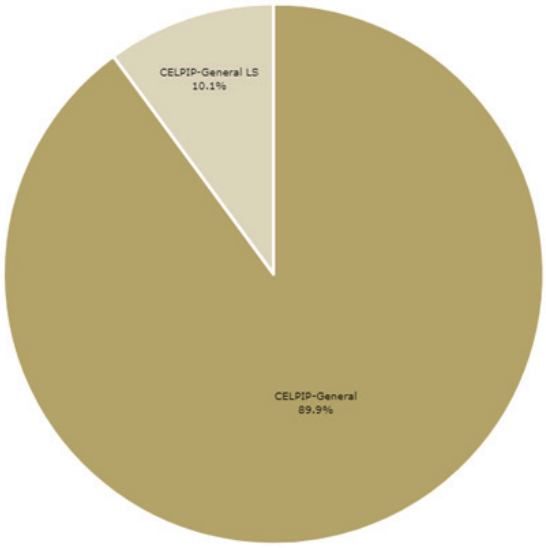

Figure 2: Distribution (in %) of CELPIP Test Takers by Test

Table 3 shows the age group distributions of test-taker populations by CELPIP test type. The majority of CELPIP test takers were between 20 and 40 years old (80.7%). A comparison of age distributions between the CELPIP-General and CELPIP-General LS populations reveals that CELPIP-General has a younger demographic, with larger groups between ages 20 and 30. In contrast, the CELPIP-General LS test-taker population has a slightly greater distribution between ages 36 and 55.

Table 3: Distribution of CELPIP Test Takers by Age

Age Group	% Test Population		
	Overall	CELPIP-General	CELPIP-General LS
< 20	0.3	0.3	0.3
20-25	13.7	14.8	4.3
26-30	29.1	30.9	13.7
31-35	21.6	21.5	22.6
36-40	16.3	15.4	24.2
41-45	10.6	9.7	18.5
46-50	5.4	4.7	11.7
51-55	2.2	2.0	4.7
> 55	0.8	0.8	0.1

3 Test Statistics

3.1 Score Distributions

Table 4 presents the mean and standard deviation for the CELPIP-General and CELPIP-General LS component scores. The mean score is the simple average of all test takers' scores, and the standard deviation quantifies how scores are spread out from the mean.

Table 4: CELPIP-General and CELPIP-General LS Test Scores

Test	Component	Mean	Standard Deviation
CELPIP-General	Listening	8.29	2.34
	Reading	7.52	2.55
	Writing	7.99	1.96
	Speaking	7.85	2.10
CELPIP-LS	Listening	6.48	2.56
	Speaking	6.60	2.35

Tables 5a and 5b show the percentage breakdown of test takers who received each component band score on the CELPIP-General and CELPIP-General LS Tests.

Table 5a: Score Distributions (%) on CELPIP-General Test

Band	Reading	Writing	Speaking	Listening
0-2	1.8	0.3	0.3	0.4
3	3.3	0.8	0.9	1.5
4	7.1	2.0	3.1	4.3
5	11.9	5.2	7.2	8.1
6	14.4	10.9	16.1	10.6
7	13.1	22.8	16.1	11.7
8	10.1	27.0	23.1	12.6
9	14.0	9.7	12.5	15.5
10	8.4	8.8	6.4	16.7
11	9.3	5.2	7.3	10.3
12	6.6	7.4	6.8	8.3

Table 5b: Score Distributions (%) on CELPIP-General LS Test

Band	Listening	Speaking
0-2	3.3	3.2
3	7.7	5.3
4	14.3	9.5
5	16.4	15.1
6	13.8	19.7
7	10.9	13.4
8	9.0	14.4
9	8.5	7.1
10	8.2	4.5
11	4.8	4.5
12	3.1	3.2

The above tables (Tables 4, 5a, and 5b) suggest that a larger proportion of test takers achieve lower bands on the CELPIP-General LS Test than on the CELPIP-General Test. It is important to note that, generally, individuals take the CELPIP-General and the CELPIP-General LS to meet different immigration requirements. The CELPIP-General Test is typically taken by the primary applicant for Canadian Permanent Residency. The CELPIP-General LS is required for applications for Canadian Citizenship. The minimum language proficiency requirements are currently different for permanent residency and citizenship applications. The different requirements may be a primary contributor to the observed difference in score profiles between the CELPIP-General and the CELPIP-General LS Tests.

Appendices A - E offer a more detailed breakdown of test taker performance by gender, first language, and country of citizenship.

3.2 Measurement Consistency

In statistics and psychometrics, measurement consistency is often referred to as reliability. It is an important component when evaluating the quality of a test. A reliable test gives us the same result consistently, assuming no change in the individual's ability. For example, if a test is designed to measure English language proficiency and has been administered to the same individual multiple times, the test scores should be approximately the same if the test taker has not significantly improved their English proficiency during the period of time. In contrast, an unreliable test produces inconsistent results each time, which greatly limits the value of the test scores.

Reliability can be estimated in multiple ways. In general, a higher value suggests a greater reliability of the test scores. A reliability of 0.80 and above is recommended according to the literature and industry standards (e.g., George & Mallery, 2016). In this section, we report internal consistency for CELPIP Listening and Reading components and rater agreement for CELPIP Speaking and Writing components.

3.2.1 Internal Consistency for the Listening and Reading Components

Internal consistency is a measure of whether test items designed to measure the same construct produce similar results. It is suitable for quantifying the reliability of tests that consist of many items, such as the Listening and Reading components on the CELPIP Test. To measure internal consistency, Cronbach's alpha was calculated for each CELPIP Listening and Reading form administered in 2023 (see Table 6).

Table 6: Reliability Estimates for the Listening and Reading forms on the CELPIP Test

Component	Mean Reliability	Standard Deviation	Mean Distances from the Median
Listening	0.86	0.03	0.02
Reading	0.88	0.03	0.02

Table 6 shows that the mean internal consistency for the Listening and Reading forms was 0.86 and 0.88, respectively. These values suggest that, on average, there was high internal consistency within each of the CELPIP Listening and Reading forms. Additionally, the standard deviations of the mean reliability (0.03 for Listening and 0.03 for Reading) and the mean distances from the median were small (0.02 for Listening and 0.02 for Reading), indicating the forms performed similarly in terms of their internal consistency.

3.2.2 Rater Agreement for the Writing and Speaking Components

The reliability of the Writing and Speaking scores of the CELPIP Test is maintained operationally by assigning multiple raters to independently assess a test taker's performance and regular monitoring of rater agreement. The raters for the Writing and Speaking components of the CELPIP Test are highly proficient in English and are fully trained and certified by Paragon. Each writing task is rated independently by two accredited raters. If the scores awarded by the raters disagree, the task is evaluated by a third rater. Each test-taker's speaking performance is rated independently by three accredited raters. Two additional raters will evaluate the responses if the scores awarded by the original raters disagree.

Since the evaluation process for the Writing and Speaking components relies on human judgment and the interpretation and application of a rating scale, variations in judgments are to be expected. Paragon constantly monitors rater agreement for quality control purposes. Table 7 shows the rater agreement for the Writing and Speaking components in 2023. Overall, the results indicate great consistency of judgment among raters.

Table 7: CELPIP Rater Agreement for Writing and Speaking (%)

	Speaking	Writing
Rater Agreement	85.3%	85.1%

4 Closing Remarks

The CELPIP Test continues to be an industry leader in language assessment. The test is fully computer-delivered and available at over 140 locations across Canada and internationally, with test dates available every week. Paragon offers a range of test preparation materials, both for free and for purchase, to help you prepare for your test. Our CELPIP experts work in collaboration with official CELPIP Test Centers to provide the CELPIP Preparation Program to help test takers build the skills for each component of the CELPIP Test. For more information about test registration and preparation, visit the CELPIP website at https://www.celpip.ca/.

Reference

George, D., & Mallery, P. (2016). IBM SPSS statistics 23 step by step: A simple guide and reference. Routledge.

Appendices

The following appendices are provided as an overview of the outcomes of test-takers based on different demographics.

Appendix A: CELPIP Scores by Test Purpose

Test Purpose	Listening		Reading		Writing		Speaking	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
IRCC	8.07	2.44	7.50	2.56	7.96	1.96	7.67	2.14
Non-IRCC	8.83	2.06	7.93	2.40	8.48	1.91	8.93	2.27

Appendix B: CELPIP-General Scores by the Top 10 Declared Nationality

Country	Listening		Reading		Writing		Speaking	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
India	8.35	2.05	7.40	2.27	7.89	1.62	7.81	1.65
Philippines	7.20	2.30	6.25	2.43	7.33	1.92	6.85	1.83
China	7.88	2.63	7.53	2.70	7.62	1.85	6.94	1.86
Hong Kong, China	8.26	2.25	7.91	2.42	7.98	1.64	7.17	1.67
Nigeria	8.65	1.99	7.78	2.28	8.88	1.80	9.40	1.66
Ukraine	7.36	2.51	6.55	2.59	7.05	1.91	6.93	1.91
Brazil	8.89	2.30	8.43	2.49	8.16	1.83	7.94	1.88
Canada	9.33	1.87	8.43	2.34	9.02	1.90	9.82	2.13
Iran	8.47	2.23	7.59	2.40	8.00	1.74	7.73	1.71
Mexico	8.40	2.43	7.75	2.62	7.70	2.00	7.79	1.98

Appendix C: CELPIP-General Scores by the Top 10 Declared First Languages

Language	Listening		Reading		Writing		Speaking	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
English	9.18	2.16	8.43	2.53	9.16	2.08	9.62	2.13
Chinese	7.95	2.57	7.55	2.66	7.66	1.83	6.99	1.84
Panjabi	7.89	2.07	6.75	2.17	7.40	1.47	7.47	1.63
Tagalog	7.32	2.31	6.39	2.46	7.42	1.91	6.91	1.82
Hindi	8.65	2.02	7.73	2.29	8.17	1.67	8.18	1.65
Spanish	8.56	2.39	8.00	2.56	7.95	1.97	7.86	1.98
Cantonese (Chinese)	8.33	2.21	7.99	2.37	8.00	1.61	7.20	1.63
Gujarati	8.29	2.22	7.31	2.34	7.64	1.71	7.48	1.76
Ukrainian	7.34	2.50	6.52	2.57	7.03	1.9	6.92	1.90
Portuguese	8.88	2.29	8.41	2.49	8.16	1.85	7.97	1.90

Appendix D: CELPIP-LS Scores by the Top 10 Declared Nations/Regions

Country	Listening		Speaking	
	Mean	SD	Mean	SD
India	6.26	2.53	6.4	2.25
Iran	5.98	2.34	6.16	1.90
Brazil	8.45	2.19	7.91	1.86
Pakistan	5.73	2.04	6.21	2.09
Syrian Arab Republic (Syria)	5.75	2.45	6.01	2.20
Mexico	7.47	2.41	7.44	2.15
China	6.70	2.40	6.14	1.71
Philippines	6.83	2.11	6.74	1.87
Eritrea	4.48	1.71	4.54	1.67
Afghanistan	4.92	1.77	5.63	1.91

Appendix E: CELPIP-General LS Scores by the Top 10 Declared First Languages

Language	Listening		Speaking	
	Mean	SD	Mean	SD
Arabic	5.96	2.42	6.22	2.23
Spanish	7.30	2.46	7.20	2.16
Farsi	5.74	2.25	6.02	1.92
English	7.36	2.62	8.26	2.65
Portuguese	8.32	2.23	7.86	1.90
Chinese	6.73	2.36	6.17	1.67
Urdu	5.77	2.08	6.24	2.07
Panjabi	5.37	2.25	5.57	2.07
Russian	7.06	2.59	6.94	2.22
Turkish	6.03	2.36	5.88	2.13