RTCC Language-competency Assessment

Phase II: Results of the Pilot-Testing Process

FINAL REPORT

directions
Evidence + Policy Research Group

cmec
Council of Ministers of Education, Canada
Conseil des ministres de l'Éducation (Canada)

Founded in 1967, the Council of Ministers of Education, Canada (CMEC) is the collective voice of Canada's ministers of education. It provides leadership in education at the pan-Canadian and international levels and contributes to the exercise of the exclusive jurisdiction of provinces and territories over education.

•

The Registrars for Teacher Certification Canada (RTCC) is a committee established in 1999, at the request of CMEC, to exchange information concerning the regulation of the teaching profession throughout Canada. Registrars also coordinate the implementation of the Canadian Free Trade Agreement (CFTA) for the teaching profession. The committee is composed of the registrars for teacher certification in all provinces and territories.

•

The Language Competencies Working Group (LCWG) is a time-limited working group tasked with supporting the Pan-Canadian Assessment Centre (PAC) and Integration of Internationally Educated Teachers (IETs) project, under the leadership of the RTCC.

Authors

This report was authored by *Directions* Evidence and Policy Research Group, LLP. It was funded as part of the PAC and IETs project under the Foreign Credential Recognition Program (FCRP) of Employment and Social Development Canada (ESDC) and the Registrars for Teacher Certification Canada (RTCC).

Disclaimer

The opinions, interpretations, findings, and recommendations expressed in this report are those of the authors. They do not necessarily reflect the official policy, positions, or views of the Council of Ministers of Education, Canada (CMEC), provincial and territorial governments in Canada, or provincial and territorial regulatory bodies for the teaching profession in Canada.

Acknowledgements

The analysis captured in this report has been supported by participation from the following:

- Registrars for Teacher Certification Canada (RTCC)
- Language Competencies Working Group (LCWG)
- Wired Solutions
- Eunice Eunhee Jang, PhD

Ce document est également disponible en français sous le titre :

Évaluation des compétences linguistiques des RAPEC - Phase II : Résultats de la mise à l'essai

Contents

Executive Summary

Internationally educated teachers (IETs) seeking certification in Canada must demonstrate their ability to communicate in one of Canada's official languages if their teacher education was not in French or English. The provinces and territories, which have jurisdiction over education, do not have methods to determine language proficiency specific to the demands of teaching. The Registrars of Teacher Certification Canada (RTCC, under the auspices of the Council of Ministers of Education, Canada — CMEC) sought an English or French language-proficiency assessment that would be common to all provinces and territories and that would help to meet their obligations for labour mobility in Canada. A review of existing assessments showed that none were sufficiently focused on the language competencies required for teaching. Therefore, the RTCC developed a teaching-specific assessment.

The RTCC language-competency assessment project has three phases: Phase I (2010–2013), included a literature review of the language competencies required for teaching, developing a framework of language competencies in the teaching profession, and developing English and French language-competency assessments to test those competencies. Phase II (2019–2022) included pilot testing the assessments to ensure they were psychometrically defensible and developing an implementation model for the test. Phase III will implement the test.

Directions Evidence and Policy Research Group was engaged by the Corporation of CMEC (CCMEC) as the lead consultant for Phase II pilot testing. Its main activities were administering the English and French tests to a pilot-test population, coding test responses (by experienced teachers in Canada), conducting psychometric analyses to ensure the assessments were reliable and unbiased, and setting the standard that an IET must meet to receive certification. Standards were set by a registrar or by an individual nominated by the registrar who was experienced in teaching and familiar with standardized testing.

Using a pragmatic framework applicable to licensure examinations, *Directions* conducted a study on the pilot-test data to ensure the RTCC language-competency assessments measure appropriate language skills, that measurements were reliable and unbiased, and that standards were appropriately set.

Directions strongly recommends the adoption of the RTCC language-competency assessments (French and English) because thoughtfully using the assessments provides a fair and defensible method of determining which IETs meet the standards set by the RTCC for the language competencies necessary for teaching in Canada. After the assessment's adoption, refinements to the test should continue to be guided by the RTCC language-competency framework. It's also important to create a coding guide, and test data will need to be collected and analyzed on an ongoing basis.

Background

The teaching profession is the largest regulated profession in Canada and one of the 14 occupations identified by the Forum of Labour Market Ministers in "A Pan-Canadian Framework for the Assessment Recognition of Foreign Qualifications."[1] The Registrars of Teacher Certification, Canada (RTCC)[2] has made significant progress in improving the fairness, transparency, consistency, and timeliness of the profession's assessment and recognition procedures.

The RTCC receives over 5,000 applications for certification each year from internationally educated teachers (IETs) who face significant challenges entering and moving within the Canadian labour market. A persistent barrier is the assessment of language proficiency. Almost half of Canada's provinces and territories have no language requirement for teacher certification, and most have no consistent French test to assess the language skills of prospective teachers in francophone settings. Registrars who require demonstration of language proficiency for certification rely on several different language-proficiency tests (e.g., IELTS, TEF Canada, CELPIP) that were not designed for the teaching profession and do not specifically address the competencies demanded by the profession.

The Council of Ministers of Education, Canada (CMEC), under the leadership of the RTCC, has been pursuing the development of a language-competency assessment for prospective IETs that would be taken as part of an initial assessment of their academic and professional qualifications.[3] The language-competency assessment specific to the teaching profession would be used to assess the language skills of IETs who have not completed an acceptable teacher-education program in English or French. The test is intended to ensure that candidates have the language competencies required to teach in English-first-language and French-first-language majority and minority contexts.

The RTCC language-competency assessment project has three phases:

- Phase I (2010–2013)
  - Literature review of language competencies required for teaching practice
  - Interjurisdictional scan to examine whether a language-proficiency assessment was available that is suitable for internationally prepared professionals seeking teacher certification in Canada
  - Development of a framework of language competencies in the teaching profession
  - Development of English and French language-competency assessment items and test versions to test those competencies
- Phase II (2019–2022)
  - Pilot testing the assessments to ensure they are reliable and valid and that the decisions based upon them are defensible
  - Developing an implementation model for the test
- Phase III (2023)
  - Implementing the RTCC language-competency assessment.

This report focuses on the Phase II pilot testing conducted by *Directions* Evidence and Policy Research Group (*Directions*), the lead consultant for this phase.

[1] "A Pan-Canadian Framework for the Assessment and Recognition of Foreign Qualifications," Forum of Labour Market Ministers, 2009. https://www.canada.ca/content/dam/esdc-edsc/documents/programs/foreign-credential-recognition/CA-561-11-09-EN.pdf

[2] Established in 1999 at the request of CMEC, the Registrars for Teacher Certification Canada is a committee whose aim is to exchange information on the regulation of the teaching profession in Canada. Registrars also coordinate the implementation of the Canadian Free Trade Agreement (CFTA) for the teaching profession. The committee is composed of the registrars for teacher certification in all provinces and territories.

[3] The RTCC refers to an assessment of language competency in the context of teaching. In psychometric terminology, a test is broadly defined and can include assessment tasks that do not look like traditional pencil-and-paper tests. The terms *test* and *assessment* are used interchangeably in this report but should be understood to refer to the assessment of language proficiency within the RTCC mandate.

Phase I: Development of the RTCC Language-competency Assessment

Literature review and interjurisdictional scan

The first step toward developing an assessment was to review the research literature to discover what language competencies K-to-12 teachers in English-as-a-first-language and French-as-a-first-language schools in Canada require for effective professional practice. *Directions* examined research from language teaching, language learning, and teaching competencies (761 papers in English, 394 papers in French). The review indicated that teachers require a broad and diverse set of language competencies to be successful in their professional practice.[4] Teachers in English or French as first-language contexts require the same set of competencies, but the realities of different linguistic contexts (e.g., majority-language, minority-language) can place different demands on teachers' language skills and knowledge.

A review of existing assessments for language competency revealed several limitations and no assessment examined the language competencies in reading, writing, speaking, and listening as they apply to teaching in Canadian K-to-12 anglophone and francophone contexts. Occupation-based general second-language assessments did not assess language competencies as they apply to teaching in the domains of instructing and assessing, managing the classroom and student behaviour, and communicating with professionals and parents. Teaching-specific language assessments did not assess all four of the language modalities and did not assess all the language domains specific to teaching (instructing and assessing, managing the classroom and student behaviour, communicating with professionals and parents). They were also designed for teaching in contexts outside of Canada, and/or had unclear reliability and validity information.

Framework for language competencies

The literature review informed the development of the Framework for Language Competencies and Benchmarks for teachers.[5] The language competencies are a set of statements describing linguistic abilities in English or French in each of four modalities: speaking, listening, reading, and writing. These modalities are commonly represented in language-proficiency frameworks. For each competency, the framework specifies performance outcomes for three domains of practice: (A) instructing and assessing, (B) managing the classroom and student behaviour, and (C) communicating with parents and other professionals.

For example, the first competency in the writing modality is: *Write coherent formal and informal texts by synthesizing and evaluating complex information and ideas from multiple sources*. The associated performance outcomes in each domain of practice include:

- **Instructing and assessing** — Write lesson plans, course outlines, course descriptions, handouts, and/or teaching materials.
- **Managing the classroom and student behaviour** — Write summaries of classroom expectations and goals.
- **Communicating with parents and professionals** — Write emails, letters, or reports to other professionals using technical or nontechnical language.

[4] "Speaking for Excellence: Language Competencies for Effective Teaching Practice," Council of Ministers of Education, Canada, 2013. https://www.cmec.ca/Publications/Lists/Publications/Attachments/320/Speaking_for_Excellence.pdf

[5] Ibid.

Test item development and test construction

In 2012, *Directions* developed approximately 1,600 French and 1,600 English assessment items based upon the language-competency framework. For instance, items in the writing modality that address the instructing and assessing domain included writing lesson plans, course outlines, and other teaching materials. Items were independently developed in French and English by a range of educators (e.g., current teachers, former teachers still working in education, and language-assessment experts). From this pool of items, *Directions* developed five French and five English test versions that were reviewed by external experts (internationally recognized experts in educational assessment, teacher professional development, language in education, and curriculum development) in French and English to confirm authenticity, face validity, and content validity and to provide suggestions for improving items. Incorporating feedback from both the external reviewers and from the RTCC Language Competency Subcommittee, the team revised or replaced test items.

In 2013, one English and one French version of the assessment was tested online in an alpha-testing process to ensure interface usability and that assessment demands were clear to the intended individual test takers.[6] The interface/platform and test versions were further revised based on the alpha tests, completing Phase I of the project.

[6] Alpha test takers were 38 volunteers, most of whom were internationally educated teachers.

Phase II: Pilot Testing the Language-competency Assessment

Goals

Directions was engaged as the lead consultant to pilot test versions of the French and English assessments to lead a pan-Canadian process to validate the reliability of the RTCC On-line Language Assessment for the teaching profession in Canada. The original Phase II plan included establishment of regional pilot-testing centres; recruitment by the Language Competencies Working Group (LCWG) of up to 2,400 test takers (1,200 each in English and French); preparation of up to three English- and three French-language assessments using test items developed in Phase I; coordinating in-person training and coding; conducting psychometric analyses on up to 120,000 pilot-test item responses (60,000 each in English and French); revising items based on findings; and facilitating the setting of minimum standards of language competencies for certification purposes at a pan-Canadian level.

Planning and implementation: Impacts of the COVID-19 pandemic

Planning for in-person pilot-test administration and coding sessions to take place in summer 2020 was underway when the global COVID-19 pandemic was declared in March 2020. Table 1 summarizes the impacts of the pandemic on the project and its timelines.

Table 1. Impacts of the COVID-19 pandemic on phase II pilot testing and timelines

Planned	Actual	Implementation Dates
Test administration: Four regional testing centres (BC, the Prairies, Ontario, Atlantic provinces) using XpressLab testing platform	At-home test administration across Canada using XpressLab testing platform and Proctortrack online invigilation services	• January – March 2021 • September – December 2021
Test-taker recruitment: Recruitment from all provinces and territories of test takers new to the teaching profession (teacher training in English or French) and internationally educated teaching professionals (teacher training not in English or French)[7] through targeted recruitment of: • teacher candidates in final stages of teacher education program • recent graduates of teacher education program • internationally educated teachers, including those applying for certification • certified teachers in first five years of teaching One round of recruitment.	Recruitment from all provinces and territories of test takers in originally planned categories. In early 2021, low registration numbers and recruitment challenges at the provincial and territorial level necessitated the expansion of potential test takers to include recruitment of any individuals: • preparing to teach • currently teaching • retired from teaching There were insufficient test takers (70 French, 143 English) between January and March 2021 so a second round of recruitment by RTCC took place from August to December 2021.	• November 2020 – March 2021 • August – December 2021

[7] The beginning English/French teacher data provide insight into standard setting so that a test for IETs is not setting a higher standard than for candidates prepared in English or French.

Table 1. Impacts of the COVID-19 pandemic on phase II pilot testing and timelines (cont'd)

Planned	Actual	Implementation Dates
Test-taker numbers: To reduce measurement error to acceptable levels for a high-stakes test, CMEC intended to recruit a minimum of 400 test takers for each test version (English Versions 2, 3, 4; French Versions 2, 3, 4) for a total of 1,200 English test takers and 1,200 French test takers.	Because of low registrations in early recruiting, the decision was made to test only two versions of the English and two versions of the French language-competency assessment. This would yield more test takers distributed over fewer versions. Final test-taker numbers were: • 589 test takers for English Versions 2 & 3 • 349 test takers for French Versions 2 & 3	• January – March 2021 • September – December 2021
Coding of writing and speaking items: In-person coder training and coding using the coding interface on the XpressLab testing platform	Virtual coder training and coding, two sessions involving 91 coders, using the coding interface on the XpressLab testing platform and the Slack online communication tool	• March – April 2021 • December 2021
Standard setting: In-person meetings	Virtual meetings (5 standard setters for English tests, 5 standard setters for French tests)	• February – March 2022

Test structure

Test versions developed in 2012 and revised after alpha testing in 2013 were reviewed for any required changes. Item-level scoring criteria for speaking and writing items were revised for Phase II to simplify the task of scoring the items. To test a larger pool of items, the *Directions* team incorporated additional items from Versions 1 and 5 of the tests that were developed in Phase I into the modules for Versions 2, 3, and 4. Each test version was also reviewed to ensure good coverage of the performance outcomes from the Framework for Language Competencies and Benchmarks.

For Phase II piloting, four versions of the test were ultimately used (English Version 2 [V2] and Version 3 [V3], French Version 2 [V2] and Version 3 [V3]). For both English and French versions, there is a set of items in each modality that are common across both versions (e.g., for listening, five items are the same in V2 and V3).[8] The remaining items are unique.

Test questions

Each test item was developed with reference to a performance outcome from the Framework for Language Competencies and Benchmarks. Each test item includes a stimulus (e.g., reading passage, scenario, or sound clip) and one or more questions or tasks for the test taker. Test items include selected-response (multiple-choice) questions, cloze (fill-in-the-blank) questions, and constructed-response questions (write text or record audio in response to instructions).

Sample items

What follows are sample test items from each modality. The performance outcomes for each item are provided for reference but are not shown to test takers.

[8] During Phase I test construction, an equal number of strong items in English and French was identified for each modality. These became the common items across English and French test versions. The French common items were translated into English, and English common items were translated into French. Subsequent revisions to the English common items were independent from revisions to the French common items, so items should be considered common only within a language rather than between languages.

Reading

Performance outcome: Read and evaluate a variety of primary and secondary subject-specific sources that use a range of visual, tabular, and textual information to gain subject expertise and to select materials for classroom study.

Stimulus

You have an upcoming unit in wood carving in your intermediate art class. As part of your preparation, you consult the following excerpt from a safety manual for carving so that you can understand instructions for the safe use of the equipment.

> *If the wood has been cut into workable pieces for the students, then you will need the following equipment for carving: steel files and wet/dry silicone sanding paper of different grades. All students need to wear a dust mask and goggles. When using files, students should wear gloves. A steel brush used by the teacher will keep the files clean and easier to use. The carving area must be well ventilated, and all dust must be cleaned up with a vacuum or wet mop, not a broom. Teach students to always "carve away from you," that is, to carve away from their body.*

Question

Given the manual excerpt, which statement is **NOT** a recommendation for ensuring that students safely carve wood?

a. Teach students the "carve away from you" rule.
b. Teach students to wear dust masks, goggles, and gloves when carving.
c. Teach students to vacuum or wet mop after they have swept the carving area.
d. Teach students to carve in a well-ventilated area.

Writing

Performance outcome: Provide written feedback on student assignments.

Stimulus

You are providing written feedback on assignments in which music students were asked to spend time at a sound-rich site and to sample and record as many sounds as they could. When you introduced the assignment, the examples given were a subway station, a shopping mall, a parade, and an elementary playground. Students were to take all the sounds and intersperse or overlay them with traditional instruments to form a composition that was to be recorded in an audio file and submitted to you.

Peter submitted a very interesting collection of sounds from the public wharf on a local waterfront. He demonstrated initiative in finding the site and collecting the sounds, and the quality of the recording of those sounds was good. However, he did not combine the sounds with traditional instruments, so the assignment was incomplete.

Instructions

Write two to three sentences to Peter giving him feedback on his work and explaining why he needs to complete and resubmit this assignment.

Suggested length: 90 words
Suggested writing time: 5 minutes

Coding criteria: (9 points)

1. Text is coherent and satisfies the task requirements. (3 points)
2. Text employs language appropriate for the audience. (3 points)
3. Text employs error-free spelling, punctuation, and grammar. (3 points)

Listening

Performance outcome: Listen to, interpret, and assess student answers to open-ended questions expressed in their own words.

Stimulus

You are assessing whether your students understand "reusing" and "recycling" following a presentation from a local environment organization, The Green Team.

> Green Team's presentation: *[linked audio: Many resources go into making things. Look at this box that a toy came in. Making the toy required resources like wood and paint and metal. But making the box requires resources too — wood pulp from trees is used to make the cardboard. This box might just be thrown away into the garbage. But it could also be re-used — as a box to store something, like winter clothes that won't be needed during the summer. The box could also be recycled or turned into something else altogether — the box can be chopped up and made into pulp again that can be made into new boxes.]*

To assess understanding, you ask your students, "What does it mean to reuse and recycle?"

Instructions

Listen to the responses of four students and indicate which student has the **MOST** complete understanding of the concepts *reuse* and *recycle*.

a. Lindsay *[linked audio: We shouldn't throw things away because it is a waste. My mom says "Waste not, want not"; reusing and recycling mean that we should hang onto things for longer.]*
b. Scott *[linked audio: Reusing is finding more things you can do with something instead of just putting it in the garbage. We don't have many trees so we need to not waste them. I have a recycle bin at my house and it's my job to put tins and bottles in it. I think someone crushes them and makes new things.]*
c. Nicole *[linked audio: Recycling and reusing are the same thing, to use it again. I put stuff in our blue recycling bin and the city takes it away and uses it again. At my Saturday art class I once reused newspaper by making new paper out of it.]*
d. Hugo *[linked audio: Recycling is to find new uses for old things, like when my mom uses my old pyjamas to wash the car or uses empty yogurt tubs to store cookies. My mom recycles everything! She says soon that if we don't do our part, we will have too much garbage.]*

Speaking

Performance outcome: Discuss student academic progress, social concerns, and other school-related issues with parents.

Stimulus

You will introduce field trip guidelines to parent volunteers who are helping you chaperone a field trip to a nature park with your class next month. The trip will be on a Saturday. You will travel by charter bus to the park and spend a day hiking and taking part in a variety of educational activities. The bus will pick up students at the school site at 7 a.m.

Three parents have volunteered. They will be coming to your classroom tomorrow to discuss the trip. There are several guidelines and rules that you want to highlight in your conversation:

1. All volunteers need to complete a background check. You have the forms that they will need to complete.
2. Each volunteer will be assigned a certain number of students. They need to look after all students, not just their own child.
3. Each volunteer needs to be aware of possible safety concerns. For example, while at the park, students need to follow instructions of the park staff, always be in groups with an adult, and stay in the areas identified by the staff.

Instructions

Record an introduction to your conversation with parent volunteers. Provide information about the field trip and summarize the three key rules and procedures that volunteers need to be aware of.

Suggested preparation: 2 minutes
Suggested speaking: 2 minutes

Coding criteria: (9 points)

1. Speech is coherent, satisfies the task requirements, and is appropriate to the intended audience. (3 points)
2. Speech is intelligible with clear pronunciation. (3 points)
3. Speech is fluent (uses stress, articulation, rate, and tone of voice appropriate for the task). (3 points)

Test interface

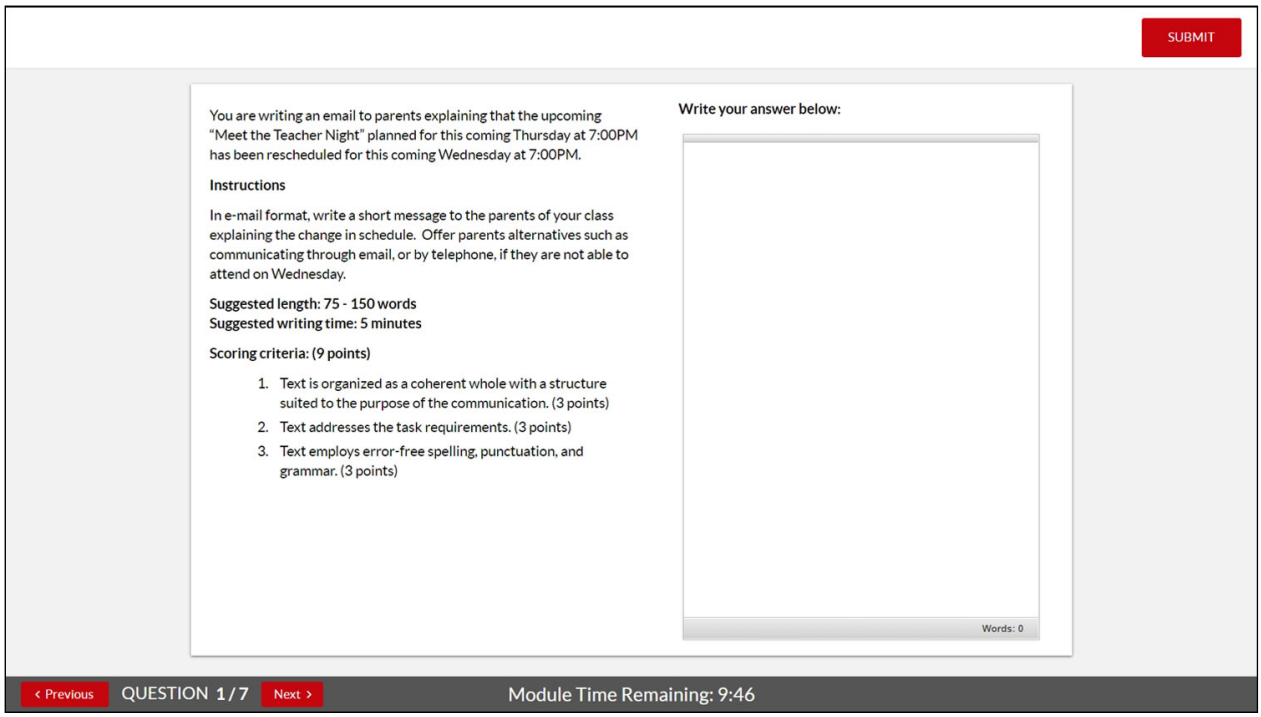
Test takers took the assessment using XpressLab, a customizable, cloud-based, language-testing software.[9] Proctortrack by Verificient, a remote invigilation service, live proctored the tests. The assessment has four hour-long modules (one hour each for reading, writing, speaking, and listening) with a maximum 15-minute break between modules. Test takers could initiate modules in any order that they chose.

[9] Wired Solutions, the IT firm that provided the XpressLab software for Phase I, was engaged for Phase II piloting.

The following question formats were used in the assessment:

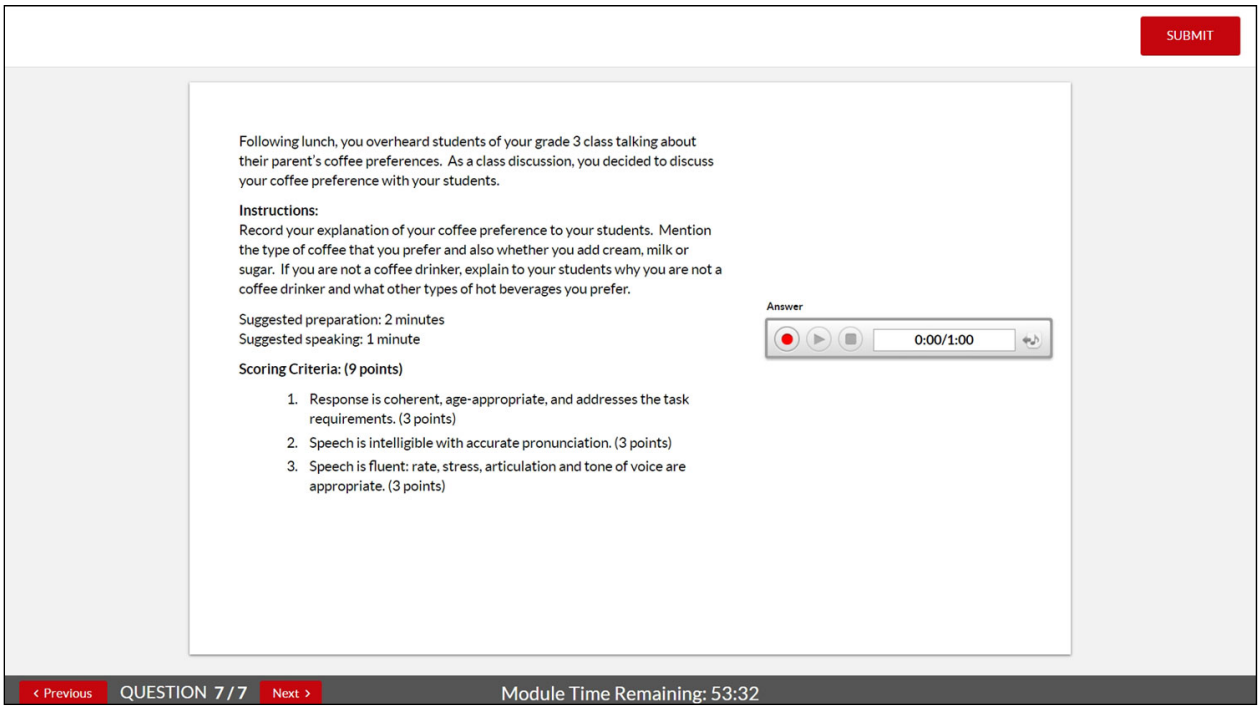
Writing: Tasks contained a stimulus, instructions, and coding criteria. Test takers provided a written response.

Figure 1. Writing response interface



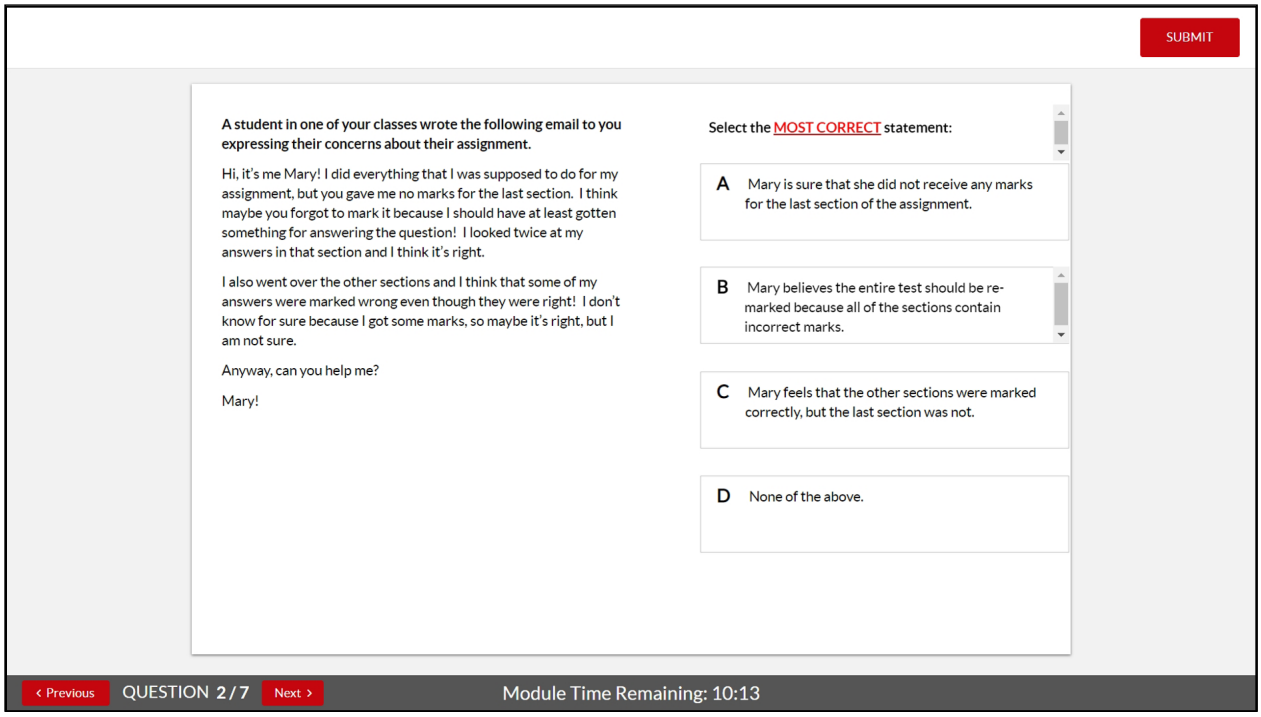
Speaking: Tasks contained a stimulus, instructions, and coding criteria. Test takers recorded a spoken response.

Figure 2. Speaking response interface



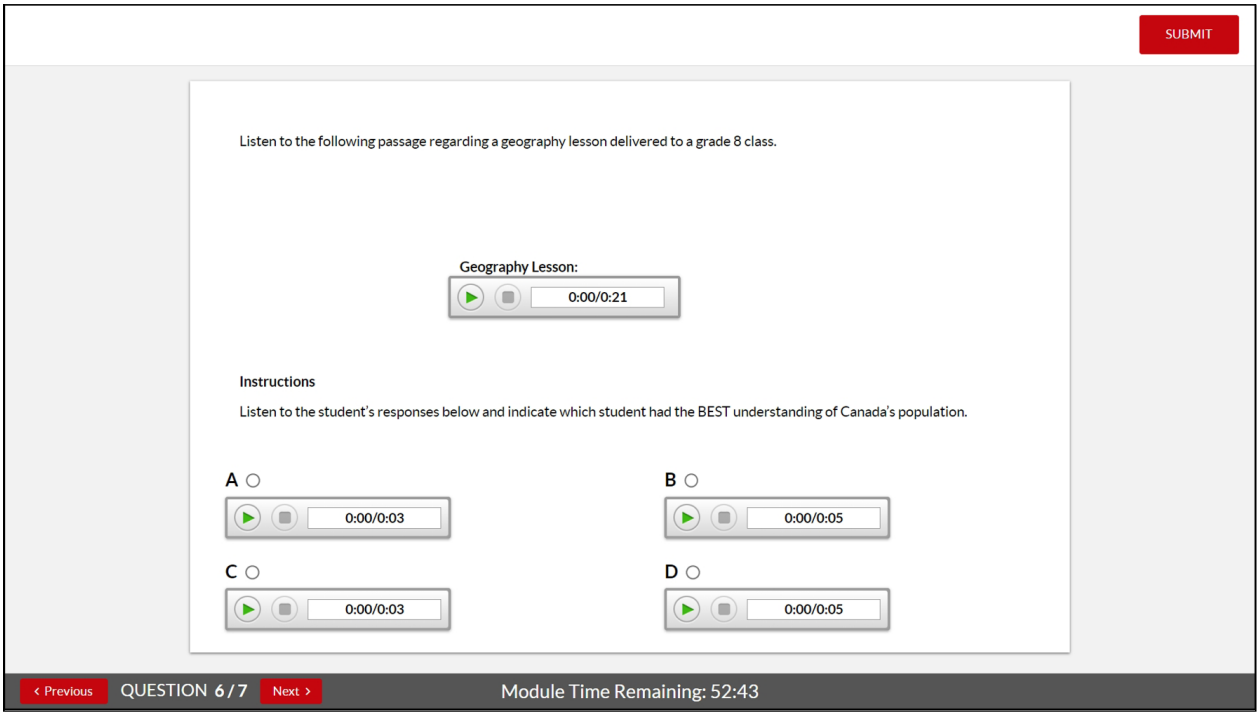
Reading or listening — Text-based multiple-choice response options: Tasks contained a text or audio stimulus and instructions on how to select a response from provided text options.

Figure 3. Text-based multiple-choice response interface



Listening — Audio-based multiple-choice response options: Tasks contained a text or audio stimulus along with instructions on how to select a response from provided audio options.

Figure 4. Audio-based multiple-choice response interface



Pilot-test administration and test takers

The RTCC recruited test takers between November 2020 and March 2021 and again from August to December 2021.

The pilot tests were remotely administered in two sessions. The first session (January to March 2021) had 70 test takers in French (across V2 and V3) and 143 in English (across V2 and V3). While psychometric analyses were completed after the first session, low numbers of test takers meant that statistical power was limited, especially among low-achieving test takers, because most test takers performed well on the test. Thus, a second pilot-testing session was conducted (between September and December 2021). This boosted total numbers across both sessions to 349 tests taken in French (across V2 and V3) and 589 tests taken in English (across V2 and V3).

Pilot test takers were primarily female certified teachers who had completed their teacher education in Canada in the testing language (e.g., English test takers had primarily taken teacher education in Canada and in English). Test takers were most proficient in the language of the test (e.g., French test takers were most proficient in French).

Coding the listening and reading items

The listening and reading modules were multiple-choice tests that were machine scored (0 for incorrect response, 1 for correct response) by the XpressLab testing software. The English V2 listening module had a single fill-in-the-blank item that was also machine scored (0 for incorrect response, 1 for correct response).

Coding the speaking and writing items

Two remote coding sessions for speaking and writing items took place: (1) from March to April 2021 (coding of January to March 2021 test data) and (2) in December 2021 (coding of September to December 2021 test data). Response coding was conducted online using an interface in XpressLab.

Coders

Coders and table leaders were experienced teachers, preferably with at least five years of teaching experience, with representation from teachers of a wide range of subject matters (e.g., art, music, language arts, science, mathematics, social studies) and levels (elementary, middle, high school). Many were experienced coders. Coders included nominees of the provinces/territories. Ninety-one individuals participated in one or two coding sessions. Coders were drawn from Alberta, Manitoba, Newfoundland and Labrador, Northwest Territories, Ontario, Prince Edward Island, and Quebec.

Training and coding

Coders took part in a live remote two-day training session in English or French, as appropriate, which included: an explanation of the coding process; independent coding of a common sample of speaking and writing responses drawn from actual test-taker responses; and table-based (groups of 4 to 10 coders with a table leader), table-leader, and large-group discussions of issues, coding disagreements, and challenges. Training materials included: videos on the assessment, coding process, and coding software; sample test items and responses; and a technical guide for coding and adjudication.

To enhance consistency among coders, a multi-step process was used. The first step involved coders examining three responses of low, medium, and high quality to one speaking and one writing item. This was used to establish an initial conversation and understanding about what separates different levels of response quality. The second step gave coders a common set of 10 speaking and 10 writing items to code. Coders and table leaders were given the coding data so they could discuss sources of disagreement and decide how to proceed and what coding criteria to use. Once this process had been established, coders were given 30 to 50 actual responses to speaking and writing items to code. When this set of responses was coded, the coding session paused. The *Directions* team met with table leaders (who also coded a selection of responses so they could understand the process and any challenges that their coders faced) to discuss emergent issues. The initial analyses and discussions provided information that was used to provide further written direction to coders on how to apply the scoring criteria. At this point coders could work at their own pace to complete their quota by the end of the coding session. Throughout the training and coding process, coders were encouraged to have frequent discussions within their tables using the Slack online communication tool on any issues that arose.

In assessments for which cut scores have been established, there are detailed coding guides. However, because no cut scores had been established for this assessment, there was no detailed coding guide for each item with examples of items that aligned with established standards. In assessment development, the coders effectively put the coding "standards" into practice and refine them as they code. Therefore, this was a more complex, interpretive process than coding an established assessment and it required continuous dialogue throughout the process among coders, table leaders, and *Directions*. This is a normal process during instrument piloting and informs the future development of a coding guide and standard setting.

Coders were asked to assign a score between 0 and 9 for each speaking or writing response, considering three coding criteria that accompanied each item. The three coding criteria varied across items but were in the following general format:

Writing: Coding criteria

- 1: Does the response address the task requirements? (3 points)
- 2: Is the communication audience appropriate? (3 points)
- 3: Conventions, grammar, punctuation, spelling, etc. are appropriate. (3 points)

Speaking: Coding criteria

- 1: Speech is coherent, age-appropriate, and addresses the task requirements. (3 points)
- 2: Speech is intelligible with accurate pronunciation. (3 points)
- 3: Speech is fluent: uses stress, articulation, rate, and tone of voice appropriate for the task. (3 points)

Coders were asked to provide a written rationale for each score to allow for analysis of how scores were assigned.

Two coders independently coded each writing and speaking response. If the pair of assigned scores differed by two or more points, it was considered a coding disagreement and the item, response, scores, and coder comments were reviewed by an adjudicator (table leader) who assigned the final score. If scores differed by only one point, the final score was an automatic average of the two scores.

Standard setting

Test data were subject to psychometric analysis before standard setting took place.

Standards (cut scores that determined a passing score for each module) were set by two separate standard-setting committees (one for English and one for French test versions). A modified Angoff process was used to set the cut scores for all four modules in both English and French tests.

Participants

Registrars in each province/territory were invited to nominate a standard setter in English and French. Standard setters had to have at least five years of experience teaching in Canadian classrooms and be familiar with standardized testing. For English standard setting, five individuals (nominated by Alberta, Nova Scotia, Northwest Territories, Ontario, Saskatchewan) participated. French standard setting also had five participants (nominated from Alberta, New Brunswick, Ontario, Quebec, Northwest Territories), one of whom also participated in the English standard setting.

Methods

Standard setters were informed that the purpose of the standard-setting exercise was to establish an entry-to-practice standard on the RTCC language-competency assessment for internationally educated teachers who did not complete a teacher education program in English or French.

Preliminary work: Before meeting as a group, standard setters were provided with material on the modified Angoff method and completed all V2 and V3 test items in English or French as test takers. Immediately after completing each item, standard setters were also asked to provide an initial estimate of a minimally competent candidate's performance on each item. For reading and listening items, the question was: What percentage of minimally competent candidates would answer this item correctly? For speaking and writing items, the question was: What score (out of 9) would you expect a minimally competent candidate to score on this item?

Standard-setting meetings: The virtual meetings took place over two days in each language. Meetings began with a review of the standard-setting process and reminders that standard setters needed to think of minimally competent candidates when setting the cut score, not average or competent candidates.

Listening modules were discussed first, followed by speaking, writing, and reading modules. For each module, the first round of discussion used the initially estimated cut scores to stimulate discussion about the items, their characteristics, and what types of performance or errors minimally competent test takers might make on the items. After seeing the score estimates for each item set by the other standard setters and discussing the rationale behind those score estimates, standard setters were given the opportunity to revise their cut scores.

In the second round of discussion for each module, the new cut score choices were shared along with statistical data about how the test takers performed on each item (mean score and standard deviation, native versus non-native speaker performance on individual items and on entire module). Standard setters were also told that no cut score is a perfect boundary between those who are minimally competent and those who are not competent enough, so they needed to consider whether it was more desirable to err on the side of leniency (allowing test takers to pass who may not be competent) or severity (failing test takers who may be competent). As cut scores were being discussed, standard setters were informed how many pilot test takers would pass or fail at that cut score.

For listening and reading modules, the discussion ended and cut scores were set. For speaking and writing modules, standard setters were given a selection of responses below, at, and above the cut score that was currently set. Based on those responses, standard setters could further reflect upon and change the cut scores for the speaking and writing modules.

Post-meeting reflection: Once cut scores had been established at the end of the two-day meetings, standard setters were given approximately two weeks to discuss the experience and cut scores set with their registrars and reflect on their satisfaction about whether the cut scores were appropriately set. English standard setters met again and made a slight adjustment to the cut score for listening module V2 but made no changes to the other cut scores. French standard setters made no changes to the cut scores after the period of reflection.

Definitions of competence: The standard setters described a minimally competent candidate in terms of their ability to meet the professional obligations of a teacher. This included being able to communicate clearly and effectively with different audiences such as parents, students, and administrators, as well as serving as a language role model to students. These descriptions of a minimally competent candidate aligned with elements of the conceptual framework such as "model appropriate language use," "give presentations to small and large groups of parents and other professionals," and "write emails or letters to parents in nontechnical language." The cut scores were set at a level that corresponds to Level 2 of the "Stages of Language Proficiency" described in the Framework for Language Competencies and Benchmarks.

Cut scores: In setting the cut scores, the standard setters were aware that no cut score represents a perfect boundary between competent and not competent language users. English and French standard setters agreed that including candidates whose language proficiency may be marginal was preferable to excluding competent candidates. The rationale was that (a) candidates' language proficiency often improves over time and with experience in the field and (b) passing the test was not a guarantee of employment. Thus, the cut scores represent consensus among standard setters about the lowest acceptable score required by a teacher whose language was minimally competent.

The cut scores resulting from standard-setting deliberations are presented in the section discussing Question 7. Are standards (cut scores) appropriately set? (see Table 9. Cut scores).

Validity study of the RTCC language-competency assessment

Tests are created to serve a purpose. A validity study examines to what extent a test is fit for its purpose. A validity study does not give a yes/no decision about whether a test is valid or not, but instead it presents an argument. The argument guides the test user and other stakeholders in their thinking and decision making about what the test measures, how reliable those measures are, and what types of decisions can be defensibly made from the test scores.

Several frameworks are available to anchor a validity study, but this study is most influenced by the work of Kane.[10] In Kane's framework, validity arguments are informed by a range of evidence and analyses. It is commonly used because of its pragmatic nature and direct applicability to licensure examinations.

The purpose of the RTCC language-competency assessment is to determine whether IETs who did not complete their teacher education in English or French have the language competencies required to be effective in Canadian K-to-12 classrooms. Thus, it is important to know that the tests measure appropriate language skills, that measurements are reliable and unbiased, and that standards are appropriately set. To that end, this validity study investigated the following questions:

[10] M.T. Kane, "Validating Interpretive Arguments for Licensure and Certification Examinations," *Evaluation and the Health Professions* 17, no. 2 (1994): 133–59.

1. Are the tests founded on an appropriate language-proficiency framework?
2. Do test items reflect the ways that teachers use language in Canadian schools?
3. Who were the test takers in the pilot tests?
4. What are the psychometric properties of the test?
   A. Do the individual modules have acceptable test reliability?
   B. Do the speaking and writing items demonstrate good interrater reliability?
   C. What are the item properties?
   D. At what ability levels do the tests provide good information?
   E. Do the tests measure what they claim to measure?
5. Are specific groups of test takers advantaged or disadvantaged by the tests?
6. Do test results correlate with other measures of language proficiency?
7. Are standards (cut scores) appropriately set?

The study's desired result is to allow registrars of teacher certification to make an evidence-informed decision about whether this test should be implemented as a means of determining whether IETs have the language skills required to be successful in a Canadian classroom. Success on this test would not be a guarantee, or even a predictor, of teaching effectiveness. It would simply indicate that a person's language skills will not *prevent* them from being an effective teacher in Canadian K-to-12 classrooms.

Question 1. Are the tests founded on an appropriate language-proficiency framework?

In Phase I of the project, the test developers conducted a literature review to inform the Framework for Language Competencies and Benchmarks (see the section on Phase I: Development of the RTCC Language-competency Assessment in this report). The review and framework were published by CMEC. The framework provided clear guidance on how to incorporate a range of occupation-specific elements into the test items. This ensured the tests provided an authentic assessment of language use in the context of teaching in Canadian classrooms. Examination of the processes for developing the framework,[11] items, and tests showed that the tests are founded on a theoretical framework that is research informed, relevant to practice, and available to all stakeholders.

Question 2. Do test items reflect the ways that teachers use language in Canadian schools?

Teaching is a complex activity that places complex demands on language use. In Phase I of the project, test items were created by individuals with direct and thorough knowledge of teaching in Canada to maximize the likelihood that items represented how teachers in Canada use language to accomplish their work effectively (see Test Item Development and Test Construction in this report). Ten assessment versions (five English, five French) were created from the item pool and reviewed by external, internationally recognized experts from the fields of educational assessment, teacher professional development, language in education, and curriculum development. Overall, the external reviewers judged the test versions positively with respect to authenticity, face validity, and content validity and provided suggestions for how to improve some items.[12] Item revisions incorporated feedback from the external reviewers (improvements to authenticity, face validity, and content validity), the RTCC Language Competency Subcommittee (ensuring no specific pedagogical or subject knowledge was required, not having separate tests for elementary and secondary teachers), and alpha testing (improvements to structure, format, and content of test items). In the process, items were revised according to

[11] "Speaking for Excellence: Language Competencies for Effective Teaching Practice," Council of Ministers of Education, Canada, 2013. https://www.cmec.ca/Publications/Lists/Publications/Attachments/320/Speaking_for_Excellence.pdf

[12] Authenticity: Do the items realistically represent tasks that a beginning teacher might encounter in the course of carrying out her or his responsibilities? Face validity: Do the items appear to measure the performance outcomes or competencies with which they are associated? Content validity: Taken as a whole, do the items include those competencies that are reasonable to assess given the constraints of the assessment?

suggestions or replaced with items deemed to meet the assessment standards. Phase II revisions (increasing the number of items in each module, simplifying coding criteria for speaking and writing items, copy-editing items) did not affect the content of the test questions.

During Phase II pilot testing, two data collection methods were used to further determine whether test items reflect a representative range of language uses in Canadian teaching contexts. The first was a survey that CMEC gave to test takers (n = 275). One survey question asked them to rate the test content on a five-point scale.[13] The mean rating was 4.04 and 82 percent of test takers rated the items as "excellent" or "good" (top two points on a five-point scale). The survey also asked for general feedback on the test. Of the 275 respondents, 28 chose to respond to the call for open-ended feedback on test items and commented on their connection to the realities of teaching in Canada; 26 of the 28 respondents indicated that the items were a good reflection of how language is used in the teaching profession. One respondent (English) felt the items did not reflect the "context and nuance of an actual classroom" and another respondent "feared the content did not 100 percent reflect the reality on the ground."

The other data were collected from two meetings (one English, one French) of coding table leaders. Table leaders were asked to gather and share comments from the coders in their table. For both English and French meetings, all table leaders agreed that the test items reflected typical tasks undertaken by teachers and sampled an adequate range of language skills.

No test can incorporate all the ways that language is used in an occupation, but it appears that test items realistically reflect a range of the ways that teachers in Canada use language in their work.

Question 3. Who were the test takers in the pilot tests?

Teachers who completed their teacher education internationally in a language other than English or French are the intended population for the RTCC language-competency assessment. To best generalize results from the pilot tests to the intended population, it is helpful if the characteristics of the pilot test takers reflect those of the intended population of test takers.

The pilot tests were administered in two sessions yielding a total of 349 tests taken in French (across versions V2 and V3) and 589 tests taken in English (across versions V2 and V3). Among test takers, 73 percent identified as female, close to the 75 percent proportion among teachers in Canada.[14] Three quarters (75%) of the test takers were fully certified teachers, with the remaining 25 percent enrolled in a teacher education program. Although 34 percent (n = 314) of the tests were completed by people who completed their teacher education internationally (across both languages and all test versions), only 5 percent (n = 50) of the tests (across both languages and all test versions) were completed by people who completed their teacher education in a language other than English or French. This is an important difference between the pilot and intended test populations. Another important difference is that only 4 percent of the test takers reported having a language other than English or French as their strongest language.

Differences between the intended test population and the pilot-test sample suggest that there is a strong possibility that mean item scores achieved by the pilot-test sample will be higher than for the intended test population. This can introduce two effects — the first is that some items may see a ceiling effect where the mean score is very high, reducing variance and the item's ability to discriminate between test takers with different language abilities. Regular data collection and analysis during test implementation are important aspects of test implementation. This process will provide ongoing data to help monitor item performance and appropriateness of cut scores.

The second effect is that standard setters may be swayed by high mean scores for items and set cut scores that are higher than desirable. Standard setters were given initial instructions and reminders to use their

[13] The survey question was "How would you rate the CONTENT of the Online Language Assessment specific to the teaching profession?"

[14] "Back to School … By the Numbers," Statistics Canada, 2018. https://www.statcan.gc.ca/en/dai/smr08/2018/smr08_220_2018#a8

concept of a minimally competent candidate to inform the cut scores. While standard setters were made aware of mean item scores, they were also made aware of who the pilot test takers were and were given information about how different groups of test takers performed on each item and each module (e.g., native vs. non-native speakers or those who completed their teacher education in Canada vs. those who completed their teacher education internationally).

Question 4. What are the psychometric properties of the test?

Two independent psychometric analyses were completed. *Directions* completed one set of analyses and used the Classical Test Theory (CTT) framework. CTT is an older psychometric model but it is well established and has been used often in test development. CTT is an easily understood framework that is useful when sample sizes are small. An external psychometrician conducted the second set of analyses, using the Item Response Theory (IRT) framework.[15] IRT is a powerful framework capable of giving detailed psychometric information about test items but it requires large sample sizes. There are different IRT models, each of which operates under its own set of assumptions and data requirements. Results from both CTT and IRT analyses are reported here. In most cases, the two different analyses yielded the same conclusions, but where differences exist, they are noted.

A. Do the individual modules have acceptable test reliability?

To examine the reliability of each test module, analysts used a measure called Cronbach's alpha. It is a commonly reported measure of internal consistency, meaning that it indicates to what extent the items in a test are measuring the same construct. The external psychometrician's analysis used an IRT model to determine a measure of test reliability. Table 2 presents both results. Generally, values of Cronbach's alpha below 0.70 are considered problematic and values below 0.75 for the IRT analysis are considered problematic.

Table 2. Test reliability

Test Version	Module	English		French	
		Cronbach's Alpha	IRT Test Reliability	Cronbach's Alpha	IRT Test Reliability
V2	Listening	0.25	0.42	0.42	0.42
	Reading	0.65	0.67	0.78	0.60
	Speaking	0.87	0.94	0.88	0.87
	Writing	0.85	0.90	0.77	0.79
V3	Listening	0.46	0.54	0.42	0.47
	Reading	0.61	0.71	0.70	0.65
	Speaking	0.87	0.88	0.85	0.85
	Writing	0.82	0.82	0.73	0.75

As Table 2 indicates, the listening modules had low reliability across all versions of the test. The reliability figures for the reading modules were better, but still problematic. The speaking and writing modules show acceptable reliability values across all versions of the test.

[15] While the external psychometrician completed the IRT analyses, the conclusions and recommendations presented in this report come from *Directions*. The external psychometrician is not responsible for, nor liable for, any of the content presented here. The external psychometric report is available from CMEC upon request.

As a result of the psychometric analyses, specific items were reviewed because of their psychometric properties.[16] As well, items were reviewed to ensure they are error free and culturally appropriate. This recommendation is based upon findings from a test-taker exit survey and meetings with table leaders.

Overall, the listening and reading modules had problematic reliability and their items did not cohere as a one-dimensional scale in their current form. They were also quite easy for the pilot test takers. The items in the speaking and writing modules acted as scales with good reliability.

B. Do the speaking and writing items demonstrate good interrater reliability?

Before describing the interrater reliability, it is important to remind readers how items were coded. Speaking and writing items were coded on a 10-point scale ranging from 0 to 9. Each item was coded based upon three criteria; each of these criteria was coded on a four-point scale ranging from 0 to 3. All responses were coded by two coders. If the two coders were within 1 on the total item score, the final score was an average of the two scores (e.g., a 6 and a 7 were averaged to 6.5). If the disagreement between the two coders was two or more points, an adjudicator (table leader) reviewed the disagreement. After considering the two assigned scores and rationales, the adjudicator assigned a final score. Thus, every speaking and writing score is the result of the combined judgment of two or three coders.

Three measures were used to examine interrater reliability (Table 3):

1. **Percentage agreement.** Two raters were deemed to have agreed if they were within one point of each other in their total scores assigned to a response. Across all modules, agreement levels were encouraging, but this number includes nonresponses / blank responses, which were awarded a zero. It is very easy for coders to agree on a score of zero for a nonresponse, which inflates the level of agreement.
2. **Cohen's Kappa.** This statistic looks at the amount of exact agreement between raters, correcting for the possibility of obtaining the same rating by random chance. Values of Cohen's Kappa below 0.20 are considered problematic.
3. **Intraclass correlation coefficient.** This statistic examines the correlation between the scores of the two raters for each item. Values below 0.80 are considered problematic.

Table 3. Interrater reliability: Range of values

Language	Test Version	Module	Percentage Agreement	Cohen's Kappa	Intraclass Correlation Coefficient
English	V2	Speaking	71% – 81%	0.20 – 0.40	0.72 – 0.90
		Writing	62% – 77%	0.22 – 0.30	0.76 – 0.88
	V3	Speaking	70% – 85%	0.22 – 0.43	0.80 – 0.92
		Writing	68% – 76%	0.19 – 0.33	0.80 – 0.88
French	V2	Speaking	61% – 79%	0.18 – 0.37	0.82 – 0.91
		Writing	62% – 72%	0.22 – 0.35	0.80 – 0.87
	V3	Speaking	69% – 80%	0.24 – 0.42	0.83 – 0.92
		Writing	61% – 76%	0.17 – 0.34	0.71 – 0.87

Details on interrater reliability, especially relating to individual items, are provided in the full psychometric report. In general, the speaking and writing items had good interrater reliability.

[16] Detailed psychometric analyses are available in the full psychometric report, which is available upon request.

C. What are the item properties?

This summary is intended to give readers an understanding of the psychometric strengths and weaknesses of each module and to develop an overall understanding of how well each module performs. A full description of all item properties can be found in the full *Directions* and external psychometric reports. These reports include detailed descriptions of item properties for every item on every test.

Listening

The listening modules consisted of 13 items and were easy for the pilot test takers. The mean score for the English listening modules was 77 percent for V2 and 83 percent for V3. The mean score for the French listening module scores was 71 percent for V2 and 77 percent for V3. A subset of items were flagged for review in the English and French tests. To be flagged, an item had to have a mean score higher than 95 percent, lower than 50 percent, or be a poor discriminator (i.e., high-ability and low-ability test takers achieve similar scores on the item). Some items had negative discrimination meaning that high-ability test takers performed worse on the item than low-ability test takers. Generally, corrected item-total correlations above 0.20 are considered desirable so the test developers reviewed items whose corrected item-total correlations were below 0.20 and especially items with negative corrected item-total correlations.[17] Higher values of the corrected item-total correlation and discrimination coefficient indicate a better ability for the item to discriminate.

There is nothing inherently wrong with mean item scores above 95 percent or below 50 percent. These items were flagged because very easy items tend to be poor discriminators and having too many easy items on a test tends to reduce the information the test provides. Thus, easy items were reviewed to determine if they could be improved or removed from the test. Likewise, there is nothing inherently wrong with items having a mean score below 50 percent, but because test takers found most items to be easy, more difficult items were reviewed to ensure the difficulty was not because of confusing wording or attractive distractors.

Reading

The reading modules were also easy for the test takers. The English modules had 22 items and the mean score was 87 percent for V2 and 89 percent for V3. The French modules had 19 items, and the mean score was 82 percent for V2 and 78 percent for V3.[18] A subset of items was flagged for review in the English and French tests. As with the listening modules, flagged items had mean scores above 95 percent, below 50 percent, or negative corrected item-total correlations. The results for listening and reading come from the CTT analyses. The IRT analyses conducted by the external psychometrician yielded results that generally agreed with the CTT analyses.

Speaking

The English V2 speaking module had 12 items with mean scores ranging from 7.64 to 7.97 on the scale of 0 to 9; the English V3 speaking module had 13 items with mean scores ranging from 7.18 to 8.27; scores of 0 from nonresponses were excluded. Corrected-item total correlations ranged from 0.49 to 0.64 for V2 and 0.49 to 0.59 for V3. These numbers indicate the speaking modules had good ability to discriminate.

The IRT analyses conducted by the external psychometrician demonstrated that most test takers did well on the speaking items so there was a high probability of test takers receiving a 9 out 9 for the items. This is

[17] *Corrected item-total correlation:* A measure of the discrimination of each item, calculated as a Pearson correlation coefficient between the item score and the scale score with a given item deleted. A Pearson correlation coefficient is an index of the degree of linear relationship between two variables. It is often known as the Pearson product-moment correlation coefficient (Pearson's r) and is one of the most used sample correlation coefficients. It is scaled so that the value of +1 indicates a perfect positive relationship (such that high scores on variable x are associated with high scores on variable y), –1 indicates a perfect negative relationship (such that high scores on variable x are associated with low scores on variable y, or vice versa), and 0 indicates no relationship.

[18] Not all modules within a modality have the same number of questions. This is because during item development, an estimated test-taker completion time was assigned to each item and, during test construction, each module was constructed so that all modules had approximately the same estimated length.

likely because there were few test takers with low scores so there were insufficient data at low ability levels to generate accurate statistical information about the module's ability to discriminate at low ability levels.

Both French speaking modules had 12 items. Item mean scores ranged from 5.98 to 6.88 for the French V2 speaking module and 5.77 to 6.83 for French V3. Corrected item-total correlations ranged from 0.45 to 0.56 for V2 and 0.41 to 0.48 for V3.

The IRT analyses conducted by the external psychometrician showed the same pattern in French as in English. Higher abilities were associated with higher scores, but discrimination is weak at lower ability levels. Again, this is likely because of the low number of respondents who received low scores. For example, the French V3 speaking module had only three items with any responses that were awarded a total score of 1 (and only one person received this score for each of these three items), and only seven items had any responses that were awarded a total score of 2 (with a maximum of three people receiving this score). With so little data at the low end of the score scale, it is impossible to calculate robust statistics related to those score points.

In summary, the CTT analyses indicate the speaking items were easy for the pilot test takers, had good interrater reliability, and discriminated well. The IRT analyses demonstrate the items were easy for the test takers, but items struggled to discriminate at lower ability levels.

Writing

The English V2 and V3 writing modules both had seven items each and mean item scores ranged from 6.72 to 7.52 (V2) and from 6.89 to 7.56 (V3); scores of 0 from nonresponses were excluded. While this cohort of test takers still performed well in the writing modules, the mean scores were lower than for the speaking modules. Corrected-item total correlations ranged from 0.56 to 0.65 (V2) and from 0.53 to 0.59 (V3); these numbers indicate the writing modules had good ability to discriminate.

The IRT analyses conducted by the external psychometrician demonstrated that most test takers did well on English writing items so there was a high probability of test takers receiving a 9 out 9 for items. For moderate ability levels there was a higher probability of being coded as an 8 on an item instead of 9. This is different than for speaking items and suggests the writing items have a better ability to discriminate at moderate ability levels. As with speaking modules, the analyses indicate that higher scores on writing items correspond to higher ability levels, but the items struggle to discriminate at lower levels.

Like the English modules, the French writing modules had seven items each; mean item scores ranged from 5.98 to 6.75 (V2) and from 5.77 to 6.83 (V3); scores of 0 from nonresponses were excluded. While this cohort of test takers still performed well in the French writing modules, the mean scores were lower than for French speaking modules. Corrected-item total correlations ranged from 0.45 to 0.56 (V2) and 0.41 to 0.48 (V3); these numbers indicate the writing modules had good ability to discriminate.

The IRT analyses conducted by the external psychometrician demonstrated that most test takers did well on writing items, but compared to English tests, there was a higher probability that test takers received a score of 8 out 9 for the French items at higher ability levels. This suggests the writing items were more difficult for French than English test takers and that, compared to speaking items, the writing items have a better ability to discriminate at moderate ability levels. As with the speaking modules, higher scores on writing items correspond to higher ability levels but the items struggle to discriminate at lower levels. IRT analyses for all items can be found in the external psychometrician's report.

D. At what ability levels does the test provide good information?

The CTT analyses performed by *Directions* do not allow a good response to this question, but the IRT analyses conducted by the external psychometrician provide some insight. (Full graphs of test information functions can be found in the external psychometric report.) According to that work's general conclusions,

the listening and reading modules are most accurate at low ability levels and provide poor test information at average or high ability levels. This is true across all four tests. The English V3 listening module (test) information function is particularly narrow, meaning the module provides good information about test takers only within a narrow range of low-ability learners. These findings are not necessarily problematic if the cut score is set at a level where the module has good information and low standard error of measurement. If the cut score is set outside of the ability range where the module provides good information, then the module would not be performing well at the cut score. The cut score set by standard setters (see Question 7. Are standards (cut scores) appropriately set?) is below the level where the module is providing good information. While this is not ideal from a psychometric standpoint, test takers are very unlikely to be disadvantaged because the cut score is set very low.

For the speaking modules, the test information functions indicate the test performs best from below- to above-average abilities but does not perform well with test takers who are at the extremes (i.e., two standard deviations below or above average). This is an appropriate range of ability in which the speaking module performs well, but it should be noted that standard errors of measurement are higher than ideal. The results are the same for the writing modules, although the French V2 writing module information function is not interpretable so no conclusions can be made about the performance of this module across ability levels.

E. Do the tests measure what they claim to measure?

Two different types of factor analysis were conducted to determine the tests' structure. For language tests, there is an expectation that scores across modalities are correlated. This is because while listening, reading, speaking, and writing are distinct skills, people who are good at one language skill tend to be good at all of them. For example, it would be difficult to be an excellent writer in a language without also being an excellent reader.

The first type of factor analysis — exploratory factor analysis (EFA) — is useful when the structure of a test is unknown. The EFA found the tests to be one dimensional, with moderate to strong correlations between modules.[19] Table 4 shows the factor loading for each module onto a single factor. The factor loading gives an indication of how well the module scores correspond to general language ability in the context of teaching. Ideally, factor loadings should be above 0.30.

Table 4. Factor loading of the four modules in each test

Language	Test Version	Listening	Reading	Speaking	Writing
English	V2	0.54	0.75	0.82	0.82
	V3	0.61	0.70	0.66	0.75
French	V2	0.50	0.56	0.70	0.84
	V3	0.49	0.53	0.63	0.75

Table 5 shows the Pearson correlation coefficients between different modules for English V2 and V3 tests; Table 6 presents them for the French V2 and V3 tests. Low correlation coefficients (below 0.20) suggest that the two skills are independent of each other, whereas very high coefficients (above 0.80) suggest the two modules are measuring the same skill. The strength of the correlations between modules suggest that the modules are measuring skills that are related, but not identical.

[19] A one-dimensional or unidimensional scale measures a single construct, trait, or attribute.

Table 5. Pearson correlation coefficients between modules of the English test

		V2			
		Listening	**Reading**	**Speaking**	**Writing**
V3	**Listening**		0.49	0.36	0.40
	Reading	0.61		0.60	0.56
	Speaking	0.28	0.38		0.73
	Writing	0.38	0.45	0.64	

All correlations are significant at $p < .01$.

Table 6. Pearson correlation coefficients between modules of the French test

		V2			
		Listening	**Reading**	**Speaking**	**Writing**
V3	**Listening**		0.35	0.38	0.37
	Reading	0.39		0.30	0.49
	Speaking	0.23	0.33		0.59
	Writing	0.41	0.43	0.53	

All correlations are significant at $p < .01$.

The external psychometrician carried out a Confirmatory Factor Analysis (CFA) that tests which of several predetermined models best fits the data. They found that for English V2, French V2, and French V3 tests, the best model has a general language factor with four components (i.e., listening, reading, speaking, writing). For the English V3 test, the best model had one underlying factor for the whole test (i.e., all modules were measuring the same general language skill). These results suggest that the four modules are measuring separate, but related skills. This aligns with the conceptual framework used to design the RTCC language-competency assessment. This framework describes an overall skill (language proficiency in the context of teaching) as having four components (listening, reading, speaking, and writing). Please see the external psychometrician's report for a full description of the models and fit indices.

Question 5. Are specific groups of test takers advantaged or disadvantaged by the tests?

To examine whether specific groups of test takers were advantaged or disadvantaged by the tests, two analyses were conducted: test impact and Differential Item Functioning (DIF).

Test impact

Test impact compares how different groups performed on the test (i.e., mean score for members of that group). Test impact examined two different demographic factors: gender and strongest language (Table 7). T-tests were used to find statistically significant differences in the mean score per item completed.

Table 7. Gender and language differences in performance

Language	Test Version	Gender Differences	Language Differences
English	V2	No differences in any module	Those who identified English as their strongest language outscored those who did not identify English as their strongest language. This is an expected finding and not evidence of test bias.
	V3	Females outscored males on speaking items (Cohen's $d = 0.31$)	
French	V2	Females outscored males on reading (Cohen's $d = 0.35$) and writing items (Cohen's $d = 0.42$)	Those who identified French as their strongest language outscored other test takers on the speaking and writing items, but not on the listening and reading items.
	V3	Females outscored males on reading (Cohen's $d = 0.42$) and writing items (Cohen's $d = 0.49$)	

Differential item functioning (DIF)

While test impact can provide some clues about where biases may exist in a test, because it does not control for test-taker ability, it can also produce misleading results. For example, on language tests, native speakers usually outscore non-native speakers. This is not necessarily bias but could be because native speakers genuinely have higher language ability than non-native speakers. DIF is a more sophisticated approach to examining bias that incorporates test-taker ability into the analysis.

Three statistical approaches were used to find DIF in the test items. *Directions* used Mantel-Haenszel statistics and logistic regression, while the external psychometrician used IRT methods. Mantel-Haenszel statistics work well with small sample sizes and are useful for finding uniform DIF, which is when an item is biased against a group across all ability levels. Logistic regression is also useful with smaller sample sizes but is a more sophisticated technique that can find both uniform and non-uniform DIF. Non-uniform DIF is when an item behaves differently across ability levels. For example, an item may exhibit no DIF for low-ability test takers but be advantageous toward males in high-ability test takers. The IRT analyses conducted by the external psychometrician are a very effective approach for finding DIF, but they require large sample sizes and the assumptions of the IRT model to be met.

Different methods of calculating DIF yield different results (e.g., for one item the logistic regression analysis showed DIF across genders, whereas for another item both the Mantel-Haenszel test and logistic regression showed evidence of DIF across genders). Thus, items where two or more methods demonstrated DIF were given highest priority for review. Note that DIF analyses reported here do not examine which group is advantaged or disadvantaged by items demonstrating DIF. This is a further analysis that should be conducted for any items where DIF is present. Items that demonstrate DIF did not necessarily need to be discarded, but item review was conducted to ensure that any group differences in achievement on an item were not severe and that the same group was not consistently disadvantaged.

Overall, there were relatively few items that demonstrated DIF, especially where more than one method of analysis found DIF. These cases were examined as part of the item review because it is important from a fairness and equity perspective that items demonstrating DIF were examined closely to determine what group is disadvantaged and to what extent, as well as why the DIF exists.

Demographic information related to race, sexual orientation, or equity-deserving groups was not collected so no analyses could be completed for these groups. More demographic information should be collected during test implementation to allow for more thorough DIF analyses to be conducted to ensure the tests are not biased or disadvantaging specific groups of candidates.

Question 6. Do test results correlate with other measures of language proficiency?

Ideally, in a test-validation study, test scores are compared to other measures of the construct. In this pilot testing, the only other measure of language proficiency was a self-rating where candidates rated their language proficiency on a six-point scale. The association between candidates' test performance and self-rating was investigated using linear regression (Table 8). For both English tests, the self-rating was significantly associated with mean item scores for all modules. For both French tests, the self-rating was significantly associated with mean item scores for speaking and writing. Neither French test had a significant association between the test taker's self-rating and listening mean item score, and only the French V2 test had a significant association between the test taker's self-rating and reading mean item score.

Table 8. Linear regression results with self-reported ability as the independent variable and mean item score as the dependent variable

Test Version	Listening	Reading	Speaking	Writing
English V2	Constant = 0.570 Slope = 0.040*	Constant = 0.692 Slope = 0.035*	Constant = 3.547 Slope = 0.772*	Constant = 1.059 Slope = 1.102*
English V3	Constant = 0.552 Slope = 0.052*	Constant = 0.682 Slope = 0.040*	Constant = 3.441 Slope = 0.782*	Constant = 1.987 Slope = 0.920*
French V2	Constant = 0.604 Slope = 0.019	Constant = 0.777 Slope = 0.013*	Constant = 3.017 Slope = 0.814*	Constant = 3.772 Slope = 0.503*
French V3	Constant = 0.660 Slope = 0.021	Constant = 0.688 Slope = 0.000	Constant = 4.547 Slope = 0.546*	Constant = 3.303 Slope = 0.580*

The constant is the expected score if someone rated their ability as zero. The slope is the increase in expected score for every one-point increase in self-rating. For instance, for English V2, someone who rated their proficiency as 5 would have an expected mean item score of 0.570 + 5 x 0.040 = 0.77. For French V3 speaking, someone with a self-rating of 4 would have an expected mean item score of 4.547 + 4 x 0.546 = 6.731.
*Statistically significant association between mean item score on module and self-rating of language proficiency.

In the future, it would be worthwhile to consider asking test takers to report the results of other language tests they have taken (e.g., DELF, TOEFL, language tests for immigration) to provide a better data set to investigate the correlations between performance on the RTCC language-competency assessment and other language-proficiency tests.

Question 7. Are standards (cut scores) appropriately set?

Directions used a modified Angoff method to set standards.[20] The Angoff method is a robust and legally defensible method commonly used for standard setting, especially for examinations with different item types. Modified Angoff methods of setting standards (cut scores) rely upon the judgment of a panel of experts. The cut scores' defensibility thus depends upon the panel members' expertise. The panel members for standard setting had a minimum of five years of teaching experience in Canadian classrooms and were nominated as experts by the registrars in their respective province. Thus, all panel members had classroom teaching experience, as well as experience with teacher certification. Initial cut scores were based upon the panel's judgment of the proportion of minimally competent test takers who would answer listening and reading items correctly, and on the expected score for a minimally competent candidate on the speaking and writing items.

The standard-setting process implicitly assumed that test takers would complete all items in a module. Analysis of test-takers' performance, combined with results of the exit survey that they completed, revealed that many test takers struggled to complete the speaking and writing modules within the 60-minute time limit. In the interest of fairness, *Directions* reduced the number of items in the speaking and writing modules and adjusted the cut scores accordingly. The new speaking and writing cut scores were established so that

[20] For a description of the cut-score-setting process, please see the section on standard setting.

the average score per item remained consistent with the initial cut scores that were set. As a result of item review,[21] some items were also removed from the reading modules. The cut scores for reading were then revised based on the reduced number of items and consistent with the initial cut scores that were set for the proportion of minimally competent test takers expected to answer reading items correctly.

The standard setters then reviewed and approved the revised cut scores and justifications for the revisions. When the RTCC language competency assessment is implemented, ongoing data collection and analysis should include examination and revision of the cut scores, which is a normal part of ensuring a test's ongoing quality. The results of standard-setting discussions and deliberations are shown in Table 9.

Table 9. Cut scores

Language	Module	V2 Test	V3 Test
English	Listening (13 items)	6.5 (max = 13) * 20 of 288 pilot test takers fail	6.5 (max = 13) 7 of 291 pilot test takers fail
	Reading (18 items)	12.5 (max = 18) 26 of 283 pilot test takers fail	12.5 (max = 18) 24 of 293 pilot test takers fail
	Speaking (8 items)	47.7 (max = 72) 48 (max) of 271 pilot test takers fail	47.7 (max = 72) 51 (max) of 277 pilot test takers fail
	Writing (5 items)	24.7 (max = 45) 49 (max) of 272 pilot test takers fail	24.7 (max = 45) 58 (max) of 283 pilot test takers fail
French	Listening (13 items)	7.5 (max = 13) 26 of 163 pilot test takers fail	7.5 (max = 13) 18 of 176 pilot test takers fail
	Reading (18 items)	12.5 (max = 18) 17 of 161 pilot test takers fail	12.5 (max = 18) 34 of 176 pilot test takers fail
	Speaking (8 items)	55.7 (max = 72) 66 (max) of 153 pilot test takers fail	55.7 (max = 72) 87 (max) of 163 pilot test takers fail
	Writing (5 items)	29.7 (max = 45) 70 (max) of 158 pilot test takers fail	29.7 (max = 45) 87 (max) of 165 pilot test takers fail

*In each table cell for the tests, the first line gives the cut score and the second line gives the number of pilot test takers who would fail with that cut score for the module.

The cut scores are identical between V2 and V3 test versions within English and French. Given that the two versions of the test appeared to be of equivalent difficulties, this is appropriate.

With the current cut scores, a larger proportion of French pilot test takers would fail the test than English pilot test takers. While this is not necessarily problematic (e.g., there may be legitimate reasons such as different cohorts of test takers in English and French during pilot tests), for equity purposes it would be important to monitor pass rates for the two languages to ensure that French test takers are not disadvantaged compared to their peers taking the English test (or vice versa).

[21] After the psychometric analyses were completed, the *Directions*' team examined evidence from their own and external psychometrician's reports in an overall review of the items and to make suggestions about which items should be most targeted for review. These items were then reviewed by a team of experts in language education who made recommendations for revising or removing items in cases where appropriate revisions were not possible.

Recommendations and Considerations for Enhancing Pan-Canadian Labour Mobility in the Teaching Profession

The adoption of the RTCC Language-Competency Assessment would meaningfully contribute to interjurisdictional cooperation to ensure equitable labour mobility in Canada. The journey began when the registrars responsible for teacher certification met to identify obstacles to pan-Canadian labour mobility.[22] The first significant milestone was an agreement to ensure that the certificate issued to a teacher in one province or territory would also be issued in another province or territory when the teacher moved.

When the registrars decided to examine and remove obstacles to certification for internationally educated teachers, they achieved another milestone. The registrars sought to determine whether the assessment of language competency for internationally educated teachers was an obstacle to their certification.

Working under the auspices of CMEC and with support from Employment and Social Development Canada (ESDC, previously Human Resources and Skills Development Canada), the registrars committed to developing assessments to assess the language competencies of internationally educated teacher candidates who cannot provide evidence of having completed an acceptable teacher-education program delivered in English or French.

The registrars commissioned *Directions* to conduct a literature review to identify the specific competencies that K-to-12 teachers in English-first-language and French-first-language schools in Canada need to teach effectively. The literature review and the Canadian Language Benchmarks helped to inform the Framework for Language Competencies and Benchmarks that *Directions* developed on behalf of the registrars. The framework identifies competencies in each of four modalities: speaking, listening, reading, and writing. Each competency in the framework includes specification of performance outcomes in three domains of teaching practice: instructing and assessing, managing the classroom and student behaviour, and communicating with parents and other professionals.

The registrars also engaged *Directions* to appraise the language assessments that were in use at the time (2011) and conduct an interjurisdictional scan to examine whether a language-proficiency assessment was available that is suitable for internationally educated teachers seeking teacher certification in Canada. The conclusion was that no existing assessments met all the requirements for the teaching profession (see section on Phase I: Development of the RTCC Language-competency Assessment).

Having reached another milestone in its journey to ensure fairness in certification with the production of the Framework for Language Competencies and Benchmarks, the registrars commissioned the development of assessments to assess competence in each language modality and for each domain of teaching practice. *Directions* developed and piloted the RTCC Language-Competency Assessment and confirmed its validity for assessing language competencies for teaching in IETs whose teacher education was in neither French nor English.

The next significant milestone for fairness in certification is to adopt and employ the RTCC Language-competency Assessment. The following recommendation and accompanying considerations are based on the psychometric evidence summarized in this report, as well as the evidence in the full psychometric reports by *Directions* and the external psychometrician. The recommendation and considerations are intended to help the registrars decide whether to implement this test, and if so, what conditions should be placed upon implementation.

[22] The Canadian Free Trade Agreement (CFTA) reaffirms the labour mobility provisions and obligations established under the 1995 Agreement on Internal Trade (AIT). CFTA Labour Mobility provisions (Chapter 7) state that certified workers have to be recognized as qualified to work by a regulatory body in another province or territory which regulates that occupation, without having to go through significant additional training, work experience, examination or assessment, unless an exception has been posted.

Recommendation

Directions strongly recommends the adoption of the RTCC Language-competency Assessment (French and English) because thoughtfully using the assessments provides a fair and defensible method of determining which internationally educated teachers meet the standards set by the RTCC for language competencies needed for teaching in Canada.

Implementation considerations

Consideration 1: Use the language-competency framework that informed the development of these assessments in further refinements of the assessments

The RTCC Language-competency Assessment is founded on a research-informed and useful language-competency framework that effectively guided test development and interpretation. The assessment items reflect the language skills and contexts that Canadian K-to-12 teachers face in their work environment. The framework should continue to guide further work on the assessment.

Consideration 2: Create a coding guide

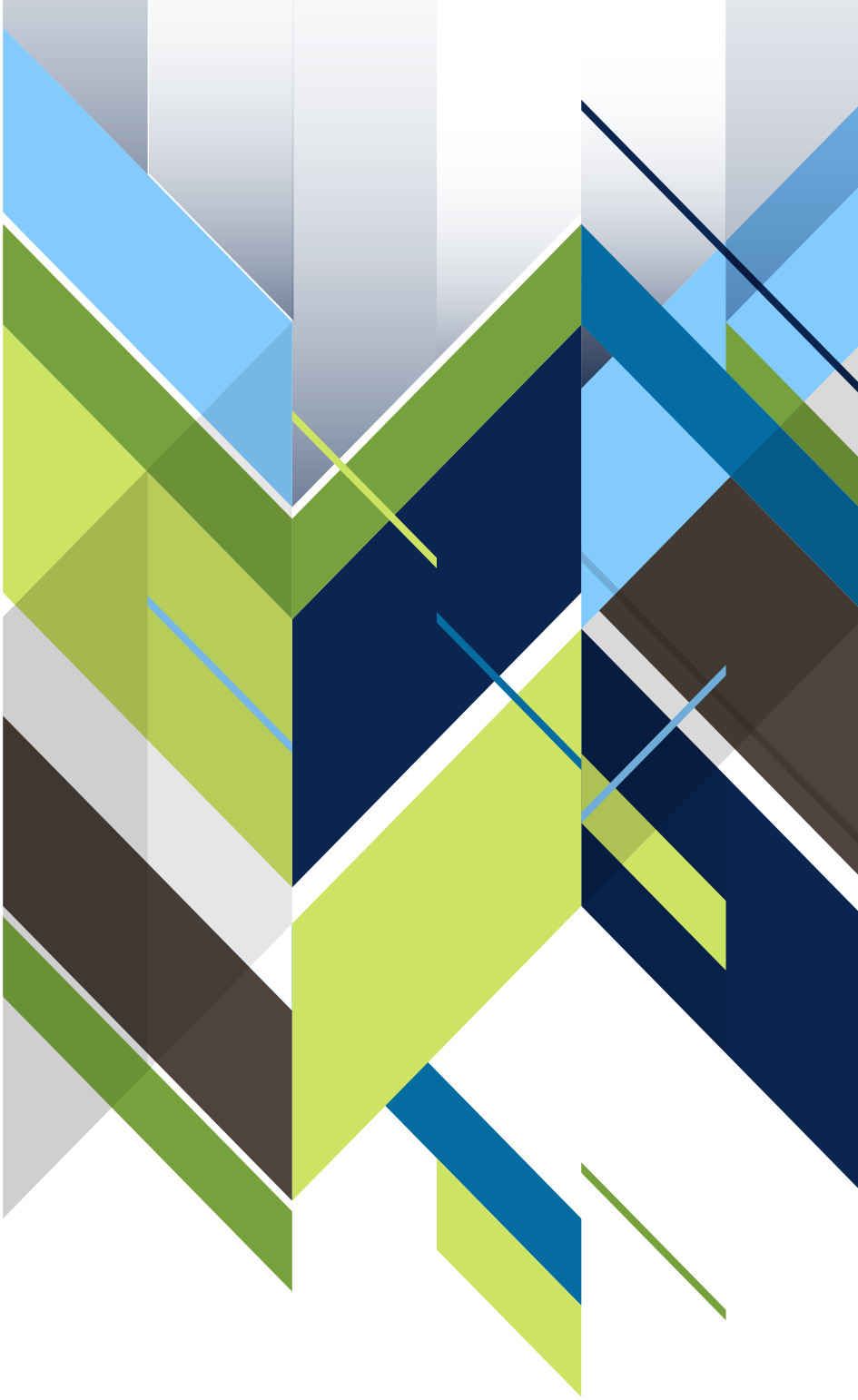
A coding guide for speaking and writing items is an essential component of test implementation. While the initial advice given to coders seemed to be effective in promoting consistent, reliable coding, providing them with a coding guide would align with known effective practices in testing and would also likely yield improved item reliability and interrater reliability. The coding guide would also add clarity for standard setters whenever standards are revised. The advice given to coders during the pilot-test coding sessions provides a good start for a coding guide, but this document needs to be formalized and should be reviewed by language-assessment experts, projected test users, and coders (or table leaders) from the test-coding sessions.

Consideration 3: Collect and analyze test data on an ongoing basis

The ongoing collection and analysis of test data is a normal part of any quality testing program. It is especially important with the RTCC language-competency assessment because the pilot-test population did not match the intended test population. Thus, psychometric properties of the items may change during implementation. It would be useful if, during implementation, test takers were asked to submit scores they received on other language tests (e.g., DELF, IELTS, TOEFL). These scores would provide information useful for understanding how the RTCC language-competency assessments compare to other language tests and for understanding how the cut scores relate to other proficiency standards. The ongoing data collection and analysis should look at item performance, coder performance, test bias, and whether cut scores are appropriately set.

Consideration 4: Report format for test takers

Language-competency-assessment test takers will be provided with a report that describes whether they have achieved the minimum required score (standard) for each one of the modalities. Test takers must achieve the minimum score in all four modalities to pass the assessment.



RTCC Language-competency Assessment

Phase II: Results of the Pilot-Testing Process

FINAL REPORT

www.cmec.ca

Funded in part by
the Government of Canada's
Foreign Credential Recognition Program

Canada