An evidence-based review of Celpe-Bras: The exam for certification of proficiency in Portuguese as a foreign language

Cecilia Guanfang Zhao
University of Macau, China

Carina Jiayu Liu
University of Macau, China

Keywords
Celpe-Bras, communicative language testing, large-scale proficiency test, Portuguese as a foreign language, test fairness and justice framework

Celpe-Bras, the exam for the certification of proficiency in Portuguese as a foreign language, is the only Portuguese proficiency test recognized by the Brazilian government (Ministério da Educação, 2013). The exam was first launched in 1998 with 127 examinees taking the test in five testing sites (Ministério da Educação, 2013). By 2018, Celpe-Bras had been administered to a total of 113,172 test takers at 94 testing sites across 37 countries around the world (UFRGS, n.d.). Given the recent growth of interest and also its unique design as a large-scale proficiency test, this review seeks to examine the overall usefulness and fairness of Celpe-Bras, using Kunnan’s (2018) most recent model of test fairness and justice. As the current field of language testing still tends to focus on the assessment practices and issues of English language testing, it is also our hope that this review, together with other such recent reviews (e.g., Elder, 2018), could help to raise awareness about the assessment of world languages other than English, hence expanding our profession’s scope and impact.

Corresponding author:
Cecilia Guanfang Zhao, Department of English, University of Macau, Faculty of Arts and Humanities, E21-4074, Avenida da Universidade, Taipa, Macau, China.
Email: czhao@um.edu.mo
General description

Test purpose: According to the Candidates’ Manual (Ministério da Educação, 2015a), Celpe-Bras is used to determine a candidate’s Portuguese language proficiency for the purposes of study and employment in Brazil. The test may also be used to demonstrate test takers’ ability to use Portuguese for communication purposes in their home country outside of Brazil.

Structure of the test: Celpe-Bras measures communicative language ability with fully integrated performance-based test tasks. Overall, it comprises a written part and an oral part. The former includes four tasks that integrate listening, reading, and writing skills. Two such tasks require test takers to write in response to a video or an audio of approximately 3 minutes long (played twice), whereas the other two tasks require test takers to write in response to one or more reading passages. All the writing prompts specify four rhetorical elements: the action (what to do), the writer’s role (who writes), the interlocutor (writing to whom), and the objective of the action (writing for what purpose). Test takers should therefore write in this specified rhetorical situation and select adequate and appropriate information from various inputs to fulfil specific communicative purposes. The genres of writing focus primarily on argumentation, exposition, and description, based on Silva and Cunha’s (2017) analysis of 88 writing tasks in Celpe-Bras between 2005 and 2015.

Unlike the written part that is administered collectively with all the test takers at a testing site, the oral part is carried out individually. It is a 20-minute face-to-face audio-taped interaction between a candidate and two examiners (an interlocutor and an observer). The first section is a 5-minute basic interpersonal conversation about a test taker’s personal background. In the second section, the interlocutor provides such stimuli as pictures, advertisements, and short texts for the candidate to study in one minute before answering questions related to such stimuli. Normally, this section includes three to four stimuli in total, which together aim to examine the candidate’s overall oral communicative competence across various rhetorical contexts and purposes.

Length and administration: Celpe-Bras is developed and administered by the Brazil Ministry of Education through the National Institute for Educational Studies and Research Anísio Teixeira (INEP). As a large-scale assessment, each of the testing sites around the world has its own coordinator for test administration and management. The total duration of the exam is approximately 200 minutes: 3 hours for the written part and 20 minutes for the oral part.

Scoring procedure: Celpe-Bras certifies four levels of Portuguese proficiency: intermediate, high intermediate, advanced, and high advanced. In order to be certified at all, one needs to reach, both in the written and the oral part, at least the intermediate level. This is because each part on Celpe-Bras receives an independent certification of proficiency level, and when the certifications on the two parts differ, the final certification of one’s proficiency is decided by the lower level reported (INEP, 2017). Specifically, both parts are scored on a scale of 0–5, and Table 1 presents the equivalence between the final scores and the certified levels.

According to the edict published in the official journal of the federal government of Brazil, Diário Oficial da União (DOU) (INEP, 2017), each task in the written part is
evaluated by two raters independently (INEP, 2017). If the scores differ by no more than 1 point, the final score for the task will be the average of the scores assigned by the two raters. In other cases, a third rater will be used. If the third score converges with a score assigned earlier, the final score of the task will be the average between the two convergent scores. On the other hand, if the third score shows an equidistant difference from the two scores assigned earlier, the third score will be used as the final score. Ultimately, the score for the entire written part will be the average of all these task-level scores.

The scoring rubrics for the written part encompass three main dimensions: contextual adequacy, discursive adequacy, and linguistic adequacy. Evaluative criteria for discursive adequacy (clarity and cohesion of the writing) and linguistic adequacy (lexical and grammatical accuracy) remain constant across the rubrics for all the writing tasks. The contextual adequacy dimension, however, involves the format, interlocutor, and purpose, as well as the rhetorical situation specific to each writing task (Schoffen, 2009). Consequently, this part of the rubric is developed ad hoc by the Technical Committee after each administration, based on sample texts produced by test takers.

With regard to the oral part, each candidate is evaluated by an interlocutor and an observer independently. The observer, using an analytic rubric, assigns a score of 0–5 in each of the three weighted analytic dimensions: (a) comprehension, interactional competence, and fluency (50%), (b) lexical and grammatical adequacy (42%), as well as (c) pronunciation (8%). Meanwhile, the interlocutor evaluates the candidate’s performance in the interaction and assigns a single score on a scale of 0–5, using a holistic rubric. If there is no significant discrepancy (1.5 points difference and above) between the two scores, the final score will be the average of the two individual ratings. In the case of significant discrepancy, the oral part will be re-evaluated by two independent raters on the Technical Committee, based on the audio recording of the interaction. Again, one of them will use an analytic rubric and the other a holistic rubric.

**Price:** The test registration fee is charged by the local testing site according to the following guidelines developed by INEP and published in DOU: (a) at the testing sites accredited in Brazil, a maximum of 200 Reais (approximately 54 USD) can be charged; (b) at the testing sites in other countries, for private institutions and/or institutions accredited by the Ministry of Foreign Affairs of Brazil, no more than 100 USD should be charged.
Appraisal of Celpe-Bras

This review adopts Kunnan’s (2018) most recent fairness and justice framework as an evaluative lens. It should be noted that although the theoretical framework indicates a primary focus on fairness and justice, the model itself and the key arguments for fairness and justice still encompass all relevant discussions of test quality and usefulness. According to Kunnan (2018), for example, the Principle of Fairness implies that an assessment ought to provide adequate opportunities for all test takers to learn about and prepare for the test, be consistent, meaningful, and free of bias, and provide proper access and administration, whereas the Principle of Justice requires an assessment to promote positive values to the community and advance justice through public justification and reasoning. In our review, therefore, an evidence-based appraisal of various claims in relation to opportunity-to-learn, consistency and meaningfulness, absence of bias, access, as well as washback and consequences brought about by the development and administration of Celpe-Bras, will be presented.

Claim 1: Celpe-Bras provides adequate opportunity and time for test takers to learn about and practice the exam prior to taking it.

Information about Celpe-Bras is available to the general public at its official website as well as at a federal-university-maintained database of public documents, applied tests, statistics, and studies associated with Celpe-Bras (www.ufrrgs.br/acervocelpebras). Important documents for the preparation of the test include the following: the actual tests administered over the years with videos and audios; the Guide for Participants, which contains a sample test with authentic answers from actual test takers at different proficiency levels and the respective comments given by the Technical Committee; and the Candidates’ Manual, which introduces the process of registration and the structure of the exam. Since all the information can be accessed at any time by the general public, it is convincing that Celpe-Bras offers test takers sufficient opportunities to learn about and prepare for the test prior to taking it.

One potential rebuttal to this claim, however, may be that the information provided is presented in Portuguese only, using language at a fairly advanced level. This could be problematic to some lower-proficiency test takers who would like to prepare for the test at an early stage. Based on our analysis of the linguistic complexity in the Guide for Participants, the Portuguese Flesch index score for the section explaining evaluative criteria, for example, reached 31, indicating that the level of Portuguese would be mostly suitable for readers at high-school and university levels (Tedrus & Steinberger-Elias, 2013). According to Brazil’s Ministry of Education, however, the target test taker population of Celpe-Bras includes all the L2 speakers of Portuguese aged over 16 years who have an educational level equivalent to the Brazilian middle-school level (Ministério da...
Educação, 2015a). Although we understand that the use of complex integrated tasks assumes “a threshold level of language proficiency for examinees to perform on them competently” (Cumming, 2014, p. 222), the relatively high level of language complexity in important manuals and documents may still pose a challenge for some test takers to become adequately prepared for the test.

**Claim 2:** Celpe-Bras is fair to all test takers, as it ensures consistency in its scoring procedure and reliability in its scores.

Official information released by the test developers shows that Celpe-Bras adopts a rather thorough scoring procedure to ensure consistency and reliability in the test scores. As mentioned previously, each task is evaluated by two raters independently. If the scores from the two raters show any discrepancy, more raters will be involved. In total, there are at least 10 raters for the assessment of any candidates’ written and oral performance (Coura-Sobrinho, 2004), which could help minimize the idiosyncrasy in any individual rater’s interpretation of test taker performance. Meanwhile, since different tasks are evaluated by different sets of raters, it may reduce the potential halo effect carried by the same rater(s) from the evaluation of one task to that of another. To ensure consistency in the scoring procedure, especially when so many raters are involved, Celpe-Bras only recruits certified teachers of Portuguese who are experienced and familiar with the exam. Additionally, prior to the actual scoring, all raters are required to participate in a rigorous training and norming session (Coura-Sobrinho, 2004). Such scoring procedures, hence, make it likely that Celpe-Bras could provide rather consistent results, serving as the basis for valid interpretations of test performance. The lack of publically available technical reports containing actual reliability indices, however, makes it difficult for us to draw a conclusive judgment regarding the overall reliability of the test.

In addition, the lack of clear and specific enough descriptors in the rubrics may also pose a potential threat to reliability and validity. Take the writing rubric reported in Schoffen (2009) as an example. Different score levels (levels 4 and 5, for instance) for the “contextual adequacy” dimension often share the exact same one-word-only descriptor (e.g., “adequate”). Such undiscriminating and vague descriptors could make the scoring difficult, hence causing reliability and validity concerns. Moreover, descriptors at a particular level (such as level 1) may even contain all-inclusive evaluative adjectives (e.g., “adequate or partially adequate or inadequate”). It is not clear how a score of 1 point should be interpreted and assigned, based on such descriptions. Likewise, the descriptor for the “lexical and grammatical adequacy” dimension at the high-intermediate level (“Few inadequacies and/or interference of mother language”) seems to be even more demanding than that at the advanced level (“Some inadequacies and/or interference of mother language”), which would certainly lead to reliability and validity concerns. Finally, it is not clear how these five analytic ratings are translated into a final score for any given response. Therefore, even though the test developers claim that rigorous rater training is carried out before actual scoring, it is still unclear how such descriptors are interpreted and applied with consistency and accuracy.

**Claim 3:** Celpe-Bras is meaningful as it measures what it purports to measure.
According to the test developers, Celpe-Bras measures communicative language ability using integrated tasks. Written responses are therefore evaluated on contextual, discursive, and linguistic levels. According to Moutinho (2006) and Schoffen (2009), contextual adequacy serves as the primary focus of evaluation. In other words, if a written response received four points on the contextual adequacy dimension, the final score of this text could not be higher than four points, even if the discursive and linguistic dimensions may receive higher ratings. This focus on contextual adequacy reflects the way in which the test prioritizes the construct of communicative language use ability (Bachman, 1990), which corresponds to the stated construct of the exam (Schlatter, 2014).

Meanwhile, empirical research into other areas, such as test content and topics, seems to provide further validity evidence. For example, Pileggi’s (2017) analysis of writing tasks from one administration showed how the integrated tasks provided various opportunities for test takers to demonstrate their ability to use Portuguese for specific communicative purposes, which contributes to the authenticity and ultimately construct validity of the test. Likewise, Schoffen and Mendel (2018) analyzed 156 writing tasks on Celpe-Bras between 1998 and 2016, and concluded that the genre demand and change over the years on the exam faithfully reflected the change of genre types and demands in the real world. For example, the authors noticed how writing genres such as blog posting and email writing are becoming more recurrent in the exam. Of course, the authors also called for further refinement of the writing tasks to ensure more valid inferences about test takers’ actual ability based on test performance.

With regard to the oral part, according to the Ministry of Education (Ministério da Educação, 2013), it is important that the conversation between the candidate and the examiner should make sense, instead of being a mechanical interaction out of contexts. For this purpose, the examinees are expected not only to answer the questions posed by the examiner, but also to engage meaningfully in the conversation and, when appropriate, initiate new points of discussion that help to move the dialogue along naturally. According to Cândido (2015), the purpose for this part of the exam is to assess test takers’ ability to carry on conversations that resemble authentic daily interactions, which is indeed the intended construct of the exam. Although it seems that the test tasks are designed according to principles of good practice, more data-based evidence from empirical validation work is needed to provide further support for the validity argument.

**Claim 4:** Celpe-Bras is free of bias for different test taker groups of similar ability.

According to Cota (2013), before designing specific tasks, all the members of the Technical Committee will convene to review research on previous editions of the exam in order to address and respond to identified issues when developing the new edition. They are also instructed to avoid pejorative, polemical, or offensive materials and questions, with the aim of eliminating the cognitive and affective sources of construct-irrelevant variance. However, some would argue that the current formation of the Technical Committee may lead to potential bias. As reported in Lima’s (2008) study, the members
on that committee who are responsible for choosing materials and input and developing test tasks are representatives of Brazilian states that are more affluent and politically powerful. Since their perspectives are likely to be influenced by their own political, cultural, and socioeconomic backgrounds, it is argued that the tasks they developed could also be biased toward test takers from a similar background. Of course, INEP is probably aware of this potential source of bias, and therefore stresses that the recruitment of the Technical Committee members is open to all certified Brazilian teachers of Portuguese. Nevertheless, no further evidence detailing the recruitment process and criteria is available. In addition, although the recruitment of only professional teachers of Portuguese to serve on the Technical Committee is understandable, given their expertise and experience, no information is available on how the committee involves other stakeholders in the process of test design and development to ensure the absence of bias. Therefore, more research should look into the certification of the committee members and examine how, if at all, this may lead to potential bias issues in task design, content and input selection, as well as topical knowledge required in the test-taking process. In the meantime, analysis of test performance data, particularly DIF studies, could help to indicate any potential test items or tasks that may function differently for different test taker groups of similar ability.

**Claim 5:** Celpe-Bras provides proper access, administration, and accommodation to test takers.

According to Kunnan (2004), access to an assessment includes educational, financial, geographical, personal, as well as equipment access. In terms of the educational access, analysis in the previous section about opportunity-to-learn shows that test takers overall do have access to information and resources for learning and test preparation. Regarding financial access, however, although INEP provides some rough guidelines on how different registration fees may apply to test takers at different testing sites, no further information is available to clarify how the coordinators at different testing sites make their own decisions on how much to charge the test takers. Thus, despite the fact that the test developers of Celpe-Bras do seem to have considered and attempted to address the fairness issue from a financial point of view, we still could not conclude for sure that the test is actually financially accessible to all the potential test takers.

In terms of geographical access, although testing sites are available in about 37 countries and regions by 2018 (UFRGS, n.d.), Celpe-Bras is admittedly not as widely spread as some large-scale international English tests. In addition to the registration fee, therefore, test takers may need to pay an extra (sometimes a considerable) amount to take the test, as they may need to travel to another city, or even another country, to reach a testing site. In Asia, for example, only five testing sites are available in Korea, China, and Japan (INEP, 2017). Hence, all the applicants in Asia need to travel to, if they are not already living in, these five cities to take the exam. The geographical locations of these testing sites also seem to be in rather expensive metropolises. In China, for instance, the only two testing sites are located in Beijing and Macau. The limited geographical access to Celpe-Bras for now could, therefore, pose a challenge to test takers from a less affluent family/region.
Regarding personal access, Celpe-Bras takes care to provide appropriate accommodations for test takers who may have special test-taking requirements. According to INEP (2017), any test taker with low or monocular vision, blindness, deafness, hearing impairment, physical disability, autism, attention deficit, intellectual disability or dyslexia, or any test taker who is elderly, pregnant or has to breastfeed or lactate, can request a special accommodation at the time of registration with relevant proof and declaration. The accommodation options that the coordinator of the testing site can offer include: braille test paper, test paper with enlarged or super enlarged font, easy access room, desk and chair without armrest, as well as desks specifically designed for wheelchair users. In the case of candidates with dyslexia or attention deficit, they can apply for an extra 60 minutes for the written exam. The Candidates’ Manual (Ministério da Educação, 2015a) also explains in detail how to apply for a special accommodation.

In addition, Celpe-Bras ensures proper and fair administrative procedures. The responsibilities of INEP and the coordinator of each testing site are well specified in the 44-page-long Orientation Manual for Coordinators of Celpe-Bras (Ministério da Educação, 2015b), including the detailed procedures of conducting the exam, and what ought to be done before and after the exam. For example, the coordinators are required, in general, to be responsible for maintaining and updating data at their testing sites, verifying the information provided by applicants during registration, approving their applications, and verifying the special condition described by any applicant as potentially relevant for considerations of special accommodation, in addition to other more specific requirements for the management of the testing site and administration. Hence, all the responsibilities and administrative procedures are standardized to a large extent. Meanwhile, Celpe-Bras is administered with a paper-and-pencil format, which demands no special equipment familiarity on the test takers’ part. Given such observations, it is reasonable to conclude that Celpe-Bras provides proper administration that is fair to all test takers.

**Claim 6:** Celpe-Bras is beneficial to the community and society, and the testing institution promotes positive values and advances justice.

Since Celpe-Bras is an exam of communicative language ability with integrated performance-based tasks only, the most obvious washback of the exam is on the method of teaching and learning Portuguese as a foreign language. Moutinho (2006), for example, interviewed four candidates taking Celpe-Bras at three different stages (before the first contact with the exam, during the preparation period, and after taking the exam) and found that the strong belief that grammar knowledge should be at the center of language teaching and learning had been significantly weakened, if not completely eradicated, after taking the exam, and that the test takers reported their willingness to learn Portuguese through interactions with native speakers in authentic communicative situations instead of focusing primarily on studying grammar. Such results provided evidence that Celpe-Bras could have a positive washback on learning. Likewise, Almeida (2012) observed that the exam influences and guides the way in which future language teachers conceptualize language teaching method and curricular design. The test tasks on Celpe-Bras prompt Portuguese language teachers to focus
more on the development of communicative competence than on the traditional structuralized teaching of grammar rules.

In addition, results of Celpe-Bras are published in the DOU and on the website of INEP about two months after each administration. Candidates’ names will be listed accordingly under the four certified levels for their reference. According to INEP (2015), the test takers can make an appeal via email within five working days after the publication of test results, if they are dissatisfied with or have questions about the results. The Technical Committee will reanalyze their performance upon the receipt of test takers’ appeals, and the final decision will be issued within two weeks. The availability of this appeal process itself reflects that the testing institution cares about test takers’ concerns, and is willing to promote positive values and ensure fairness and justice through public justification and reasoning.

**Summary and discussion**

Overall, our review shows that Celpe-Bras test developers do take great care in their design, development, and administration of the test to address core issues of validity, reliability, and authenticity. In addition, a rather transparent mechanism is adopted, which certainly helps promote fairness and justice in the implementation and use of this large-scale, high-stakes language test. However, it would help reviewers and test users to make more confident and informed evaluative judgments about the overall quality of the test, if technical reports or more validation research results could be made available to the public.

There are also areas, based on our current review, that need further improvement to enhance the overall test usefulness. For example, the current test results released by the test developers only shows the final certified proficiency level without specifying scores for each individual part or task. As a result, test takers who receive the final report would not know how to improve their Portuguese language use or whether it is worthwhile to make an appeal. It would be more meaningful and useful to the test takers, therefore, if the score report could include component scores for individual parts and tasks, and provide more discursive evaluations of the test takers’ language competence in the three main areas of the communicative language ability as defined by the rubric (i.e., contextual, discursive, and linguistic adequacy). Moreover, as discussed earlier in our appraisal of the test, relevant research is still in need for different stakeholders to evaluate the exam in more depth, particularly the potential impact of the following factors based on actual test performance data: (1) the complexity of language in the informational handbooks and manuals; (2) inter- and intra-rater consistency in the application of rubrics; (3) the formation of the Technical Committee and involvement of stakeholders; and (4) tasks that may function differently for different test taker groups of similar ability.

As a large-scale, high-stakes international language proficiency test, Celpe-Bras is different from other such large-scale language assessments currently in use. Although almost all the current large-scale language tests claim that they intend to assess the communicative competence of language learners, relatively few have adopted a fully integrated performance-based assessment design as Celpe-Bras does. Oftentimes, receptive language knowledge is still assessed using discrete-point test items. Although assessing
receptive language knowledge in isolation could certainly provide information about L2 learners’ language proficiency, especially at the lower level for beginning learners, integrating the assessment of receptive language knowledge into that of meaningful and contextualized communicative interactions may provide more relevant, authentic, meaningful, and hence useful information for all the stakeholders involved.

Of course, some would argue that large-scale, high-stakes assessments would inevitably face various constraints in test design owing to considerations of practicality and logistics; consequently, more selective and/or short-response items, which could be machine scored, would dominate many of the language tests. However, over-prioritizing logistic considerations and practicality at the expense of validity and authenticity may just be a form of unjust testing practice. Additionally, if the purpose of language teaching and learning is ultimately developing learners’ ability to use the language for actual communication, then it sounds only logical that the assessments also focus on the actual use in authentic communicative scenarios. It is to be hoped that our review of the Portuguese test here, with its unique test design, could open up more discussions of whether and how such a design should or could be implemented in large-scale language assessments in our future practice.

Acknowledgements
The authors would like to thank Mr Júlio Reis Jatobá and Professor Ricardo Moutinho for directing us to the relevant information on Celpe-Bras and their support of this project.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Cecilia Guanfang Zhao https://orcid.org/0000-0003-2444-0994

References


