

SECTION ARTICLES

A study of the dimensionality of assessment scales for oral proficiency for the Certificate of Proficiency in Portuguese as a Foreign Language*

Laura Márcia Luiza Ferreira¹
ORCID: 0000-0001-7632-0834

Abstract

By using Celp-Bras assessment scales, a construct of verbal proficiency in Portuguese for foreign-language speakers is operationalized and measured. The oral Celpe-Bras's score model is organized in seven items distributed in two scales through which the assessors – the interviewer and the observer – rate six each of the following items: comprehension, interactional competence, fluency, lexical adequacy, grammatical adequacy, and pronunciation. In order to analyze the dimensions of the scales, evidences to discuss the oral scale's dimensionality, I perform the exploratory factorial analysis of a set of scores obtained by 1,000 participants who sat for the exam in its first edition in 2016. R^2 was 0.9617 and the Tucker Lewis Index (TLI) was 0.896. Only one factor explained the variables because the loading values ranged from 0.65 to 0.94. The measure was found unidimensional. According to the communality values, only comprehension was slightly below 0.5, indicating the need for further investigation. The weight values of each item were, in decreasing order: interviewer's score 0.36, lexical adequacy 0.19, fluency 0.18, grammatical adequacy 0.13, interactional competence 0.09, pronunciation 0.06, and comprehension 0.04. Based on factorial analyses, I discuss a proposal for the composition of the individual scores in the oral test well as the implications of changing the weight of the items in the new proposal to rank participants on a per certification range basis.

Keywords

Dimensionality – Validity – Assessment of an additional language – Oral proficiency – Factorial analysis.

* Translated by Glaucia Roberta Rocha Fernandes and Martin Clowes

1 - Universidade Federal da Integração Latino-Americana (Unila), Foz do Iguaçu, Paraná, Brasil.

Contato: laura.ferreira@unila.edu.br



DOI: <http://dx.doi.org/10.1590/S1678-4634201945202512>
This content is licensed under a Creative Commons attribution-type BY-NC.

Introduction

The Certificate of Proficiency in Portuguese as a Foreign Language, hereinafter Celpe-Bras, is the official examination of the Brazilian Government for proof of proficiency of foreigners in the Portuguese language. The examination is currently supervised by the Anísio Teixeira National Institute of Educational Study and Research (Inep). Certification of proficiency attested by means of the Celpe-Bras may be required of foreigners under certain circumstances, such as application for educational cooperation programs financed by the Brazilian Government and revalidation processes of diplomas, depending on the requirements of professional councils.

By means of the Celpe-Bras, it is possible to obtain certification of proficiency in the levels intermediate, upper intermediate, advanced and upper advanced after completion of a single examination. The Celpe-Bras consists of two sections: the written examination and the oral examination. The former consists of four open-ended items, i.e., the drafting of four texts, which are evaluated using a holistic scale that generates a score. The final score of the written section is calculated from the simple arithmetic mean for four items. The score for the oral examination is composed in a slightly more complex manner. The lower score of the two examination phases is the value for purposes of certification of proficiency (BRASIL, 2016a).

The oral examination consists of seven items organised into two assessment scales. After a face-to-face oral interaction with a duration of 20 minutes between the interlocutor-rater and the candidate, the score is given by two raters in the places where the interactions take place. Both the interlocutor-rater and the observer-rater give independent scores for the candidate's oral performance. The interlocutor-rater assigns a single score from a holistic matrix (Figure 1) and the observer-rater uses an analytical matrix (Figure 2), in which the score is composed of six assessment items: oral comprehension, interactional competence, fluency, lexical adequacy, grammatical adequacy and pronunciation. In other words, in total, the score of the oral examination consists of seven items that compose the two assessment matrices: the observer's and the interlocutor's.

In the Candidate's Manual (BRASIL, 2010), the oral section is divided into two phases: in the first, the interaction is based on information from the candidate's enrolment questionnaire, with a duration of five minutes, and the second is based on three trigger elements selected by the interlocutor-rater. In the second phase, the interaction lasts 15 minutes and is divided into three parts, with five minutes dedicated to each of the three trigger elements, which are mostly clippings from news reports circulated in printed media.

The assessment is audio-recorded and sent to Inep. According to the registration public notice (BRASIL, 2016a) for the examination, the final score of the oral examination is calculated from an arithmetic mean between the scores of the interlocutor-rater and the observer-rater, i.e., each score has a weighting of 50% in the composition of the final score of the oral examination. If the scores awarded by the raters diverge by more than one and a half points (1.5), the interaction is reassessed in examination correction events. In addition, the oral examination may be reassessed by a third rater if: the result diverges by up to two points in relation to the score of the written examination; the difference in

scores between the two examination types involves a change in the certification level; or the final score in the written examination is superior to that of the oral examination.

In performance examinations, as is the case of the Celpe-Bras, the score is attributed by the raters, who use scales to guide the judgement of the candidate's performance. Eckes (2015) explains that, in the face-to-face interaction, there are at least five facets² and various forms of interaction between them, which may impact on the final result, namely: the candidate, the task, the scale or model for assigning the score and the rater.

This study discusses the model for assigning scores of the Celpe-Bras oral examination with the aim of evaluating the unidimensionality of the measure composed of the seven items of the assessment scales for the oral examination. To this end, the results of an exploratory factorial analysis are presented.

This will be followed by a brief presentation of the concept of 'construct' in the context of oral assessments of foreign language.

Construct in oral assessment of additional languages

According to Bygate (2009), theories on methodology for teaching foreign languages fell short in the attempt to draw up a construct on the development of orality. In making a brief retrospective, the author points out that, in structuralist approaches, the focus was on grammar. Although speech occupied a central position in the audio-lingual method, orality was seen more as a means of acquiring foreign languages rather than as an end in itself. The author also argues that, even in the most diverse ways of looking at the communicative approach to language teaching, orality is still seen more as a means rather than a goal to be set and achieved.

In the context of assessment studies, such goals approximate the notion of a construct. A construct is something that can be observed and measured. By seeking to operationalise the nature of the construct of orality, examinations of oral proficiency, such as the Celpe-Bras, provide highly useful guidance for the discussion of central issues in orality, such as the constructs of speech, task, performance criterion, and development of speech (BYGATE, 2009). According to the author, the tests are valuable analytical tools for discussing the comprehension of the possible parameters of oral proficiency.

Fulcher (2003) emphasizes that, in the case of oral proficiency being a construct to be measured and observed, it is necessary to associate it with something that can be observed and measured. Fulcher (2003), like Bygate, points out the problem that there is no ready and efficient construct for oral proficiency in a foreign language, and argues that there cannot be a consensus on it between theorists and teachers. The definition of a construct, according to Fulcher (2003), is a matter of choosing some theories and trying to operationalise them in an assessment context with their specific purposes, providing a theoretical and empirical basis for the choices made. McNamara (2004) also stresses that constructs will always be controversial and targets of criticism, so they must be linked with arguments that defend the validity of the examination.

2- According to Eckes (2015), facets are synonymous with factors, variables or components that are part of the assessment situation and that affect the scores in a systematic way.

Figure 1- Observer-rater's assessment matrix



GRADE DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE
OBSERVADOR



Ministério da
Educação




PRONÚNCIA *	ADEQUAÇÃO GRAMATICAL	ADEQUAÇÃO LEXICAL	FLUÊNCIA	COMPETÊNCIA INTERACIONAL	COMPREENSÃO	
Pronúncia (sons, ritmo e entonação) adequada .	Uso de variedade ampla de estruturas. Raras inadequações na utilização de estruturas.	Vocabulário amplo e adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Raras interferências de outras línguas.	Pausas e hesitações para organização do pensamento e, eventualmente , para resolver algum problema de construção linguística, sem interrupções no fluxo da conversa.	Apresenta muita desenvoltura e autonomia , contribuindo muito para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Compreensão do fluxo natural da fala. Rara necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala .	5
Pronúncia (sons, ritmo e entonação) com algumas inadequações e/ou interferências de outras línguas .	Uso de variedade ampla de estruturas. Poucas inadequações na utilização de estruturas complexas e raras inadequações no uso de estruturas básicas.	Vocabulário amplo e adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Poucas interferências de outras línguas.	Pausas e hesitações para organização do pensamento e, eventualmente , para resolver algum problema de construção linguística, com poucas interrupções no fluxo da conversa.	Apresenta desenvoltura e autonomia . Não se limita a respostas breves, contribuindo para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Compreensão do fluxo natural da fala. Alguma necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala .	4
Pronúncia (sons, ritmo e entonação) com inadequações e/ou interferências de outras línguas .	Uso de variedade de estruturas. Algumas inadequações na utilização de estruturas complexas e poucas inadequações no uso de estruturas básicas.	Vocabulário adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Algumas interferências de outras línguas, com ocasional comprometimento da interação.	Pausas e hesitações para organização do pensamento e, algumas vezes, para resolver algum problema de construção linguística, com algumas interrupções no fluxo da conversa.	Não se limita a respostas breves , contribuindo para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Alguns problemas na compreensão do fluxo natural da fala. Necessidade de repetição e/ou reestruturação ocasionada por palavras de uso frequente, em ritmo normal da fala .	3
Pronúncia (sons, ritmo e entonação) com inadequações e/ou interferências frequentes de outras línguas .	Uso de variedade limitada de estruturas. Inadequações mais frequentes tanto na utilização de estruturas complexas quanto nas básicas.	Vocabulário adequado para a discussão de tópicos do cotidiano e para expressar ideias e opiniões sobre assuntos variados. Algumas interferências, da língua materna, ocasionando algum comprometimento da interação.	Pausas e hesitações para organização do pensamento e, mais frequentemente , para resolver algum problema de construção linguística, com interrupções no fluxo da conversa.	Pode se limitar a respostas breves , mas contribui para o desenvolvimento da conversa. Mesmo quando necessário, faz pouco uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Alguns problemas na compreensão do fluxo natural da fala. Necessidade frequente de repetição e/ou reestruturação ocasionada por palavras de uso frequente, em ritmo normal da fala .	2
Pronúncia (sons, ritmo e entonação) inadequada e/ou interferências acentuadas de outras línguas .	Uso de variedade limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas.	Vocabulário inadequado e/ou limitado para a discussão de tópicos do cotidiano e para expressar ideias e opiniões sobre assuntos variados. Muitas interferências de outras línguas, ocasionando frequente comprometimento da interação.	Pausas e hesitações frequentes exigem um grande esforço do interlocutor , ou alternância no fluxo da fala entre língua portuguesa e outra língua.	Limita-se a respostas breves , contribuindo pouco para o desenvolvimento da conversa. Mesmo quando necessário, faz pouco uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Muitos problemas na compreensão do fluxo natural da fala. Necessidade muito frequente de repetição e/ou reestruturação ocasionada por palavras básicas, em ritmo normal da fala .	1
Pronúncia (sons, ritmo e entonação) inadequada e/ou interferências muito acentuadas de outras línguas .	Uso de variedade bastante limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas, comprometendo a interação.	Vocabulário muito inadequado e/ou limitado para a discussão de tópicos do cotidiano e para expressar ideias e opiniões sobre assuntos variados. Muitas interferências e outras línguas, comprometendo a interação.	Pausas e hesitações muito frequentes interrompem o fluxo da conversa, ou fluxo de fala em outra língua.	Limita-se a respostas breves, raramente contribuindo para o desenvolvimento da conversa, que fica totalmente dependente do avaliador. Mesmo quando necessário, não faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Problemas sérios na compreensão do fluxo natural da fala. Necessidade constante de repetição e/ou reestruturação, mesmo em situação de fala simplificada e muito pausada .	0




* Não se espera uma fala sem sotaque nem mesmo nos níveis mais altos de certificação.



Figura 2- Holistic matrix or interlocutor-rater's assessment



**FICHA DE AVALIAÇÃO DA INTERAÇÃO
FACE A FACE | ENTREVISTADOR**






Nome do examinando: XXXXXXXXXXXXXXXXXXXX XX XXXXXXXXXXXXXXXXXXXX XX XXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXX

Nacionalidade: XXXXXXXXXXXXXXXXXXXXXXXXXXXX Número de inscrição: 999999999

Documento de identificação n.º: XXXXXXXXXXXXXXXXXXXX


Posto aplicador XXX



4360144035

Elementos provocadores utilizados

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



Preencha os círculos totalmente e com nitidez, utilizando caneta esferográfica de tinta azul ou preta.

Avaliação do Entrevistador

Marque o número da descrição que melhor caracteriza o desempenho do examinando



5	<input type="radio"/>	Demonstra autonomia e desenvoltura, contribuindo bastante para o desenvolvimento da interação. Apresenta fluência e variedade ampla de vocabulário e de estruturas, com raras inadequações. Sua pronúncia é adequada e demonstra compreensão do fluxo natural da fala.
4	<input type="radio"/>	Demonstra autonomia e desenvoltura, contribuindo para o desenvolvimento da interação. Apresenta fluência e variedade ampla de vocabulário e de estruturas, com inadequações ocasionais na comunicação. Sua pronúncia pode apresentar algumas inadequações. Demonstra compreensão do fluxo natural da fala.
3	<input type="radio"/>	Contribui para o desenvolvimento da interação. Apresenta fluência, mas também algumas inadequações de vocabulário, estruturas e/ou pronúncia. Demonstra compreensão do fluxo natural da fala.
2	<input type="radio"/>	Contribui para o desenvolvimento da interação. Apresenta poucas hesitações, com algumas interrupções no fluxo da conversa. Apresenta inadequações de vocabulário, estruturas e/ou pronúncia. Pode demonstrar alguns problemas de compreensão do fluxo da fala.
1	<input type="radio"/>	Contribui pouco para o desenvolvimento da interação. Apresenta muitas pausas e hesitações, ocasionando interrupções no fluxo da conversa, ou apresenta alternância no fluxo de fala entre língua portuguesa e outra língua. Apresenta muitas limitações e/ou inadequações de vocabulário, estruturas e/ou pronúncia. Demonstra problemas de compreensão do fluxo natural da fala.
0	<input type="radio"/>	Raramente contribui para o desenvolvimento da interação. Apresenta pausas e hesitações muito frequentes que interrompem o fluxo da conversa, ou apresenta fluxo de fala em outra língua. Apresenta muitas limitações e/ou inadequações de vocabulário, estruturas e/ou pronúncia, que comprometem a comunicação. Demonstra problemas de compreensão de fala simplificada e pausada.

ENTREVISTADOR

POSTO APLICADOR

DATA ____/____/____

Rubrica

Source: INEP, 2016.

Constructs can be based on theories or can be prepared from a composition of concepts. Messick (1987) questions the fact that the constructs are not based on a theory, but on a composition of concepts. For the author, it is possible to investigate the efficiency of the composition by verifying the extent to which the measures are assessing the same construct. In the case of the Celpe-Bras, as there are two different matrices evaluating the same thing, and various items that reflect various theoretical concepts about development of orality in a foreign language, it is necessary to investigate how the scores awarded by means of the holistic matrix and the analytical matrix are related. Messick (1987) states that the representation of the construct refers to the relative dependence on the test design, i.e., the scales for the items assessed. Regarding the Celpe-Bras oral examination, for example, the representation of the construct is relatively dependent on the score given by the interlocutor-rater, the score for oral comprehension, the score for pronunciation, etc.

Fulcher and Davidson (2007), when discussing the system for assigning scores in a language examination, state that the rater's judgement process is what connects the evidence of performance, which can be represented by the score, to the task and the construct. To obtain inferences, it is necessary to collect evidence that may be related to the scores. According to Messick (1987), however, in the empirical approaches to designing a test, the items should be part of its composition after the analyses of the data, whether they are internal data to the examination, demonstrating the homogeneity of the item or its factor loadings, or external data, involving the study of the correlation of the assessment parameter or the correlation of discrimination of the criterion with respect to a number of other parameters. According to Messick (1987), these analyses refer to the substantial aspect of construct validity.

In a publication of the American Educational Research Association (AERA, 2014), a substantial aspect is understood to be the evidence based on the internal structure of the examination. It is worth noting that both Messick (1987) and Standards (AERA, 2014) indicate factorial analysis as a way of analysing the unidimensionality of the measure, i.e., whether the items are measuring the same construct. For example, how much each assessment item exemplifies the construct being measured is a matter of substantial aspect of construct validity, i.e., an evidence of validity based on the internal structure of the examination.

Messick (1987) points out that factorial analysis is advisable for evaluating the relationships between assessment items via the analysis of factor loadings. Factor loadings are values that make it possible to evaluate how each assessment item composes the final score of an assessment. Messick (1987) states that factorial analysis is advisable when the aim is to combine the evaluation of theories and the construction of scales for interpreting the consistency of responses.

This work discusses the factorial analysis methodology in more detail below.

Factorial analysis and assessment in foreign languages

Factorial analysis can be used to investigate the validity of the construct, to check theoretical hypotheses and to summarise or collate a large volume of data. In the field

of statistics, the noun validity followed the term factorial. Thompson (2004) revisits Nunnally's text of 1978 and asserts that the historic term for "construct validity" is "factorial validity". Thompson (2004) suggests that, when we are developing specification documents related to a measure, such as scales for assessing and grading descriptors for levels of proficiency, factorial analysis should be used to examine the validity of the score. Thompson (2004) explains that, if the researcher aims to answer questions related to what the test measures, the answer should be given in terms of factorials.

Brown (2015) is also an enthusiast for the use of the factorial analysis methodology to verify the validity of a construct in research in social and behavioural sciences. According to the author, the analysis can provide empirical evidence about convergent and discriminant validity in relation to the theoretical constructs. For Brown (2015), a factorial analysis can reveal empirical evidence of a strong interrelationship between assessed items that are similar or overlapping from a theoretical point of view or a weak interrelationship when they are part of distinct theoretical constructs.

In regards to the assessment methodology of the Celpe-Bras examination, for example, a factorial analysis can point out which items of the analytical matrix are more strongly interrelated. The more tightly interrelated the parameters are, the greater the evidence that the assessment is being made from the same theoretical construct (that of oral proficiency in the case of the present study). Thompson (2004) points out that, although the terms related to the validity concept do not include factorial validity, factorial analysis continues to be a useful tool in the construction of issues relating to the validity.

Factorial analysis can be the starting point for many research works that analyse scores assigned to describe some kind of performance (FULCHER, 2003; BROWN, 2015). Fulcher (2003) exemplified the methodology citing the work of Hinofotis, 1983, in which he analysed 12 items to assess the communication of assistant professors at the University of California with their students in situations of oral interaction in the classroom.

By means of factorial analysis, Hinofotis (1983 apud FULCHER, 2003) investigated the relationship between vocabulary, grammar, pronunciation, flow of speech, eye contact, non-verbal aspects, confidence in manner, presence, *development of explanation*, *use of supporting evidence*, *clarity of expression* and *ability to relate to students*. The researcher started from the assumption that the items could be grouped into five factors and, after interpreting the data, he concluded that factor 1 (communication and information) is strongly influenced by the items: *development of explanation*, *use of supporting evidence*, *clarity of expression* and *ability to relate to students*; factor 2 (expression) is impacted by *flow of speech* and *ability to relate to students*; factor 3 (non-verbal aspects) by *non-verbal aspects* and *ability to relate to students*; factor 4 (language proficiency) by *vocabulary* and *grammar*, and factor 5 (pronunciation) only by *pronunciation*. Fulcher (2003) draws attention to the fact that pronunciation was not empirically part of the factor regarding the construct of linguistic proficiency, and also to the fact that the criterion *ability to relate to students* is present for two factors: 'communication and information' and 'expression'.

In the end, Fulcher (2003) concludes about the factorial analysis method that, if the researcher's argument, based on analysis, is plausible, then it will succeed in presenting evidence to substantiate an inference about the meaningfulness of the score.

In the field of foreign-language assessment, Kunnan (1992) also used the methodology to analyse the meaningfulness of the score of a placement examination at the University of California (UCLA) and used, among other methods, exploratory factorial analysis for four groups of learners of English as a second language to investigate the validity of a tool that separately assesses the skills of reading, listening and grammar. At the end of the study, Kunnan (1992) concluded from the factorial analysis that students with low proficiency tend to have low scores in different skills, since the factorial loading of this group influences a single factor, while more proficient students may have a variation in the mastery of the skills of reading, listening and use of grammar, because the factorial loading is distributed. Based on the analyses, the author suggests that the score by skill, i.e., separated by test section – reading, listening or grammar – should be used for placing students in different levels of study, and not the total score, as was usually done in the language studies program analysed by the author.

In addition to being useful to evaluate the scores of an assessment instrument that is already established, Brown (2015) affirms that factorial analysis is a popular tool for the development and construction of assessment scales. By calculating factor loadings, it is possible to define how scores should be assigned, whether by means of an assessment instrument that considers a set of dichotomous items which relate to different language skills, as in the case of the placement examination studied by Kunnan (1992), or polytomous items, as in the proposal of Hinofotis (1986 apud FULCHER, 2003), which is similar to the scales of the Celpe-Bras oral examination. According to Brown (2015), factorial analysis can be used to check, for example, the number of assessment items of the oral examination that are related to the factor, i.e., the dimension of the oral proficiency construct, and the patterns of each of the items in relation to the factor(s) or dimensions of oral proficiency. In the context of this study, factorial analysis will provide elements for discussing the relationship between the score and the construct, comprised of the items of the Celpe-Bras oral examination, and for proposing a reformulation of the weightings that compose the final oral score.

The analysis and discussion of the results are presented below.

Analysis and discussion

The data correspond to the assessment of the oral performance of 1000 candidates who took the examination in the first semester of 2016. The data set analysed has seven variables: six scores relating to the six items assessed in the observer's matrix and a total score, called the interlocutor's score. In other words, the analysis used data relating to the scores of the six items that make up the analytical matrix and to the interlocutor's score. The observer's final score and the final examination score were not considered in the calculations presented below.

The analysis was carried out in several stages so as to identify how the aspects assessed contribute to the composition of the candidate's oral score. The statistical software R (R CORE TEAM, 2018), version 3.5.0 of May 23, 2018, for Windows 10 was used for the calculations. R is a free software program that can be used for various

statistical calculations. The Psych package, version 1.8.4 (REVELLE, 2018), was used for the factorial analysis.

The confidence intervals for the weightings and factor loadings were calculated by bootstrapping (DAVISON; HINKLEY, 1997; CANTY; RIPLEY, 2017). Estimation by bootstrapping makes use of concepts of the central limit theorem. Regardless of the data distribution form, the sample distribution of the parameters of interest manages to assume a normal distribution. The Celpe-Bras oral scores are asymmetrical, but the application of theorem concepts can guarantee accurate results for the calculated values, regardless of the way the data are presented. The adoption of estimation by resampling or bootstrapping is necessary to ensure the correct application of the central limit theorem to the data. This method consists of successive samplings of the available data and calculation of the values of interest. After successive samplings, the final value will be the arithmetic mean of the calculated values. In this case, 10,000 samples with replacement of size 1000 were taken from the data in this study.

The sample is composed of high scores (Table 1). The data in the table refer to the measurements. It can be seen that 2.5% of the examinees obtained scores of up to 1.8373 and 75% of the examinees obtained scores of up to 4.29 on a scale of zero to five points. More than half of the sample refers to scores greater than or equal to 3.855. The distribution of scores higher than this value is concentrated on values greater than 4.5. Of the group of scores less than 3.85, more than 25% are around 3.25 and only 5% represent scores of 2.17, half of which are 1.709 or below, i.e., there are very few scores of 1 in the sample. It is worth noting that, in the first semester of 2016, 6222 examinees enrolled in the examination. The research sample represents 16.07% of the total number of examinees enrolled.

Table 1 – Cumulative standard normal distribution of final scores of the oral section

	Observer's score	Interviewer's score	Final score for the oral section
0%	0.250	0	0.4000
2.5%	1,709	2	1.8373
5%	2.170	2	2.0900
25%	3,250	3	3.1500
50%	3,855	4	3.9300
75%	4,500	4	4.2900
95%	5,000	5	5.0000
97.5%	5,000	5	5.0000
100%	5,000	5	5.0000

Source: prepared by the author based on data from Inep.

As the aim was to understand how the six variables of the analytical matrix and the interlocutor-rater’s score comprise the oral proficiency factor in practice, and with the use of the matrix by the raters, an exploratory factorial analysis was performed. All the scores allocated to the seven items were taken into account in the calculation based on the factorial analysis of the principal axes (principal axis factoring, PAF). In this analysis, it was concluded that the seven variables can be represented by just one factor. It was not necessary to rotate the factors, because the analysis was reduced to one factor. The hypotheses about the factorial structure of scores being organised into one or two factors were tested. It is worth noting that preliminary analyses were performed using structural equations and confirmatory analysis, but convergence problems were found.

From the factorial analysis, it is assumed that the oral proficiency factor is being explained by seven variables, which would be the six items that make up the analytical score plus the interviewer’s score, the seventh variable. To assess the local adjustment of the model, the coefficient of determination (R^2) was analysed, which refers to the percentage of variation of the variables – the scores from the oral test – that are being explained by the calculated factorial structure. The factorial structure presented had an R^2 value of 0.9617319. This shows that the data suit the analysis model and can be explained by it. To assess the adjustment of the model, the calculation of the RMSEA (root-mean-square error of approximation) is presented, the value of which was 0.18. A value of around 0.5 is suggested to be a good adjustment index. In the case analysed, a possible hypothesis for the high value of the RMSEA index may be the fact that the scores are very strongly correlated. The Tucker–Lewis index (TLI) is another way to assess the reliability of the results calculated by the analytical model. In this study, the value was 0,896, which is a satisfactory result suggesting that the data can be explained by the analytical method used.

In the analysis, the current weighting used to calculate the final score of the oral examination was disregarded. When recalculating the values to arrive at the variables that explain the factor, or construct of oral proficiency, the interviewer’s score was the most important variable.

Table 2 – Characteristics and weightings to be assigned to each analytical variable and interviewer’s score

	Load	Weighting	Communality
Oral comprehension	0.6572906	0.0455407	0.43
Interactional competence	0.8028341	0.0950375	0.64
Fluency	0.8816178	0.1836518	0.78
Lexical adequacy	0.9092421	0.1946057	0.83
Grammatical adequacy	0.8852773	0.1350449	0.78
Pronunciation	0.8004653	0.0644487	0.64
Interviewer’s score	0.9481050	0.3644149	0.90

Source: prepared by the author based on data from Inep.

From the loading values presented in Table 2, it can be stated that the variables are related or that they explain a single factor. By analysing the values for factor loading, it is suggested that a single factor is influencing the values of the scores allocated to items, and therefore we can say that the measure is unidimensional. Asserting that the measure is unidimensional, in our context, is the same as saying that the scores are related to a single thing, i.e., oral proficiency.

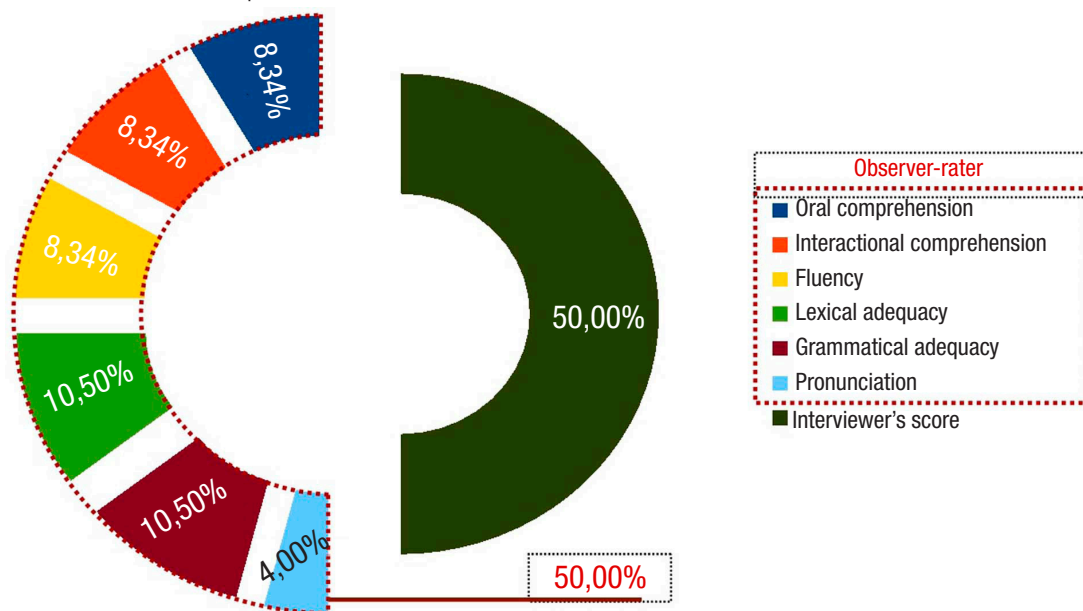
According to Kim and Mueller (1978), the calculation of the communality is based on the correlation of each variable with the rest of the set of variables, i.e., it tries to quantify how the score for oral comprehension, for example, is correlated with all the other scores. Figueiredo Filho and Silva Júnior (2010) explain that it is on the basis of the communality value that we can infer that one variable is linearly correlated with the others. The authors claim that low communality values (less than 0.50) mean that they might not be linearly correlated. With regard to the communality values in Table 2, we have a slightly low value for the oral comprehension score, suggesting that the item might be less related to the others.

Items that have higher factor loadings in the final score of the oral examination with their respective confidence intervals are, from largest to smallest: interviewer's score 0.95 (0.94–0.96); lexical adequacy 0.91 (0.90–0.92); fluency 0.88 (0.86–0.90); grammatical adequacy 0.88 (0.87–0.89); interactional competence 0.80 (0.77–0.83); pronunciation 0.80 (0.77–0.82) and comprehension 0.65 (0.60–0.69). The interlocutor's score is the item that most explains the score in the oral examination when comparing this variable separately with the others. Interactional competence and pronunciation are variables that contribute approximately equally to the oral proficiency factor, as do fluency and grammatical adequacy, by having approximate loading values. The values of 0.95 for the item interviewer's score and 0.90 for lexical adequacy stand out.

The weighting values represent how much each aspect contributes to the composition of the final score of the oral assessment. The weighting values for each of the items with their respective confidence intervals are, from largest to smallest: interviewer's score 0.36 (0.33–0.42); lexical adequacy 0.19 (0.15–0.22); fluency 0.18 (0.15–0.22), grammatical adequacy 0.13 (0.05–0.15); interactional competence 0.09 (0.07–0.11); pronunciation 0.06 (0.04–0.08) and oral comprehension 0.04 (0.03–0.05).

The weighting values calculated by the analysis are approximate and, on summing them, they would come to an approximate value of 105. The values were therefore recalculated to fit the metric of 100% (as in Charts 1 and 2, below). In the charts, the weightings used in the composition of the score and the approximate weightings proposed are compared, based on the factorial analysis for the composition of the final score from the recalculation.

Chart 1 – Current composition of the oral examination



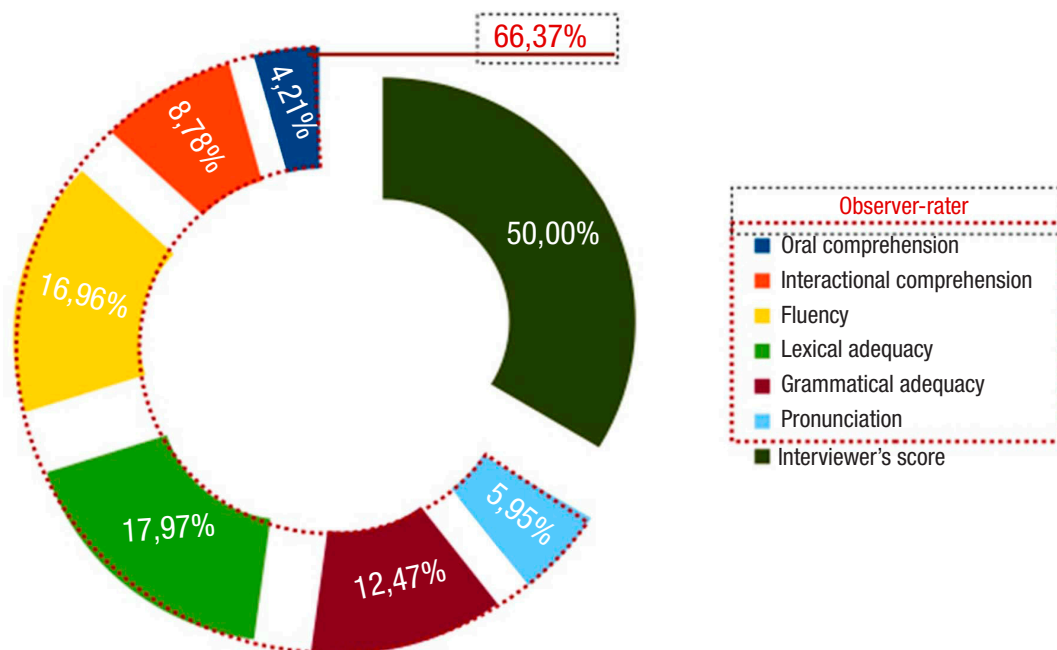
Source: prepared by the author based on data from Inep

Considering that the composition of the analytical items would have a weighting of 50% in the composition of the final score of the oral examination, Chart 1 shows how the score of the oral examination is currently calculated. In Chart 2, the weighting of 50% for the interlocutor-rater’s score is disregarded, and the values of the six parameters of the analytical matrix and the interlocutor-rater’s score form the seven variables in the composition of the final score for the oral examination, i.e., the interlocutor-rater’s score is considered as a variable without a fixed weighting value. It is worth mentioning that the item that had its weighting decreased most proportionally was that of oral comprehension. In general, the other items transferred one third of their weighting to the interlocutor’s score.

With regard to the weighting of the set of parameters that make up the observer-rater’s score and the interlocutor-rater’s single score, even though the interlocutor-rater’s score – with a value of 33.67% – is the item that most explains the score of the oral examination, the observer-rater’s score, i.e., the composition between the six other scores, is what most explains the final oral score. Adding together the weighting of the six items that make up the analytical score, 66.34% of the final score of the oral examination is explained by the sum of the weightings for oral comprehension, lexical competence, fluency, lexical adequacy, grammatical adequacy and pronunciation. Put another way, the observer’s score is more important than the interlocutor’s score, because when the

weightings of the analytical items are summed in the composition of the final score, more than 50% of the composition of the final score is explained by the observer-rater's score.

Chart 2 – Estimated composition of the oral examination score

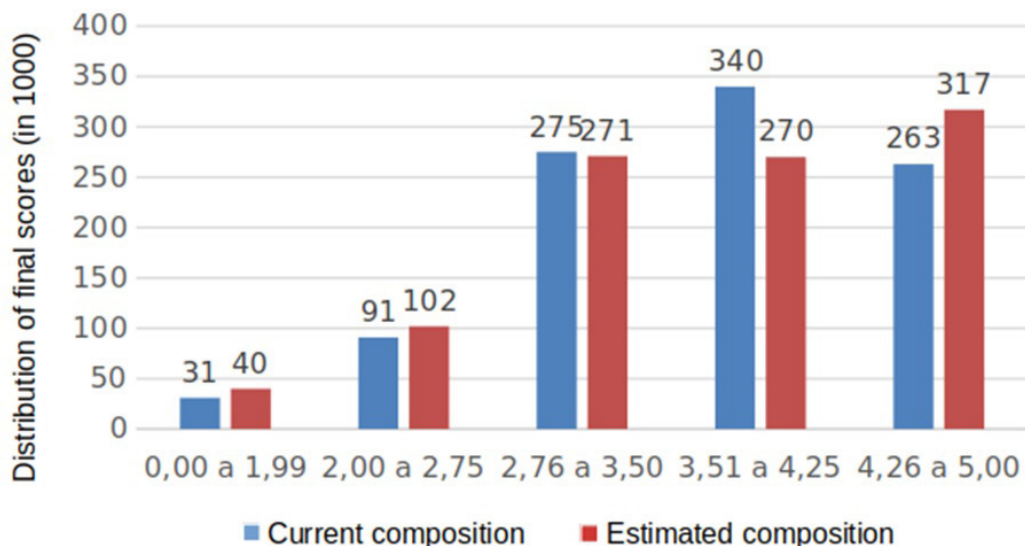


Source: prepared by the author based on data from Inep.

In Chart 3, with respect to the comparison of the score distribution into bands of proficiency, the blue bars show the classification of oral proficiency based on the current composition of the score. The band intervals are defined as follows: examinees with scores from 0.00 to 1.99 are classified as uncertified; from 2.00 to 2.75 as intermediate; from 2.76 to 3.50 as upper intermediate; from 3.51 to 4.25 as advanced; and examinees with scores from 4.26 to 5 are classified as upper advanced. The recalculation and reorganisation of examinees into the bands were based on a set of data that correspond to the scores of 1000 examinees. The classification into proficiency bands based on the final score of the oral examination was compared considering the composition of the final score based on the current weightings in the blue bars and based on the proposed new weightings in the red bars in Chart 3.

Chart 3 – Comparison of the distribution of scores into proficiency bands

Comparison of the distribution of scores into proficiency bands



Source: prepared by the author based on data from Inep

In general, the examinees in the advanced band were scattered to other classification bands, focusing especially on upper advanced. Composing the final score with the new weightings increased the uncertified and intermediate bands compared to the classification based on the composition of the score using the current weightings.

There was a tendency for an increase in examinees classified in the basic, intermediate and upper-intermediate bands when comparing the score calculated with the new weightings with the calculation currently used. After these levels, the trend reverses, because the number of examinees in these bands decreased. In other words, composing the score from the analytical items with the proposed weightings, it is likely that the number of examinees classified in the advanced and upper-advanced bands will decrease and the number classified in the basic, intermediate and upper-intermediate bands will increase. The recalculation of the items that make up the observer-rater’s score can rearrange the classifications in terms of a reduction of the examinee’s classification score, since the proposal presented implies placing more weighting on items related to linguistic adequacy, for which the participants tend to achieve lower scores, and less weight on items such as oral comprehension, for which examinees achieve high scores. Although there is this trend with respect to the observer’s score, the sum of the final score showed no increase in examinees classified as basic, intermediate and upper intermediate, nor was there a decrease in examinees in the advanced and upper-advanced bands. This is explained by the reduction of the interviewer’s score in the composition of the final score. The new final

score decreased because it consisted more of the observer's score, accounting for 66.34% in the composition of the score, than the interviewer's score, with 33.67%.

Although the weightings of the analytical items lexical adequacy, grammatical adequacy and fluency increased in the new proposal for the composition of the oral score, which probably increases the likelihood of the examinees achieving low scores, from the analysis, the interviewer's score seems to be described in a way that is difficult for examinees to be classified in the upper-advanced band. According to the band descriptors in the interviewer's matrix (Appendix 1), what differentiates between scores 4 and 5 is that 5 "presents fluency and a wide variety of vocabulary and structures, with very few inadequacies. Pronunciation is adequate" and 4 "presents fluency and a wide variety of vocabulary and structures, with occasional inadequacies in communication. Pronunciation may present some inadequacies", while in relation to autonomy, resourcefulness and understanding, the descriptors are the same. There seems to be a tendency for the interviewer to opt for the advanced level between the advanced and upper-advanced bands, and therefore, on reducing the weighting of the interviewer's score and increasing that of the observer, the number of examinees classified in the upper-advanced band increased. In other words, increasing the weightings of the analytical items related to linguistic aspects and increasing the weighting of the observer's score does not necessarily mean that the final examination score will decrease, because the interviewer's judgement seems to tend to focus on classifying examinees in the advanced band when there is a doubt as to the classification between advanced and upper advanced. Thus, on reducing the weighting of the interviewer's score in the composition of the new final score, the examinees were reorganised such that the number classified in the upper-advanced band increased.

Final considerations

The Celpe-Bras examination is a large-scale assessment tool that aims to certify proficiency in the Portuguese language for speakers of other languages. The examination consists of assessments of oral and written proficiency. In the present study, evidence was submitted to support the discussion of the relationship between score and construct. According to Messick (1987), through analysing the meaningfulness of the score, it is possible to acquire strong evidence on the construct validity of assessment tools. To this end, a factorial analysis was presented, there being more or less a consensus among statisticians that factorial calculation is related to questions that aim to investigate the construct being measured by some tool. In this sense, the factorial analysis presented here was efficient in generating evidence about what the test measures and how much each item corresponds to the measurement, i.e., how much each item contributes to the composition of the final score. Based on the factorial analysis, it was found that both the analytical matrix and that of the interlocutor-rater measure only one construct – oral proficiency. This means that the tool is one-dimensional, i.e., it measures only one thing – oral proficiency. In other words, the assessment matrix is valid from the point of view

of the construct it is intended to assess, although the analysis revealed the need to review some items.

The factor loadings of the scale items ranged from 0.65 to 0.94, suggesting that the items may be explaining the same factor (oral proficiency). The weighting values for each of the items with their respective confidence intervals are, from largest to smallest: interviewer's score 0.36 (0.33–0.42); lexical adequacy 0.19 (0.15–0.22); fluency 0.18 (0.15–0.22); grammatical adequacy 0.13 (0.05–0.15); interactional competence 0.09 (0.07–0.11); pronunciation 0.06 (0.04–0.08) and oral comprehension 0.04 (0.03–0.05). In the proposed analysis, the analytical score as a whole represents 66% of the total for the examination. Oral comprehension stood out in the analysis by disagreeing slightly with the values from the factorial analysis, indicating the need to evaluate not only its weighting, but also to think about the extent to which oral comprehension is being assessed in the examination situation proposed by the Celpe-Bras. It seems reasonable to assert that candidates are being challenged more as to their grammatical or lexical adequacy than as to their capacity for oral comprehension. In the present assessment context, would oral comprehension not be a prerequisite for interlocution to occur?

The factorial analysis generated information for the creation of a new manner to calculate the composition of the final oral score of the examination. After applying the new weightings for each of the items in the composition of the final score, the implications of the change in the composition of the new final score based on the new weighting was discussed with respect to possible changes in the classification bands for the examination. After applying the current weightings and the proposed weightings to the same set of scores and comparing them to the distribution of the candidates in each of the classification bands of the examination, it was found that the analytical scores of the advanced band were distributed among the other bands.

Regarding the final score after applying the new weightings, the most significant change was the reduction in the number of participants classified in the advanced band and increase of participants in the upper-advanced band. As the percentage of the interlocutor-rater's score was the one that suffered the greatest changes, it was argued that the interlocutor-rater's score can tend to focus candidates into the advanced band. In other words, by the way in which the descriptors are organized, it may be unlikely for a interlocutor-rater to classify the examinee in the upper-advanced band. It would be interesting for other studies to investigate candidates with a high level of proficiency for the purpose of describing their performance in the advanced bands so that there can be more clarity as to what is being assessed and how. In addition to this, other analyses – both quantitative and qualitative – are necessary to investigate the relationship between the analytical and holistic matrices in terms of a new composition of weightings and its implications for the classification of candidates.

It is hoped that the results of this research may serve to substantiate the argument of the validity of the Celpe-Bras oral examination and to refine the process of assigning the oral score.

References

AERA. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for educational and psychological testing. New York: AERA, 2014.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Edital n. 1, de 28 de janeiro de 2016 - de abertura de inscrições do exame Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras/2016.1). Brasília, DF: MEC, 2016a. Disponível em: <http://download.inep.gov.br/outras_acoes/celpe_bras/legislacao/2016/edital_n1_de28012016_celpe_Bras_2016.1.pdf>. Acesso em: 04 set. 2017.

BRASIL, Ministério da Educação. Secretaria de Ensino Superior. Certificado de Proficiência em Língua Portuguesa para estrangeiros: grades de avaliação holística e analítica. Brasília, 2016b.

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. Certificado de proficiência em língua portuguesa para estrangeiros: manual do participante. Brasília, DF: MEC, 2010.

BROWN, Timothy A. Confirmatory factor analysis for applied research. New York: Guilford Press, 2015.

BYGATE, Martin. Teaching and testing speaking. In: LONG, Michael H.; DOUGHTY, Catherine J. The handbook of language teaching. Chichester: Wiley-Blackwell, 2009. p. 411-440.

CANTY, Angelo; RIPLEY, Brian. Boot: Bootstrap R (S-Plus) functions. R package, versão 1, p. 3-20, 2017.

DAVISON, Anthony C.; HINKLEY, David Victor. Bootstrap methods and their applications. Cambridge: Cambridge University Press, 1997.

ECKES, Thomas. Introduction to many-facet rasch measurement: analyzing and evaluating rater-mediated assessments. Frankfurt: Peter Lang, 2015.

FIGUEIREDO FILHO, Dalson Brito; SILVA JÚNIOR, José Alexandre da. Visão além do alcance: uma introdução à análise fatorial. Opinião Pública, Campinas v. 16, n. 1, p. 160-185, 2010.

FULCHER, Glenn. Testing second language speaking. London: Routledge, 2003.

FULCHER, Glenn; DAVIDSON, Fred. Language testing and assessment: an advanced resource book. Routledge: New York, 2007. p. 91-114.

KIM, Jae-on, MUELLER, Charles W. Factor analysis: statistical methods and practical issues. Iowa: Sage University Press, 1978.

KUNNAN, Antony John. An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. Language Testing, Newbury Park, v. 9, p. 30-49, 1992.

McNAMARA, Tim. Language testing. In: DAVIES, Alan; ELDER, Catherine. The handbook of applied linguistics. Malden: Blackwell, 2004. p. 763-783.

MESSICK, Samuel. Validity. New Jersey: Educational Testing Service Princeton, 1987.

REVELLE, William. Psych: procedures for personality and psychological research, version 1.8.4. Evanston: Northwestern University, 2018. Disponível em: <<https://CRAN.R-project.org/>>. Acesso em: maio 2018.

R CORE TEAM. R: a language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2018. Disponível em: <<http://www.R-project.org/>>. Acesso em: maio 2018.

THOMPSON, Bruce. Exploratory and confirmatory factor analysis: understanding concepts and applications. Washington: American Psychological Association, 2004.

Received on June 06th, 2018

Approved on September 26th, 2018

Laura Márcia Luiza Ferreira is a professor at the Federal University of Latin-American Integration (UNILA), PhD from Cefet-MG, Master in Linguistics and licensed in Letters by UFMG. She was a lecturer at Chulalongkorn University, Thailand, and a teacher in Timor-Leste. She researches internal and external evaluations, especially the Celpe-Bras examination.