Educação e Pesquisa
ISSN: 1517-9702
ISSN: 1678-4634
Faculdade de Educação da Universidade de São Paulo

Ferreira, Laura Márcia Luiza
Um estudo sobre a dimensionalidade das escalas de avaliação da proficiência oral do Certificado de Proficiência em Língua Portuguesa para Estrangeiros
Educação e Pesquisa, vol. 45, e202512, 2019
Faculdade de Educação da Universidade de São Paulo

DOI: 10.1590/S1678-4634201945202512

Um estudo sobre a dimensionalidade das escalas de avaliação da proficiência oral do Certificado de Proficiência em Língua Portuguesa para Estrangeiros

Laura Márcia Luiza Ferreira[1]
ORCID: 0000-0001-7632-0834

Resumo

Por meio das escalas de avaliação do Celpe-Bras, um construto de proficiência oral em língua portuguesa para falantes de outras línguas é operacionalizado e mensurado. Na prova oral, sete itens compõem duas escalas por meio das quais o avaliador-interlocutor e o avaliador-observador atribuem uma nota para cada um dos seis itens, a saber: compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia. Com objetivo de analisar a dimensionalidade das escalas, apresento a análise fatorial exploratória do conjunto de notas de 1.000 participantes que se submeteram ao exame na primeira edição de 2016. O coeficiente de determinação R^2 foi de 0.9617 e o índice Tucker-Lewis (TLI) foi de 0.896. Os valores de carga fatorial dos itens da escala variaram de 0.65 a 0.94, sugerindo que os itens possam estar explicando o mesmo fator. O resultado da análise aponta que a nota da prova é uma medida unidimensional. A partir dos valores de comunalidade, apenas o item compreensão ficou ligeiramente abaixo 0.5, apontando a necessidade de investigação do item. Os valores de peso para cada um dos itens são, do maior para o menor; para nota do entrevistador, 0.36; adequação lexical, 0.19; fluência, 0.18; adequação gramatical, 0.13; competência interacional, 0.09; pronúncia, 0.06; compreensão, 0.04. Por meio da análise fatorial, apresento uma proposta de composição da nota da prova oral e discuto as implicações da mudança de peso dos itens na nova proposta para a classificação dos participantes por faixas de certificação.

Palavras-chave

Dimensionalidade – Avaliação de língua adicional – Proficiência oral – Validade – Análise fatorial.

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

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

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.

Introdução

O Certificado de Proficiência em Língua Portuguesa para Estrangeiros, doravante Celpe-Bras, é o exame oficial do Governo do Brasil para comprovação de proficiência de estrangeiros em língua portuguesa. Atualmente, o exame está sob tutela do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). A certificação de proficiência atestada por meio do Celpe-Bras pode ser exigida aos estrangeiros em algumas situações tais como: candidatura a programas de cooperação educacional financiados pelo Governo do Brasil e processo de revalidação de diplomas, a depender dos requisitos exigidos pelos conselhos profissionais.

Por meio do Celpe-Bras, é possível obter certificação da proficiência nos níveis intermediário, intermediário superior, avançado e avançado superior, após a realização de uma prova única. O Celpe-Bras está estruturado em duas partes: prova escrita e prova oral.

A primeira é composta por quatro itens abertos, isto é, a redação de quatro textos, os quais são avaliados a partir de uma escala holística que gera um escore. A nota final da parte escrita é calculada a partir da média aritmética simples referente aos quatro itens. A nota da prova oral é composta de maneira um pouco mais complexa. A menor nota entre as duas etapas de prova é a que vale para fins de certificação da proficiência (BRASIL, 2016a).

A prova oral é composta por sete itens que estão organizados em duas escalas de avaliação. Após uma interação oral face a face com duração de vinte minutos entre o avaliador-interlocutor e o participante, a nota é atribuída por dois examinadores nos locais onde são realizadas as interações. Tanto o avaliador-interlocutor quanto o avaliador-observador atribuem notas independentes para o desempenho oral do participante. O avaliador-interlocutor atribui uma nota única a partir de uma grade holística (Figura 1) e o avaliador-observador utiliza uma grade analítica (Figura 2) na qual a nota é composta por mais seis itens de avaliação: compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia. Ou seja, ao todo, a nota da prova oral é composta a partir de sete itens que compõem as duas grades de avaliação: a do observador e a do interlocutor.

No Manual do Participante (BRASIL, 2010), a parte oral é dividida em duas etapas: na primeira, a interação se dá a partir das informações que constam no questionário de inscrição do participante, com duração de cinco minutos, e a segunda, a partir de três elementos provocadores selecionados pelo avaliador-interlocutor. Na segunda etapa, a interação dura quinze minutos e está dividida em três partes, sendo cinco minutos dedicados a cada um dos três elementos provocadores. Os elementos provocadores são, em sua maioria, recortes de reportagem que circulam na mídia impressa brasileira.

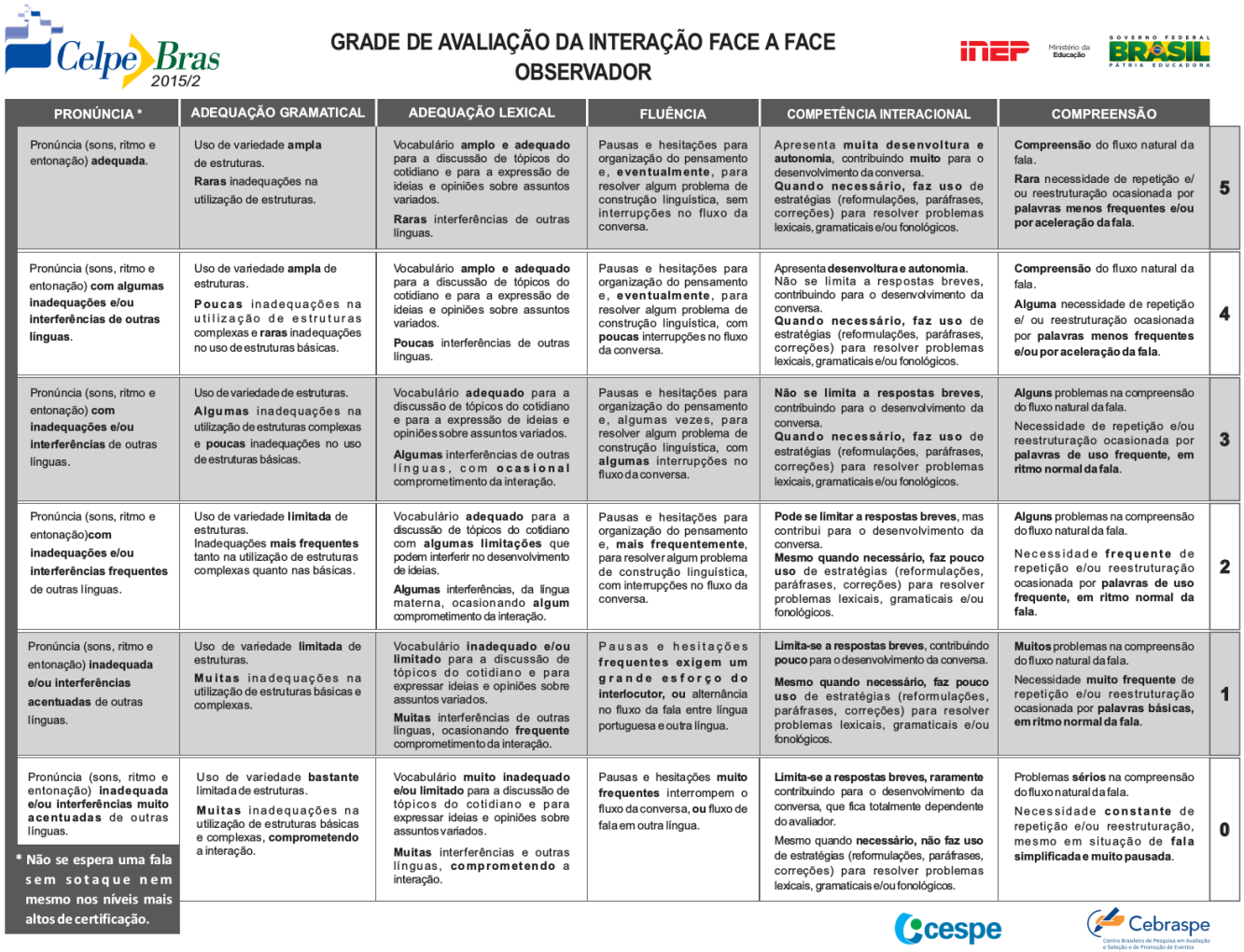
A avaliação é gravada e enviada ao Inep. Segundo o edital de inscrições (BRASIL, 2016a), a nota final da prova oral é calculada a partir de uma média entre as notas do avaliador-interlocutor e as do avaliador-observador, ou seja, cada nota tem um peso de 50% na composição da nota final da prova oral. Caso as notas atribuídas entre os avaliadores sejam divergentes em mais de um ponto e meio (1.5), a interação é reavaliada em eventos de correção do exame. Além disso, a prova oral pode ser reavaliada por um terceiro avaliador, quando o resultado for discrepante em até dois pontos em relação à nota da prova escrita, quando a diferença de notas entre as duas modalidades da prova implicar mudança do nível de certificação ou, ainda, quando a nota final na escrita for superior à oral.

Em exames de desempenho, como é o caso de Celpe-Bras, a atribuição da nota é mediada pelos avaliadores que utilizam escalas para guiar o julgamento da performance do participante. Eckes (2015) explica que, na situação de interação face a face, há pelo menos cinco facetas[2] e variadas maneiras de interação entre elas, que podem impactar no resultado final, a saber: o participante, a tarefa, o avaliador-interlocutor, a escala, ou modelo de atribuição de nota e o avaliador.

Neste trabalho, discuto o modelo de atribuição de notas da prova oral do exame Celpe-Bras com objetivo de avaliar a unidimensionalidade da medida composta pelos sete

2- Segundo Eckes (2015), facetas são sinônimos de fatores, variáveis ou componentes que fazem parte da situação de avaliação e que afetam as notas de forma sistemática.

Figura 1- Grade de avaliação do avaliador-observador



CelpeBras 2015/2

GRADE DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE
OBSERVADOR

INEP Ministério da Educação — Governo Federal Brasil Pátria Educadora

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 com **algumas limitações** que podem interferir no desenvolvimento de ideias. **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.**

cespe — Cebraspe Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

Fonte: INEP, 2016.

Figura 2- Grade holística ou avaliação do avaliador-interlocutor

Celpe Bras

FICHA DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE | ENTREVISTADOR

INEP — Ministério da Educação — GOVERNO FEDERAL BRASIL PÁTRIA EDUCADORA

Nome do examinando: **Xxxxxxxxxxxxxxxxxxxxx xx Xxxxxxxxxxxxxxxxxxxxx xx Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**

Nacionalidade: **XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX** Número de Inscrição: **9999999999**

Documento de identificação n.º: **XXXXXXXXXXXXXXXXXXXXXX**

Posto aplicador **XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX**

4360144035

Elementos provocadores utilizados

1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ 8 ◯ 9 ◯ 10 ◯ 11 ◯ 12 ◯ 13 ◯ 14 ◯ 15 ◯ 16 ◯ 17 ◯ 18 ◯ 19 ◯ 20 ◯

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	◯	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	◯	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	◯	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	◯	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	◯	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	◯	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

cespe

Cebraspe — Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

Fonte: INEP, 2016.

itens das escalas de avaliação da prova oral. Para tanto, apresento o resultado de uma análise fatorial exploratória.

A seguir, apresento brevemente o conceito de construto no âmbito das avaliações orais de língua adicional.

Construto em avaliação oral de línguas adicionais

Segundo Bygate (2009), as teorias sobre metodologia de ensino de línguas estrangeiras (LE) deixaram a desejar na tentativa de elaborar um construto sobre o desenvolvimento da oralidade. Ao fazer uma breve retrospectiva, o autor relembra que, nas abordagens estruturalistas, o foco incidia sobre a gramática. Embora no método áudio-lingual a fala ocupasse um lugar central, a oralidade era vista mais como um meio de aquisição da LE do que como um fim. O autor defende ainda que, mesmo nas mais diversas formas de encarar a abordagem comunicativa do ensino de línguas, a oralidade ainda assim é vista mais como um meio do que meta a ser estabelecida e atingida.

No contexto dos estudos sobre avaliações, tais metas se aproximam da noção de construto. Construto é o que pode ser observado e medido. Por buscar operacionalizar a natureza do construto da oralidade, testes de proficiência oral, como o Celpe-Bras, são de grande utilidade para nortear questões centrais na discussão da oralidade como os construtos da fala, da tarefa, do critério do desempenho e do desenvolvimento oral da fala (BYGATE, 2009). Segundo o autor, os testes são instrumentos valiosos de análise para discussão sobre a compreensão dos possíveis parâmetros da proficiência oral.

Fulcher (2003) salienta que, no caso de ser a proficiência oral um construto a ser medido e observado, é necessário que ela esteja associada com algo que possa ser observado e mensurado. Fulcher (2003), assim como Bygate, aponta o problema de não haver um construto pronto e eficiente sobre proficiência oral em língua adicional, e argumenta que não há como existir consenso entre os teóricos e professores sobre ele. A definição de um construto, segundo Fulcher (2003), é uma questão de escolher algumas teorias e tentar operacionalizá-las em um contexto de avaliação com seus propósitos específicos, providenciando uma fundamentação teórica e empírica para as escolhas feitas. McNamara (2004) também ressalta que os construtos serão sempre controversos e alvos de críticas, por isso devem estar articulados com os argumentos que defendem a validade do teste.

Os construtos podem ser baseados em teorias ou podem ser elaborados a partir de uma composição de conceitos. Messick (1987) questiona o fato de os construtos não serem baseados em uma teoria, mas em uma composição de conceitos. Para o autor, é possível investigar a eficiência da composição ao verificar até que ponto as medidas estão avaliando um mesmo construto. No caso do Celpe-Bras, como há duas grades distintas avaliando a mesma coisa, e variados itens que refletem diversos conceitos teóricos sobre desenvolvimento da oralidade em língua estrangeira, é preciso investigar como as notas atribuídas por meio da grade holística e da grade analítica estão relacionadas. Messick (1987) afirma que a representação do construto se refere à relativa dependência do desenho da prova, ou seja, das escalas dos itens avaliados. No caso da prova oral do Celpe-Bras, por exemplo, a representação do construto está relativamente dependente da nota do avaliador-interlocutor, da nota de compreensão, da nota de pronúncia etc.

Fulcher e Davidson (2007), ao tratarem do sistema de atribuição de notas em teste de línguas, afirmam que o processo de julgamento do avaliador é o que conecta a evidência de performance, que pode ser representada pela nota, à tarefa e ao construto. Para se chegar às inferências, é preciso coletar evidências que podem estar relacionadas com as notas. De acordo com Messick (1987), porém, nas abordagens empíricas de construção de teste, os itens deveriam entrar em sua composição depois de feitas a análise dos dados, sejam eles dados internos ao teste, que demonstrem a homogeneidade do item ou suas cargas fatoriais, ou dados externos, que envolvam o estudo da correlação do parâmetro de avaliação ou a correlação da discriminação do critério com relação a um conjunto de outros parâmetros. Tais análises se referem ao aspecto substancial da validade de construto, segundo Messick (1987).

No documento da American Educational Reseach Association (AERA, 2014), por aspecto substancial, entende-se a evidência baseada na estrutura interna do teste. Cabe ressaltar que tanto Messick (1987) quanto o Standards (AERA, 2014) apontam para a análise fatorial como forma de analisar a unidimensionalidade da medida, ou seja, se os itens estão medindo o mesmo construto. Por exemplo, o quanto cada item de avaliação exemplifica o construto que está sendo medido é uma questão de aspecto substantivo da validade de construto, ou seja, uma evidência de validade baseada na estrutura interna do teste.

Messick (1987) aponta que a análise fatorial é recomendável para avaliar as relações entre os itens de avaliação por meio da análise das cargas fatoriais. As cargas fatoriais são valores que nos permitem avaliar o quanto cada item de avaliação compõe a nota final de uma avaliação. Messick (1987) afirma que a análise fatorial é recomendada quando se quer combinar a avaliação de teorias e a construção de escalas para interpretar a consistência das respostas.

Trato com mais detalhes, a seguir, da metodologia da análise fatorial.

Análise fatorial e a avaliação em línguas adicionais

A análise fatorial pode servir para investigar a validade do construto, verificar hipóteses teóricas e resumir ou agrupar um grande volume de dados. No campo da estatística, o substantivo *validade* era seguido do termo *fatorial*. Thompson (2004) revisita o texto de Nunnally de 1978, e afirma que o termo histórico para validade de construto é validade fatorial. Thompson (2004) sugere que, quando estamos desenvolvendo documentos de especificações relacionadas a uma medida, como escalas de avaliação e graduação de descritores por níveis de proficiência, a análise fatorial deveria ser utilizada para examinar a validade da nota. Thompson (2004) explica que, se o pesquisador tem o objetivo de responder perguntas relacionadas àquilo que o teste mede, a resposta deveria ser respondida em termos fatoriais.

Brown (2015) é também um entusiasta do uso da metodologia da análise fatorial para se verificar a validade de um construto em pesquisas da área de ciências sociais e comportamentais. De acordo com o autor, a análise pode oferecer evidências empíricas sobre validade convergente ou discriminante em relação a construtos teóricos. Segundo o autor, a análise fatorial pode mostrar evidência empírica de forte inter-relação entre os

itens avaliados que são similares ou estão sobrepostos do ponto de vista teórico ou de fraca inter-relação quando fizerem parte de construtos teóricos distintos.

No caso da metodologia de avaliação do exame do Celpe-Bras, por exemplo, a análise fatorial pode apontar quais itens da grade analítica estão mais fortemente relacionados entre si. Quanto mais os parâmetros estiverem relacionados entre si, mais evidências de que a avaliação está sendo feita a partir de um mesmo construto teórico (o da proficiência oral, no caso do presente trabalho). Thompson (2004) ressalta que, embora os termos relacionados ao conceito de validade não incluam a validade fatorial, a análise fatorial continua sendo ferramenta útil na construção de questões relacionadas à validade.

A análise fatorial pode ser o ponto de partida de muitas pesquisas que analisam notas atribuídas para retratar algum tipo de desempenho (FULCHER, 2003; BROWN, 2015). Fulcher (2003) exemplificou a metodologia ao citar o trabalho de Hinofotis, de 1983, em que analisou doze parâmetros para avaliar a comunicação de professores assistentes na Universidade da Califórnia com seus estudantes em situações de interação oral em sala de aula.

Por meio da análise fatorial, Hinofotis (1983 apud FULCHER, 2003) investigou a relação entre: vocabulário, gramática, pronúncia, fluência, contato visual, aspectos não verbais, segurança, presença, desenvolvimento de argumentação, uso de evidências ao argumentar, clareza e relacionamento com os estudantes. O pesquisador partiu da hipótese de que os parâmetros poderiam ser agrupados em cinco fatores e, após interpretação dos dados, concluiu que o fator 1 (comunicação e informação) é fortemente influenciado pelos parâmetros desenvolvimento de argumentação, uso de evidências ao argumentar, clareza e relacionamento com os estudantes; o fator 2 (expressão) é impactado pela fluência e habilidade de se relacionar com os alunos; o fator 3 (aspectos não verbais), por aspectos não verbais e habilidade de relacionar com os alunos; o fator 4 (proficiência linguística), pelo vocabulário e gramática e o fator 5, (pronúncia), apenas pelo parâmetro da pronúncia. Fulcher (2003) chama atenção para o fato de a pronúncia não fazer parte empiricamente do fator que diz respeito ao construto da proficiência linguística e chama atenção para o fato do critério relação com os estudantes estar presente em dois fatores, o da comunicação e informação e o da expressão.

Ao final, Fulcher (2003) conclui sobre o método da análise fatorial que, se o argumento do pesquisador, baseado na análise, for plausível, então o ele terá êxito ao apresentar evidências para fundamentar uma inferência sobre o significado da nota.

Ainda no campo da avaliação de línguas estrangeiras, Kunnan (1992) também utilizou a metodologia para analisar o significado da nota de um teste de nivelamento da Universidade da Califórnia (UCLA) e empregou, dentre outros métodos, a análise fatorial exploratória em quatro grupos de estudantes de inglês como segunda língua, para investigar a validade de instrumento que avalia separadamente as habilidades de leitura, compreensão oral e gramática. Ao final do estudo, Kunnan (1992) concluiu, a partir da análise fatorial, que os estudantes com baixa proficiência tendem a ter notas baixas em diferentes habilidades, uma vez que a carga fatorial deste grupo pesa para um só fator, ao passo que estudantes mais proficientes podem ter variação no domínio das habilidades de ler, compreender oralmente e usar a gramática da língua, porque a carga fatorial é distribuída. Baseando-se nas análises, o autor sugere que a nota por habilidade, ou seja, separada por seção do teste – leitura, compreensão oral, gramática – deveria ser usada

para encaminhamento dos estudantes para os níveis de estudo e não a nota total, como normalmente era feito no âmbito do programa de estudos de línguas analisado pelo autor.

Além de ser útil para avaliar as notas de um instrumento que já está elaborado, Brown (2015) afirma que a análise fatorial é uma ferramenta popular para o desenvolvimento e construção de escalas de avaliação. Por meio do cálculo das cargas fatoriais, é possível definir como os escores devem ser atribuídos, seja por meio de um instrumento de avaliação que prevê um conjunto de itens dicotômicos que se referem a diferentes habilidades linguísticas, como a prova de nivelamento estudada por Kunnan (1992), ou itens politômicos, como a proposta de Hinofotis (1986 apud FULCHER, 2003), que é similar às escalas da prova oral do exame Celpe-Bras. De acordo com Brown (2015), a análise fatorial pode ser usada para verificar, por exemplo, o número de itens de avaliação da prova oral que estão relacionados com o fator, ou dimensão do construto da proficiência oral, e os padrões de cada um dos itens com relação ao(s) fator(es) ou dimensões da proficiência oral. No contexto do presente estudo, a análise fatorial fornecerá elementos para a discussão sobre a relação entre a nota e o construto, composto pelos itens da prova oral do Celpe-Bras, e para uma proposta de reformulação de pesos para composição da nota final oral.

A seguir, apresento a análise e a discussão dos resultados.

Análise e discussão

Os dados correspondem à avaliação de desempenho oral de 1.000 examinandos que se submeteram à avaliação na edição do primeiro semestre de 2016. O conjunto de dados analisados apresenta sete variáveis: seis notas referentes aos seis itens avaliados na grade do observador e uma nota total denominada nota do interlocutor. Ou seja, na análise, foram utilizados os dados referentes às notas dos seis itens que compõem a grade analítica e à nota do interlocutor. A nota final do observador e a nota final da prova não foram consideradas nos cálculos apresentados a seguir.

A análise foi realizada em várias etapas, de maneira a identificar como os aspectos avaliados contribuem para a composição da nota oral do examinando. Foi utilizado para fazer os cálculos o *software* estatístico R (R CORE TEAM, 2018), versão 3.5.0, de 23 do maio de 2018, para Windows 10. O programa R é um *software* livre que permite diversos cálculos estatísticos. Para a análise fatorial, foi utilizado o pacote Psych versão 1.8.4 (REVELLE, 2018).

O *bootstrap* foi o meio utilizado para cálculo dos intervalos de confiança para os pesos e cargas fatoriais (DAVISON; HINKLEY, 1997; CANTY; RIPLEY, 2017). A estimação por meio de *bootstrap* faz uso de conceitos do teorema central do limite. Independente da forma da distribuição dos dados, a distribuição amostral dos parâmetros de interesse consegue assumir uma distribuição normal. As notas orais do Celpe-Bras são assimétricas, porém a aplicação dos conceitos do teorema pode garantir resultados precisos para os valores calculados, independentemente da forma que os dados se apresentam. A adoção de estimação por reamostragem ou *bootstrap* se fez necessária para garantir a correta aplicação do teorema central do limite aos dados. Esse método consiste em fazer sucessivas amostragens nos dados disponíveis e calcular os valores de interesse. Após sucessivas amostragens, o valor final será a média dos valores calculados. No caso em estudo, foram retiradas 10.000 amostras com reposição, de tamanho 1.000 dos dados em estudo.

A amostra é composta por notas elevadas (Tabela 1). Os dados na tabela se referem às quantificações. Podemos notar que 2,5% dos examinandos obtiveram notas de até 1,8373, e 75% dos examinandos obtiveram notas de até 4,29, em uma escala de zero a cinco pontos. Mais da metade da amostra se refere a notas iguais ou maiores do que 3,855. A distribuição das notas maiores que este valor se concentra nos valores maiores que 4,5. Do grupo de notas menores que 3,85, mais de 25% fica em torno de 3,25 e apenas 5% representam notas 2,17 das quais metade é nota 1,709 ou abaixo, ou seja, há pouquíssimas notas 1 na amostra. Vale ressaltar, que, na edição do primeiro semestre de 2016, 6.222 examinandos se inscreveram no exame. A amostra da pesquisa representa 16,07% do total de examinandos inscritos.

Tabela 1 – Distribuição normal padrão acumulada das notas finais da parte oral

	Nota do observador	Nota do entrevistador	Nota final da prova oral
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

Fonte: elaboração da autora a partir de dados do Inep.

Como o objetivo era o de entender como as seis variáveis da grade analítica e a nota do avaliador interlocutor compõem o fator proficiência oral na prática, e com o uso da grade pelos avaliadores, foi feita uma análise fatorial exploratória. Todas as notas atribuídas aos sete itens foram levadas em conta no cálculo que teve como base a análise fatorial dos eixos principais (*principal axis factoring*, PAF). Nessa análise, concluiu-se que as sete variáveis podem ser representadas por apenas um fator. Não foi necessária a rotação de fatores, porque a análise se reduziu a um fator. Foram testadas as hipóteses sobre a estrutura fatorial das notas estarem organizadas em um ou dois fatores. Cabe ressaltar que foram feitas análises preliminares utilizando equações estruturais e análise confirmatória, porém detectaram-se problemas de convergência.

A partir da análise fatorial, pressupõe-se que o fator proficiência oral esteja sendo explicado por sete variáveis, que seriam os seis itens que compõem a nota analítica mais a nota do entrevistador, a sétima variável. Para fazer a avaliação de ajuste local do modelo, analisou-se o coeficiente de determinação, o R^2, que se refere à porcentagem de variação das variáveis: notas da prova oral; que estão sendo explicadas pela estrutura fatorial calculada. A estrutura fatorial apresentada mostrou um R^2 de 0.9617319. Isso indica que os dados se adequaram ao modelo de análise e por ele podem ser explicados. Para fazer

a avaliação do ajuste de modelo, apresento o cálculo do RMSEA (*root square erros of approximation*) cujo valor foi de 0.18. Sugere-se como índice de bom ajuste que o valor fique em torno de 0.5. No caso da análise, uma hipótese possível para o alto valor do índice de RMSEA pode ser o fato de as notas estarem muito fortemente correlacionadas. O índice Tucker-Lewis Index (TLI) é outra maneira de avaliar a confiabilidade dos resultados calculados pelo modelo da análise. No caso do presente estudo, o valor foi de 0.896, trata-se de um resultado satisfatório, que sugere que os dados podem ser explicados pelo método de análise adotado.

Na análise, desconsiderou-se o peso atual do cálculo da nota final da prova oral praticado. Ao recalcular os valores para chegar às variáveis que mais explicam o fator ou construto da proficiência oral, temos a nota do entrevistador como a variável mais importante.

Tabela 2 – Características e pesos a serem atribuídos para cada variável analítica e nota entrevistador

	Carga	Peso	Comunalidade
Compreensão	0.6572906	0.0455407	0.43
Competência interacional	0.8028341	0.0950375	0.64
Fluência	0.8816178	0.1836518	0.78
Adequação lexical	0.9092421	0.1946057	0.83
Adequação gramatical	0.8852773	0.1350449	0.78
Pronúncia	0.8004653	0.0644487	0.64
Nota do entrevistador	0.9481050	0.3644149	0.90

Fonte: elaboração da autora a partir de dados do Inep.

A partir dos valores da carga apresentados na Tabela 2, é possível afirmar que as variáveis estão relacionadas ou que explicam um mesmo fator. Por meio da análise dos valores de carga fatorial, sugere-se que um só fator esteja influenciando os valores das notas atribuídas aos parâmetros, e por isso podemos afirmar que a medida é unidimensional. Afirmar que a medida é unidimensional, no nosso contexto, é o mesmo que dizer que as notas estão relacionadas a uma coisa só, isto é, à proficiência oral.

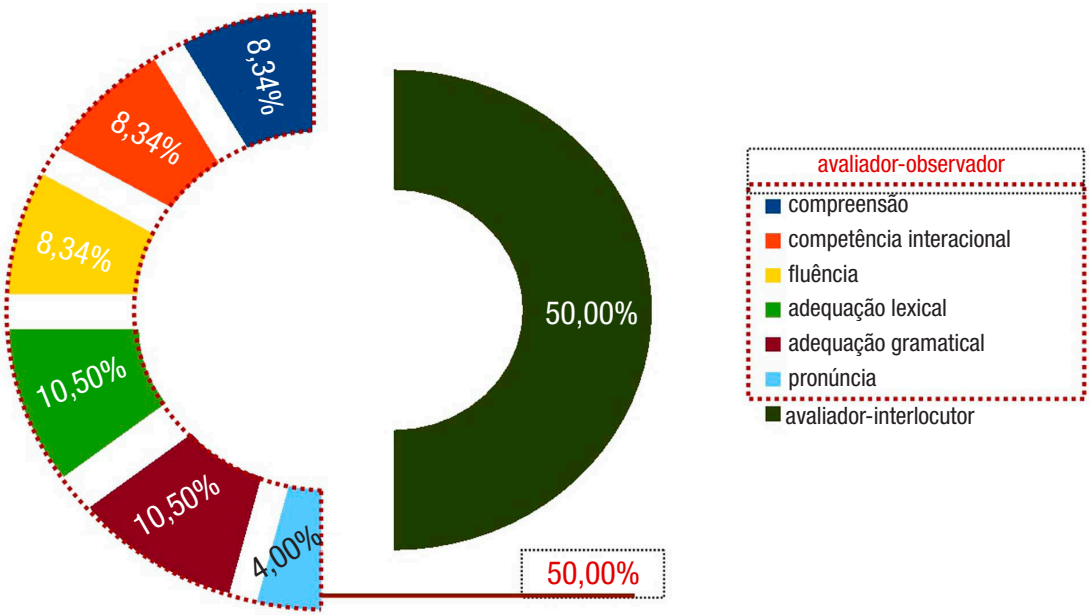
De acordo com Kim e Mueller (1978), o cálculo da comunalidade se dá a partir da correlação de cada variável com o restante do conjunto das variáveis, ou seja, ela tenta quantificar como a nota de compreensão, por exemplo, está correlacionada com o conjunto das outras notas. Figueiredo Filho e Silva Júnior (2010) explicam que é a partir do valor de comunalidade que podemos inferir que uma variável está linearmente correlacionada com as outras. Os autores afirmam que valores baixos de comunalidades (menores que 0.50) significam que elas possam não estar linearmente correlacionadas. Quanto aos valores de comunalidade da Tabela 2, temos um valor ligeiramente baixo para a nota de compreensão, sugerindo que o parâmetro possa estar menos relacionado com outros.

Os itens que têm a maior carga fatorial na nota final da prova oral com seus respectivos intervalos de confiança são, do maior para o menor: nota do entrevistador 0.95 (0.94-0.96); adequação lexical 0,91 (0.90-0.92); fluência 0,88 (0.86-0.90); adequação gramatical 0,88 (0.87-0.89); competência interacional 0,80 (0.77-0.83); pronúncia 0.80 (0.77-0.82); e compreensão 0.65 (0.60-0.69). A nota do interlocutor é o item que mais explica a nota da prova oral quando comparamos separadamente esta variável com as demais. Competência interacional e pronúncia são variáveis que contribuem de modo aproximadamente igual para o fator proficiência oral, assim como fluência e adequação gramatical, por apresentarem valores aproximados de carga. Sobressaem-se os valores 0.95 para o item nota do entrevistador e 0.90 para adequação lexical.

Os valores do peso representam o quanto cada aspecto contribui para a composição da nota final da avaliação oral. Os valores de peso para cada um dos itens, com seus respectivos intervalos de confiança são, do maior para o menor: nota do entrevistador 0.36 (0.33-0.42); adequação lexical 0.19 (0.15-0.22); fluência 0.18 (0.15-0.22), adequação gramatical 0.13 (0.05-0.15); competência interacional 0.09 (0.07-0.11); pronúncia 0.06 (0.04-0.08); compreensão 0.04 (0.03-0.05).

Os valores em peso calculados pela análise são aproximados e, ao somá-los, chegaríamos a um valor aproximado de 105. Assim sendo, os valores foram recalculados de forma a se acomodarem na métrica de 100% (conforme Gráficos 1 e 2, a seguir). Nos gráficos, os pesos que vigoram na composição da nota e os pesos aproximados propostos são comparados, fundamentados na análise fatorial para composição da nota final a partir do recálculo.

Gráfico 1 – Composição atual da prova oral

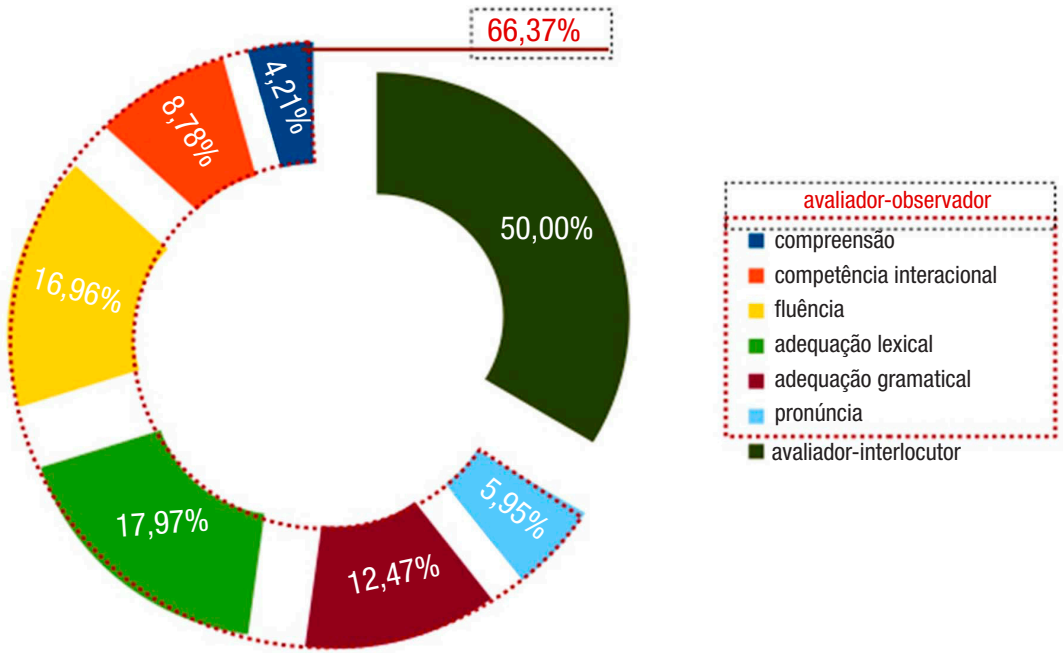


Fonte: elaboração da autora a partir de dados do Inep.

Considerando que a composição dos parâmetros analíticos teria 50% de peso na composição da nota final da prova oral, no Gráfico 1, apresenta-se a forma como é feito o cálculo da nota da prova oral que hoje vigora. No Gráfico 2, desconsidera-se o peso de 50% da nota do avaliador-interlocutor, e os valores dos seis parâmetros da grade analítica e da nota do avaliador-interlocutor formam sete variáveis na composição da nota final na prova oral, ou seja, considera-se a nota do avaliador-interlocutor como uma variável sem um valor fixo de peso. Vale ressaltar que o parâmetro que teve seu peso mais diminuído proporcionalmente foi o da compreensão. Os outros parâmetros transferiram, em geral, um terço do seu peso para nota do interlocutor.

Com relação ao peso do conjunto de parâmetros que compõem a nota do avaliador-observador e da nota única do avaliador-interlocutor, embora a nota do avaliador-interlocutor seja o parâmetro que mais explica a nota da prova oral com um valor de 33.67%, a nota do avaliador-observador, ou seja, a composição entre as outras seis notas é a que explica mais a nota oral final. Ao somarmos o peso dos seis parâmetros que compõem a nota analítica, temos 66.34% da nota final da prova oral explicada pela soma dos pesos de compreensão, competência lexical, fluência, adequação lexical, adequação gramatical e pronúncia. Dizendo de outra forma, a nota do observador é mais importante do que a nota do interlocutor, porque ao somarmos os pesos dos parâmetros analíticos na composição da nota final temos mais de 50% da composição da nota final explicada pela nota atribuída pelo avaliador-observador.

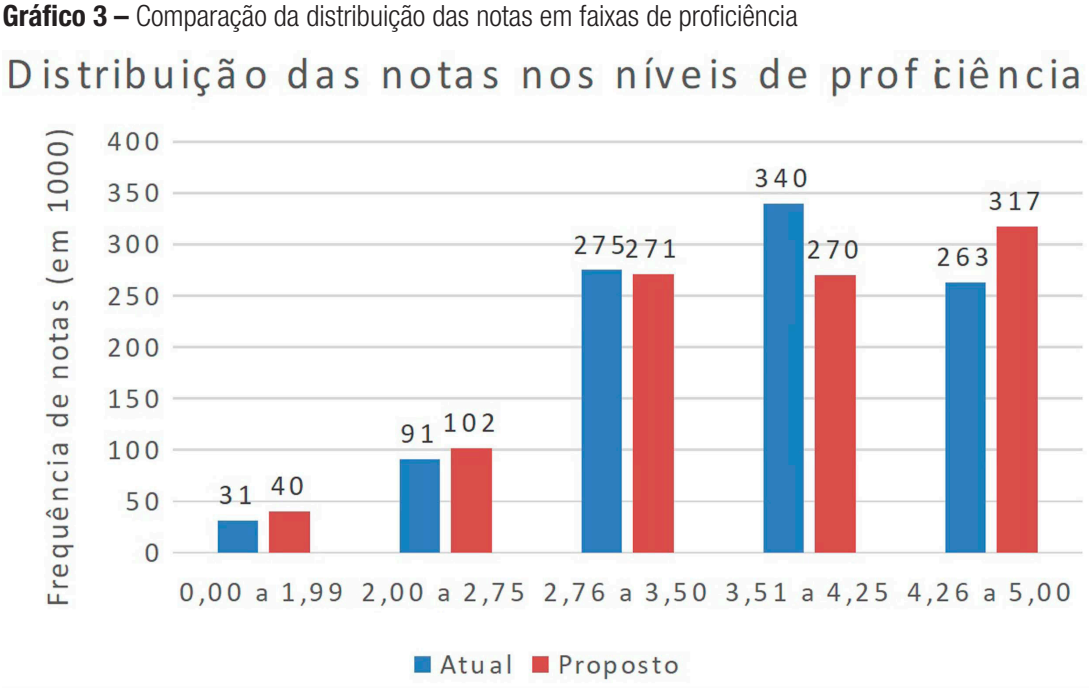
Gráfico 2 – Composição estimada da nota da prova oral



Fonte: elaboração da autora a partir de dados do Inep.

No Gráfico 3, referente à comparação da distribuição das notas em faixas de proficiência, apresenta-se nas colunas em azul a classificação da proficiência oral a partir da composição da nota que vigora atualmente. Os intervalos de faixas são definidos da seguinte forma: examinandos com notas entre 0,00 a 1,99 são classificados como sem certificação; entre 2,00 a 2,75 são classificados como intermediário; entre 2,76 a 3,50 são intermediário superior; de 3,51 a 4,25 são avançado; e entre 4,26 a 5 os examinandos são classificados como avançado superior. O recálculo e reorganização dos examinandos nas faixas foi feito a partir do conjunto de dados que correspondem às notas de 1.000 examinandos. A classificação por faixas de proficiência a partir da nota final da prova oral foi comparada, considerando-se a composição da nota final a partir dos pesos atuais, nas faixas azuis, e a partir dos novos pesos propostos, nas faixas vermelhas do Gráfico 3.

Gráfico 3 – Comparação da distribuição das notas em faixas de proficiência



Fonte: elaboração da autora a partir de dados do Inep.

De maneira geral, os examinandos da faixa avançado espalharam-se por outras faixas de classificação, concentrando-se especialmente em avançado superior. Ao compor a nota final com os novos pesos, as faixas sem certificação e intermediário aumentaram em comparação com a classificação feita a partir da composição da nota com os pesos que agora vigoram.

Houve uma tendência de aumento de examinandos classificados nas faixas básico, intermediário e intermediário superior quando comparamos a nota que foi calculada com os novos pesos e com o cálculo atualmente em vigor. Após estes níveis a tendência se inverte, pois o número de examinandos nestas faixas diminuem. Ou seja, compondo a

nota dos itens analíticos com os pesos propostos, é provável que diminua a quantidade de examinandos classificados nas faixas avançado e avançado superior e aumentem os classificados nas faixas básico, intermediário e intermediário superior. O novo cálculo dos itens que compõem a nota do avaliador-observador pode reorganizar as classificações no sentido de uma diminuição da nota de classificação do participante, uma vez que a proposta apresentada implica em colocar mais peso em itens relacionados à adequação linguística, nos quais os participantes tendem a tirar geralmente notas menores, e menos peso em parâmetros como compreensão, no qual os examinandos tiram notas altas. Embora haja essa tendência com relação à nota do observador, na soma da nota final não se verificou um aumento de examinandos classificados em básico, intermediário e intermediário superior, nem uma diminuição de examinandos em avançado e avançado superior. Isso se explica pela diminuição da nota do entrevistador na composição da nota final. A nota final nova diminuiu, porque foi composta mais pela nota do observador, responsável por 66.34%, do que pela nota do entrevistador, com 33.67% na composição da nota.

Embora os pesos dos itens analíticos adequação lexical, gramatical e fluência tenham aumentado na nova proposta de composição da nota oral, o que provavelmente aumenta a dificuldade ou a probabilidade de os examinandos tirarem notas mais baixas, pela análise, a nota do entrevistador parece estar descrita de forma que seja difícil os examinandos serem classificados na faixa avançado superior. Segundo os descritores das faixas na grade do entrevistador (Anexo 1), o que diferencia a nota 4 da 5 é que a 5 "apresenta fluência e variedade ampla de vocabulário e de estruturas, com raras inadequações. Sua pronúncia é adequada" e 4 "apresenta fluência e variedade ampla de vocabulário e de estruturas, com inadequações ocasionais na comunicação. Sua pronúncia pode apresentar algumas inadequações", enquanto que em relação à autonomia, desenvoltura e compreensão os descritores são os mesmos. Parece haver uma tendência de o avaliador-entrevistador optar pelo nível avançado entre as faixas avançado e avançado superior, e, por isso, ao diminuir o peso da nota do entrevistador e aumentar a do observador, a quantidade de examinandos classificados na faixa avançado superior aumentou. Ou seja, aumentar o peso dos itens analíticos que se referem à aspectos linguísticos e aumentar o peso da nota do observador não significa necessariamente uma diminuição da nota final do examinando, porque o julgamento do entrevistador parece tender a concentrar a classificação dos examinandos na faixa avançado, quando em dúvida quanto à classificação entre avançado e avançado-superior. Dessa forma, ao diminuir o peso da nota do entrevistador na composição da nota final nova, os examinandos foram reorganizados de forma a aumentar o número de classificados na faixa avançado superior.

Considerações finais

O exame Celpe-Bras é uma avaliação de larga escala que tem como objetivo certificar a proficiência em língua portuguesa para falantes de outras línguas. O exame é composto por avaliações da proficiência oral e escrita. No presente trabalho, foram apresentados elementos para subsidiar a discussão sobre a relação entre nota-construto. Segundo Messick (1987), por meio da análise do significado da nota, pode-se levantar

evidências fortes sobre a validade de construto dos instrumentos de avaliação. Para tanto, apresentou-se uma análise fatorial, por ser mais ou menos consenso entre os estatísticos que o cálculo fatorial está relacionado a perguntas que têm como objetivo investigar o construto que está sendo medido por algum instrumento. Nesse sentido, a análise fatorial aqui apresentada foi eficiente ao gerar evidências sobre o que o teste mede e quanto cada item corresponde à medida, ou seja, o quanto cada item contribui para composição da nota final. Com base na análise fatorial, verificou-se que tanto a grade analítica quanto a do avaliador-interlocutor medem apenas um construto, o da proficiência oral. Isso quer dizer que o instrumento é unidimensional, ou seja, mede apenas uma coisa, a proficiência oral. Ou seja, a grade de avaliação é válida do ponto de vista do construto que ela pretende avaliar, embora a análise tenha apontado a necessidade de revisão de alguns itens.

Os valores de carga fatorial dos itens da escala variaram de 0.65 a 0.94, sugerindo que itens possam estar explicando o mesmo fator (proficiência oral). Os valores de peso para cada um dos itens com seus respectivos intervalos de confiança são, do maior para o menor: nota do entrevistador 0.36 (0.33-0.42); adequação lexical 0.19 (0.15-0.22); fluência 0.18 (0.15-0.22); adequação gramatical 0.13 (0.05-0.15); competência interacional 0.09 (0.07-0.11); pronúncia 0.06 (0.04-0.08); compreensão 0.04 (0.03-0.05). Na análise proposta, a nota analítica como um todo representa 66% da nota total da prova. O item compreensão se destacou na análise, por destoar ligeiramente quanto aos valores da análise fatorial, apontando a necessidade de se avaliar não só seu peso, mas também para se pensar até que ponto a compreensão oral está sendo avaliada na situação de prova proposta pelo Celpe-Bras. Parece ser razoável afirmar que o participante está sendo desafiado mais quanto à sua adequação lexical ou gramatical do que quanto à sua capacidade de compreender oralmente. Não seria a compreensão oral, no presente contexto da avaliação, um pré-requisito para que a interlocução aconteça?

A análise fatorial gerou informações para formulação de uma nova maneira de calcular a composição da nota final oral do exame. Ao aplicar os novos pesos para cada um dos parâmetros na composição da nota final, discutiu-se as implicações da mudança na composição da nova nota final a partir dos novos pesos no que diz respeito às possíveis mudanças de faixas de classificação do exame. Ao aplicar no mesmo conjunto de notas os pesos atuais e os propostos e compará-los à distribuição dos participantes em cada uma das faixas de classificação do exame, constatou-se que as notas analíticas da faixa avançado se distribuíram entre as demais faixas.

No que diz respeito à nota final, após a aplicação dos novos pesos, a mudança mais significativa foi a diminuição de participantes classificados na faixa avançado e aumento dos participantes em avançado superior. Como a porcentagem da nota do avaliador-interlocutor foi a que sofreu maior alteração, argumentou-se que a nota do avaliador-interlocutor possa tender a concentrar os participantes na faixa avançado. Ou seja, pela maneira como os descritores estão organizados, pode ser que seja difícil que um avaliador classifique o examinando na faixa avançado-superior. Seria interessante que outras pesquisas investigassem participantes com alto nível de proficiência com a finalidade de descrever seus desempenhos nas faixas avançadas para que se possa ter mais clareza de como a o que que está sendo avaliado. Além desta, outras análises, tanto

quantitativas quanto qualitativas, fazem-se necessárias para investigar a relação entre as grades analíticas e holísticas no que diz respeito à uma nova composição de pesos e suas implicações para classificação dos participantes.

Espera-se que os resultados desta pesquisa possam servir para fundamentar o argumento da validade da prova oral do exame Celpe-Bras e para refinar o processo de atribuição da nota oral.

Referências

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.

Recebido em: 06.06.2018
Aprovado em: 26.09.2018

Laura Márcia Luiza Ferreira é professora na Universidade Federal da Integração Latino-Americana (Unila), doutora pelo Cefet-MG, mestre em linguística e licenciada em letras pela UFMG. Foi leitora na Universidade de Chulalongkorn, na Tailândia, e professora em Timor-Leste. Pesquisa avaliações internas e externas, especialmente o exame Celpe-Bras.