ARTIGOS

articles

ANÁLISE DESCRITIVA DA PARTE ESCRITA DO EXAME CELPE-BRAS[1]

DESCRIPTIVE STATISTICS OF THE WRITTEN PART OF THE CELPE-BRAS EXAM

ANÁLISIS DESCRIPTIVO DE LA PARTE ESCRITA DEL EXAMEN CELPE-BRAS

Margarete Schlatter[2]

Luciana Neves Nunes[3]

Simone Paula Kunrath[4]

RESUMO

Este trabalho descreve o comportamento de quatro edições da Parte Escrita do exame que confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) em relação à distribuição das notas finais dos examinandos em cada edição e em cada uma das quatro tarefas que compõem as edições. Os dados analisados consistem nos relatórios de correção do exame que integram o banco de dados do Celpe-Bras mantido pelo Inep. A análise descritiva das notas aponta para distribuições regulares e equivalentes entre as edições e entre as tarefas de uma mesma edição. Assimetrias da distribuição das frequências das notas em determinadas tarefas, analisadas comparativamente dentro da mesma edição, sugerem diferenças de nível de dificuldade que requerem estudos qualitativos das características das tarefas e de seus possíveis impactos nos desempenhos dos examinandos. Tais análises poderão contribuir para a explicitação fundamentada do construto do exame e de como ele é operacionalizado na Parte Escrita, provendo, desse modo, argumentos para sua validação.

Palavras-chave: Celpe-Bras, exame de proficiência, análise descritiva, análise de tarefas.

1 Agradecemos ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) pelo acesso aos relatórios das edições analisadas neste artigo e ao Núcleo de Assessoria Estatística (NAE), Departamento de Matemática, Universidade Federal do Rio Grande do Sul (UFRGS), pelas análises estatísticas dos dados.

2 UFRGS, Porto Alegre, e-mail: margarete.schlatter@ufrgs.br

3 UFRGS, Porto Alegre, e-mail: lununes@mat.ufrgs.brm

4 UFRGS, Porto Alegre, e-mail: simone.kunrath@gmail.com

ABSTRACT

This paper describes the behavior of four editions of the Written Part of the exam that grants the Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras) in relation to the distribution of the examinees' final scores in each edition and in each of the four tasks that compose them. The data analyzed consist of the exam evaluation reports that integrate Inep's Celpe-Bras database. The descriptive statistics of the scores indicate regular and equivalent distributions between editions and between tasks in the same edition. Asymmetries in the distribution of the frequency of scores in certain tasks, analyzed comparatively within the same edition, suggest differences in their levels of difficulty that require qualitative studies of the task characteristics and how they impact on the examiners' performances. Such analyzes may contribute to grounded explicitation of the exam construct and how it is operationalized in the Written Part, thus providing arguments for the exam's validation.

Keywords: Celpe-Bras, proficiency exam, descriptive statistics, task analysis.

RESUMEN

Este trabajo describe el comportamiento de cuatro ediciones de la Parte Escrita del examen que confiere el Certificado de Competencia en Portugués para Extranjeros (Celpe-Bras) en relación con la distribución de las puntuaciones finales de los examinados en cada edición y en cada una de las cuatro tareas que componen las ediciones. Los datos analizados consisten en los informes de corrección del examen que forman parte de la base de datos del Celpe-Bras mantenida por Inep. El análisis descriptivo de las puntuaciones apunta a distribuciones regulares y equivalentes entre ediciones y entre tareas en la misma edición. Las asimetrías en la distribución de la frecuencia de las puntuaciones en ciertas tareas, analizadas comparativamente dentro de la misma edición, sugieren diferencias en el nivel de dificultad que requieren estudios cualitativos de las características de las tareas y de sus posibles impactos en el desempeño de los examinandos. Dichos análisis pueden contribuir a la explicitación razonada de lo constructo del examen y cómo se opera en la Parte Escrita, proporcionando así argumentos para su validación.

Palabras clave: Celpe-Bras, examen de competencia, análisis descriptivo, análisis de tareas.

1. Introdução

O Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras), ao longo de sua trajetória de quase 30 anos, consolidou-se como um exame de alta relevância, sendo requisito para ingresso nas universidades brasileiras[5], inscrição de profissionais estrangeiros em entidades de classe[6], naturalização[7] e outros contextos acadêmicos e profissionais. Criado em 1993 e aplicado pela primeira vez em 1998, o Celpe-Bras tornou-se um marco na área de Português como Língua Adicional por ser um exame pioneiro ao testar a proficiência das habilidades de compreensão e produção (oral e escrita) de modo integrado e por certificar quatro níveis de proficiência a partir de um único instrumento de avaliação. Considerando esse construto inovador e os usos do exame,

5 O Decreto nº 7.948, de 01/03/2013 (http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2013/Decreto/D7948.htm, acessado em 20/07/2020), em substituição ao protocolo interministerial PEC-G de março de 1998, regulamentou a mais antiga cooperação educacional do país, o Programa Estudante Convênio Graduação (PEC-G), e manteve a exigência do Celpe-Bras (nível Intermediário) para estudantes estrangeiros candidatos ao programa. O Celpe-Bras também é uma das possibilidades de comprovação de proficiência em língua portuguesa exigida para ingresso no Programa Estudante Convênio Pós-Graduação (PEC-PG), conforme Protocolo de 1981, atualizado em 2006 (http://www.dce.mre.gov.br/PEC/PG/Protocolo_PECPG.pdf, acessado em 20/07/2020).

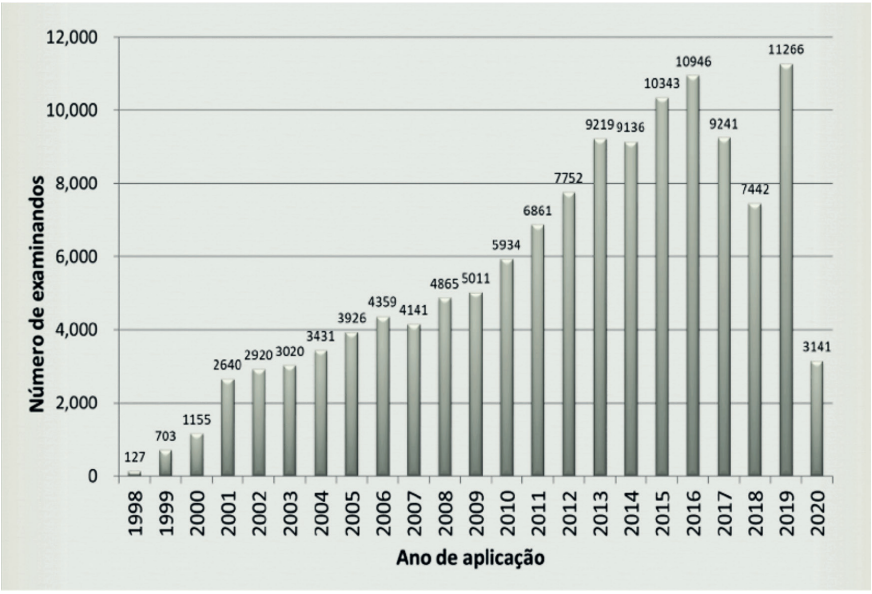
6 A Resolução nº 1.620, de 06/06/2001, do Conselho Federal de Medicina (CFM), instituiu, pela primeira vez, a exigência do exame para a inscrição de profissionais médicos estrangeiros nos Conselhos de Medicina; a Resolução nº 1.712, de 22/12/2003, estipulou o nível Avançado, que foi posteriormente alterado para nível Intermediário Superior pela Resolução nº 1.831, de 09/01/2008 (http://portal.cfm.org.br/images/stories/documentos/informe_juridico_89_15.pdf, acessado em 20/07/2020). A última decisão do CFM, Resolução 2.216, de 27/09/2018, publicada no Diário Oficial da União em 18/01/2019, dispõe que o cidadão estrangeiro cuja língua pátria não seja o português e cuja graduação em medicina não tenha sido realizada no Brasil deve comprovar o nível Intermediário do Celpe-Bras para obter o registro nos Conselhos Regionais de Medicina. (http://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/59626002/do1-2019-01-18-resolucao-2-216-de-27-de-setembro-de-2018-59625871, acessado em 20/07/2020).

7 A Portaria Interministerial nº 5 de 27/02/2018 instituiu o Celpe-Bras, em qualquer nível de certificação, como único comprovante da capacidade de se comunicar em língua portuguesa para aqueles que querem requerer a naturalização brasileira. Neste mesmo ano, essa decisão foi alterada através da Portaria Interministerial nº 16, de 03/10/2018, que acrescentou outras opções para comprovação do conhecimento em língua portuguesa além do Celpe-Bras. (http://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/43885878/do1-2018-10-04-portaria-interministerial-n-16-de-3-de-outubro-de-2018-43885761, acessado em 20/07/2020).

o Celpe-Bras tem sido considerado como um fator de impacto no ensino e na pesquisa sobre avaliação de proficiência no Brasil[8] e, gradativamente, tem se consolidado como um comprovante da capacidade de atuar em língua portuguesa em diferentes contextos.

Um reflexo dessa importância é o aumento da procura pelo certificado no cenário nacional e internacional, como mostra o número de examinandos com inscrições homologadas no gráfico a seguir.

FIGURA 1 - NÚMERO ANUAL DE EXAMINANDOS HOMOLOGADOS NO EXAME CELPE-BRAS



Fonte: MEC e Inep. In: www.ufrgs.br/acervocelpebras, acessado em 05/04/2021

Como podemos observar na figura acima, a primeira aplicação oficial do Celpe-Bras, em 1998, contou com 127 examinandos, que realizaram a prova

8 Para consulta sobre pesquisas acerca do exame Celpe-Bras que tratam de efeitos retroativos no ensino, na preparação de candidatos ao exame, na elaboração de materiais didáticos e de testes e na formação de professores, acesse

http://www.ufrgs.br/acervocelpebras/pesquisas/textos-publicados-sobre-o-exame-celpe-bras. Acessado em 16/05/2020.

em oito Postos Aplicadores (Schlatter *et al*, 2009). Atualmente, de acordo com o Documento Base do Exame Celpe-Bras (Brasil, 2020 p.14) (doravante DB), o Celpe-Bras conta com 126 postos aplicadores, 48 no Brasil e 78 no exterior, em países nos continentes americano, africano, asiático e europeu. O exame é aplicado duas vezes ao ano, e, nos anos de 2015, 2016 e 2019, o total de examinandos ultrapassou 10 mil, mantendo uma média de mais de 9 mil examinandos nos últimos 5 anos[9].

De acordo com Scaramucci (2006), os exames de proficiência são instrumentos eficientes de política linguística, exercendo influências em outras políticas, por exemplo, controlando o acesso à educação superior e ao trabalho, orientando o ensino e as inovações curriculares, promovendo mudanças na formação de profissionais da linguagem, entre outras. Além disso, como vimos acima, os resultados da avaliação são usados para tomar decisões que impactam na vida pessoal e profissional dos indivíduos. Por essas razões, os elaboradores de exames de alta relevância têm a responsabilidade de apresentar dados que possam esclarecer aos usuários quanto ao construto e aos procedimentos utilizados para permitir inferências válidas e confiáveis (Bachman; Palmer, 1996, p.95). De acordo com Kane (2012, p. 34), "a validação [de um teste] envolve uma avaliação da plausibilidade e da apropriação das interpretações e dos usos propostos para os resultados", o que requer "explicitar o que se está afirmando e avaliar a credibilidade dessas afirmações" com base em dados teóricos e empíricos[10].

Diferentes índices são relevantes para que se possa construir a validade e a confiabilidade de um exame, entre os quais a clareza dos objetivos e dos conteúdos da avaliação, a explicitação dos parâmetros utilizados para

9 Em 1998, 2018 e em 2020, houve apenas uma aplicação do exame Celpe-Bras. Em todos os outros anos, houve duas aplicações anuais.

http://www.ufrgs.br/acervocelpebras/pesquisas/textos-publicados-sobre-o-exame-celpe-bras. Acessado em 16/05/2020.

10 Chapelle (2012) apresenta um breve histórico das mudanças na concepção de validade, interpretada inicialmente como um teste para medir efetivamente o que se pretende, e depois como uma interpretação conceitual e empiricamente apropriada dos resultados do teste, evoluindo para o resultado de uma avaliação da utilidade do teste.

a elaboração dos testes e para a correção, a sistematização e transparência dos procedimentos de correção, a formação continuada de avaliadores, os estudos sistemáticos sobre as características do exame e os resultados. Visando contribuir com reflexões sobre os resultados do Celpe-Bras, este trabalho apresenta uma análise estatística da distribuição das notas da Parte Escrita de quatro edições do exame, analisando o comportamento geral de cada edição e, em cada uma, as quatro tarefas que a compõem. Segundo Brown (1996, p. 102), uma análise descritiva dos resultados é fundamental para que se possa "visualizar o comportamento médio (ou típico) do grupo como também a variação de desempenho dos examinandos" e para fornecer indícios para o aperfeiçoamento do exame em relação a, por exemplo, possíveis ajustes nos itens que o compõem e nos parâmetros de avaliação. Nesse sentido, com base na análise desenvolvida, levantamos algumas questões que podem subsidiar tomadas de decisão acerca das tarefas da Parte Escrita do Celpe-Bras.

2. Tipos de instrumentos de avaliação na área de linguagem e o Celpe-Bras

De acordo com Brown (1996), há dois tipos de instrumentos de avaliação educacionais usados na área de linguagem: os testes referenciados por critérios (TRC) e os referenciados pela norma (TRN). Os TRC são aqueles geralmente produzidos em contextos de ensino, para medir objetivos bem definidos e específicos (como os objetivos de um curso, de um programa ou de uma sequência didática), com vistas a avaliar o alcance da aprendizagem em relação aos conteúdos abordados. Os resultados são usados para embasar decisões sobre a promoção do aluno para o próximo nível, a necessidade de revisão de conteúdos ou de alterações no planejamento ou nas práticas de ensino. Como instrumentos usados para fins de diagnóstico (verificar o que os alunos já sabem sobre um determinado conteúdo antes de ensiná-lo) e para a avaliação da aprendizagem, a interpretação das pontuações de cada examinando é considerada em relação aos conteúdos e critérios adotados para medi-los, sem referência às pontuações dos outros examinandos. No uso de TRC no ensino, portanto, *critério* é entendido como o conteúdo que precisa ser aprendido e que será usado como parâmetro para medir, após o ensino, o alcance da aprendizagem (Brown, 1996, p. 3). Por outro lado, no enquadre

de exames de proficiência, tais como o Celpe-Bras, o termo *critério* é usado para se referir a um padrão de desempenho esperado para se passar ou alcançar um determinado padrão de desempenho em um teste, independentemente de quando, onde ou como esse conhecimento foi aprendido. Nesse caso, o *critério* será uma descrição de desempenho que possa ser usada como referência ou ponto de corte para a tomada de decisão para, por exemplo, "determinar se uma pessoa está qualificada para receber ou não um certificado" (Hussain *et al*., 2015, p. 28).

Já os TRN medem o desempenho de uma pessoa em relação ao desempenho de todos os participantes da avaliação, entendida como a *norma*; os resultados são, portanto, interpretados com referência às pontuações de outros participantes do teste, com vistas a "identificar o melhor candidato em comparação com os outros candidatos, e não determinar quantos dos candidatos atendem a um determinado padrão de desempenho" (Hussain *et al*., 2015, p. 26). Os TRN são usados, por exemplo, para se tomar decisões sobre a capacidade dos examinandos de ingressar numa universidade ou em uma atividade profissional, comparar resultados de diferentes instituições em relação a alguma habilidade global (leitura, produção escrita) ou nivelar examinandos para determinados programas de ensino. Esses propósitos de avaliação requerem distinguir indivíduos e colocá-los em uma escala de pontuação que expressa seu nível de proficiência em relação aos demais. Tais comparações são geralmente feitas com referência ao conceito de distribuição normal (conhecida como curva de sino). De acordo com Brown (1996),

> [...] o objetivo de um TRN é distribuir os examinandos em um contínuo de pontuações para que aqueles com habilidades "baixas" em uma habilidade global como leitura estejam em uma extremidade da distribuição normal, enquanto aqueles com habilidades "altas" estejam na outra extremidade (com a maior parte dos examinando distribuídos perto do centro da curva) (Brown, 1996, p. 2).

Isso não quer dizer que não haja critérios de acordo com os quais os desempenhos serão considerados mais e menos adequados. Nesse enquadre, no entanto, *critério* refere-se ao padrão, a uma descrição do que

seria o nível mínimo (ponto de corte) para alcançar esse padrão, e que será usado como parâmetro para avaliar cada desempenho. No Celpe-Bras, esse é o procedimento adotado para o ajuste das grades de avaliação em relação às respostas esperadas para cada tarefa da Parte Escrita. Antes de iniciar a correção das provas, "é preparada uma amostra aleatória e estratificada dos textos produzidos pelos participantes para cada tarefa, representativa dos postos aplicadores", com vistas a redigir, "a partir dela, as especificações para a avaliação" (Brasil, 2020, p.72). De acordo com o DB,

> Esse refinamento das especificações é extremamente importante para o processo de avaliação, visto que as expectativas dos elaboradores de uma tarefa em relação à compreensão por parte dos participantes nem sempre se concretizam. Além disso, é possível que participantes façam interpretações adequadas do enunciado ou dos textos-base que, entretanto, não haviam sido previstas no momento da elaboração da tarefa. (Brasil, 2020, p. 72).

Davidson (2012) alerta que a distinção entre TRC e TRN pode não ser tão clara. Mesmo que a perspectiva referenciada por critérios siga sendo avaliar o desempenho em relação a um conteúdo ou um padrão de desempenho específico, e a referenciada pela norma, a comparação dos examinandos entre si, em ambos os casos, as especificações dos instrumentos de avaliação têm sido cada vez mais detalhadas[11]. As especificações de um exame, de acordo com o autor, são descrições "[d]o desenho a partir do qual muitos itens ou tarefas equivalentes podem ser produzidos" (Davidson, 2012, p. 198). A prática de formulação das especificações (o que

11 Além disso, mesmo que, em TRC, o objetivo seja analisar o desempenho em relação a determinado critério, pode-se considerar a norma para ajustar as expectativas de alcance desses critérios, conforme explicado acima em relação ao ajuste da grade do Celpe-Bras. Por exemplo, se, numa avaliação de aprendizagem ou de proficiência, entre os examinandos, ninguém alcança o nível mais alto, isso pode ser um indício, por um lado, de que eles não estavam bem preparados ou, por outro lado, de que o instrumento é muito difícil, ou ainda de que os parâmetros estão além do que é possível alcançar e, portanto, precisam ser analisados e talvez ajustados. Da mesma forma, uma situação em que todos alcançam bons resultados também deve ser indicativa de uma análise do instrumento e dos parâmetros, pois o pressuposto de qualquer avaliação aplicada a vários participantes é que haja uma distribuição relativamente normal entre os desempenhos, com alguns examinandos apresentando resultados melhores do que outros.

pode incluir, por exemplo, lista dos conteúdos, descrição do formato do teste, exemplos de tarefas, desempenhos comentados, entre outros) é desejável não só para a construção de argumentos de validade do exame (pois é possível, por exemplo, avaliar a relação entre as especificações, as tarefas e os desempenhos dos examinandos), mas também para que os usuários possam conhecê-lo melhor e visualizar o que se pode inferir a partir dos resultados. Nesse sentido, as especificações de um exame são fundamentais para explicar os resultados de aprendizagem ou de proficiência, tanto para interpretar o que representa a posição alcançada no conjunto de examinandos quanto para compreender o que foi aprendido de determinado conteúdo. Além disso, como acrescenta Davidson (2012, p. 205), nos dois tipos de avaliação, diante de especificações detalhadas, o examinando terá condições de se autoavaliar e de preparar-se, ampliando as possibilidades de obter bons resultados no exame.

2.1 O CONSTRUTO DO CELPE-BRAS

O Celpe-Bras avalia o desempenho em relação ao uso da língua portuguesa em diferentes práticas de linguagem de que seu público-alvo necessita participar, no Brasil ou no exterior (Brasil, 2020 p. 26). De acordo com o DB (Brasil, 2020, p. 35 e 41), as habilidades de compreensão e produção de textos (orais e escritos) são testadas de modo integrado em duas etapas: uma Parte Escrita, realizada em, no máximo, 3 horas e composta por quatro tarefas de produção escrita a partir de um vídeo (compreensão multimodal), um áudio (compreensão oral) e dois textos escritos (leitura); e uma Parte Oral, composta por uma interação de 20 minutos entre avaliador e examinando, em que o examinando fala sobre si próprio e sobre assuntos da atualidade[12]. As principais características do Celpe-Bras são a ênfase no uso da língua a partir de textos autênticos que circulam na sociedade brasileira e a avaliação integrada da compreensão e da produção (oral e escrita), tendo em vista a participação em práticas sociais mediadas pela língua portuguesa (Brasil, 2020). Nesse sentido, o exame pode ser associado a uma avaliação que busca analisar o quanto o examinando transita em diferentes contextos orais e escritos, usando a língua portuguesa para participar de atividades complexas do mundo contemporâneo (Schlatter *et al*, 2009).

12 O conjunto de provas aplicadas desde 1998 pode ser acessado em http://www.ufrgs.br/acervocelpebras/acervo. Acessado em 20/07/2020.

Por meio de um único instrumento, o Celpe-Bras avalia seis níveis de proficiência, os dois primeiros, sem certificação e, para efeito de certificação, os níveis Intermediário, Intermediário Superior, Avançado e Avançado Superior. Diferentemente de outros exames de proficiência (ver, por exemplo, CAPLE [13] e DELE[14]), que delimitam a priori o que um examinando é capaz de fazer em cada nível e usam instrumentos específicos para cada um deles, os princípios que norteiam o Celpe-Bras reconhecem que as situações de interação social não são classificadas em níveis. De acordo com o DB,

> [...] o Celpe-Bras avalia distintos níveis de proficiência por meio de um exame único por reconhecer que as situações de interação social não são classificadas em níveis: o que distingue os níveis de proficiência são os recursos mobilizados pelo participante nas situações de interação propostas. Dessa forma, os enunciados das tarefas, assim como os textos que lhes servem de insumo, são iguais para todos os participantes. A diferença entre os níveis certificados espelha o desempenho do participante na realização das tarefas da Parte Escrita e na interação face a face da Parte Oral. (Brasil, 2020 p.33).

Como vimos, considerando as definições de testes referenciados por critérios e pela norma apresentadas anteriormente, o Celpe-Bras é um exame referenciado por critério: o desempenho do examinando é avaliado de acordo com padrões de desempenho que definem níveis de proficiência de uso da língua portuguesa. Em relação à interpretação dos resultados, cada examinando obtém uma nota que informa sua proficiência de acordo com a descrição dos diferentes níveis de desempenho avaliados (*critérios padrão que definem os pontos de corte de cada nível*). Embora os resultados do conjunto de examinandos possam ser dispostos uns em relação aos outros (em cada tarefa e nos resultados finais), o resultado relevante para fins de uso da certificação é ter alcançado determinado nível (e não sua posição relativa aos demais examinandos), conforme veremos em mais detalhe a seguir.

13 CAPLE - Centro de Avaliação e Certificação de Português Língua Estrangeira. Para maiores informações sobre os exames, acesse: https://caple.letras.ulisboa.pt/exames. Acessado em 02/07/2020.

14 D.E.L.E - Diploma de Espanhol como Língua Estrangeira. Para mais informações, acesse: https://www.dele.org/ Acesso em 02/07/2020.

2.2 AS FAIXAS DE CERTIFICAÇÃO DO CELPE-BRAS

Os resultados finais do Celpe-Bras são expressos pelo nome do nível, levando-se em conta a faixa de notas que os caracteriza (com intervalos de 0,75 a partir do nível Intermediário): Sem Certificação (0,0 a 1,99), Intermediário (2,0 a 2,75), Intermediário Superior (2,76 a 3,50), Avançado (3,51 a 4,25) e Avançado Superior (4,26 a 5)[15]. De acordo com o DB (Brasil, 2020 p.76), os índices de variação dentro de cada faixa resultam das notas obtidas em ambas as partes do exame. A Parte Oral é avaliada independentemente por dois avaliadores (avaliador-interlocutor e avaliador-observador) e, caso haja uma discrepância entre as notas atribuídas[16] , é reavaliada por uma nova dupla de avaliadores (e, se necessário, por uma terceira). Na Parte Escrita, o procedimento é o mesmo para cada uma das quatro tarefas, sendo que a reavaliação é feita por um terceiro avaliador. Então, o conjunto de produções (orais e escritas) de cada examinando é avaliado de forma independente por pelo menos 10 avaliadores (até 16, se houver discrepância em todas as notas, e podendo chegar a 18, caso houver necessidade de mais uma reavaliação da Parte Oral). Em cada uma das partes, é feita a média aritmética das notas obtidas, e o examinando será classificado em uma das faixas de notas. Caso seu desempenho seja diferente nas duas partes, a certificação atribuída será equivalente à nota mais baixa das partes (BRASIL, 2020 p.79), considerando que o certificado atesta a proficiência geral (e não pelas partes oral e escrita separadamente).

Sendo um exame único para avaliar e certificar diferentes níveis de proficiência, os resultados são classificatórios: o examinando alcança os critérios mínimos de desempenho padrão de determinado nível, sendo que possíveis variações dentro da faixa numérica não são relevantes para fins de certificação. Por exemplo, um examinando que alcança o nível Intermediário poderá ter obtido nota entre 2.0 e 2.75, respectivamente,

15 Para a descrição de cada nível certificado, confira o Documento base do exame Celpe-Bras (Brasil 2020, p. 67-68), disponível em: https://bit.ly/2O4o1p4. Acessado em 09/07/2020.

16 De acordo com o DB (Brasil, 2020), considera-se discrepância na Parte Escrita uma diferença maior que 1 ponto entre as duas notas originais (p. 73). Na Parte Oral, a discrepância é uma diferença entre as notas igual ou superior a 1,5 ou se houver "diferença acima de 2,0 pontos entre a nota da Parte Oral e a nota na Parte Escrita, desde que a nota final na Parte Escrita seja superior à nota na Parte Oral" (p. 76). Como a Parte Oral é avaliada independentemente com o uso de duas grades (uma holística e outra analítica), a reavaliação segue o mesmo padrão e, portanto, também é feita por uma dupla de avaliadores.

os pontos de corte inferior e superior nesse nível. Nesse enquadre, entendemos que uma análise da distribuição dos resultados, como a que fazemos neste estudo, para além de mostrar se o exame está de fato distinguindo desempenhos nos níveis avaliados, pode evidenciar comportamentos do instrumento de avaliação que mereceriam atenção. Nossa expectativa é que a análise descritiva empreendida neste estudo possa mostrar de que modo a Parte Escrita do Celpe-Bras se comporta em relação à distribuição dos desempenhos dos examinandos e, se houver assimetrias, indicar algumas questões de pesquisa com vistas a aprimorar o exame.

2.3 A AVALIAÇÃO DA PARTE ESCRITA

Foco deste estudo, a Parte Escrita do Celpe-Bras é avaliada de maneira holística, o que, de acordo com o DB (Brasil, 2020, p. 38-39), significa que vários aspectos relativos à produção escrita (adequação ao propósito e interlocução do gênero discursivo solicitado; recontextualização de informações; coesão e coerência; convenções da escrita; etc.) são considerados concomitantemente para se atribuir ao texto do examinando uma única nota (de 0 a 5), e não uma nota diferente para cada um dos aspectos que a compõem[17]. De forma independente, cada um dos avaliadores atribui, portanto, notas inteiras de 0 a 5 para a produção textual do examinando, levando em conta os padrões de desempenho dos diferentes níveis. Esses padrões são construídos a partir de parâmetros de avaliação comuns a todas as tarefas (enunciador, interlocutor, propósito, conteúdo informacional, recursos linguístico-discursivos), sendo que as especificações de cada tarefa são ajustadas a cada edição com base na leitura dos textos de uma amostra dos textos produzidos pelos participantes para cada tarefa, conforme mencionado anteriormente[18]. As notas finais de cada tarefa resultam da média das duas

17 Na Parte Oral, a produção é avaliada por meio de padrões de desempenho, holísticos e analíticos, compostos por seis critérios (compreensão oral, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia).

18 Em conformidade com protocolos para exames de alta relevância, os padrões de resposta da Parte Escrita são construídos durante a elaboração das provas e ajustados, em todos os níveis, pelos coordenadores das tarefas a partir de uma amostra representativa dos textos dos examinandos antes do início da correção. Feitos os ajustes e definidos os padrões das respostas (características dos textos de cada nível), é realizada a capacitação dos avaliadores por meio de uma simulação da avaliação com essa amostra. Nesse sentido, pode-se dizer que uma das críticas aos TRC apontada por Hussain et al. (2015, p. 29) de que "o processo de determinar níveis de proficiência e pontuação para aprovação pode ser altamente subjetivo

notas atribuídas. Em caso de discrepância, a média será calculada entre a nota do terceiro avaliador e a nota original mais próxima a essa (Brasil, 2020 p. 73). Caso a nota atribuída pelo terceiro avaliador seja equidistante às demais notas, a nota de reavaliação será considerada a nota final da tarefa (Brasil, 2020 p. 73).

Como já dissemos, nosso objetivo é descrever o comportamento de quatro edições da Parte Escrita do Celpe-Bras em relação à distribuição das notas finais do conjunto total dos examinandos em cada edição e em cada uma das quatro tarefas que a compõem. Com base nos resultados dessa análise, poderemos mostrar como os examinandos têm se distribuído nas diferentes faixas de proficiência na Parte Escrita nas quatro edições estudadas e se há alguma tarefa que se distingue das demais em cada edição. Os comportamentos observados poderão servir como indicadores de aspectos a serem analisados futuramente, de modo qualitativo, por exemplo, em relação a níveis de dificuldade de tarefas, a possíveis impactos de determinadas tarefas na nota final da Parte Escrita, entre outros.

3. Metodologia

O corpus deste trabalho é composto por todas as notas finais obtidas pela totalidade dos examinandos em quatro edições do exame. Os dados utilizados compõem o banco de dados do exame Celpe-Bras mantido pelo Inep[19] e consistem nos relatórios finais das correções das edições 2015-2, 2016-1, 2016-2 e 2017-1. O relatório final de cada edição do Celpe-Bras apresenta as notas relativas ao desempenho dos

e confuso" é confrontada por um procedimento que usa a norma para balizar e sustentar as decisões com vistas a representar, da melhor forma possível, cada nível, com desempenhos fundados no que efetivamente acontece, inclusive incorporando interpretações das tarefas e respostas consideradas adequadas e que não foram previstas pelos elaboradores da prova. Para mais detalhes do processo de avaliação da Parte Escrita, conferir Brasil, 2020, p. 72-74.

Esses e outros procedimentos adotados (equipes de avaliadores por tarefa, avaliadores com formação e experiência na área, duas avaliações independentes, controle de discrepâncias, monitoramento do desempenho dos avaliadores etc.) contribuem para garantir a confiabilidade dos resultados. (Brasil, 2020, p. 71-79).

19 Os dados foram disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – Inep a partir de solicitação cadastrada no Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC).

examinandos em todas as partes do exame e em todo o processo de correção. Os relatórios disponibilizados para este estudo trazem, portanto, por edição, as notas atribuídas aos examinandos por cada avaliador nas quatro tarefas da Parte Escrita, as notas do avaliador-observador e do avaliador-interlocutor da Parte Oral, as notas de reavaliação ou de consenso, no caso das avaliações que apresentaram notas com discrepância de dois pontos, as notas finais de cada tarefa e de cada parte do exame (Parte Escrita e Parte Oral), a nota de proficiência (nota final) e o respectivo nível de proficiência alcançado.

As análises desenvolvidas neste estudo foram feitas a partir das notas finais de cada tarefa por edição e das notas finais da Parte Escrita das edições de 2015-2, 2016-1, 2016-2 e 2017-1. O quadro a seguir apresenta o conjunto de dados que compõem o corpus por edição.

QUADRO 1. CONJUNTO DE DADOS DA PARTE ESCRITA POR EDIÇÃO

Edição	2015-2	2016-1	2016-2	2017-1
Nº de examinandos	*4471*	*4603*	*4729*	*3968*
Nº de notas finais da Parte Escrita	*4471*	*4603*	*4729*	*3968*
Nº de notas finais por tarefa da Parte Escrita	*4471*	*4603*	*4729*	*3968*
Nº total de notas atribuídas na T1	*9500*	*9798*	*10216*	*8517*
Nº total de notas atribuídas na T2	*9649*	*9923*	*10062*	*8377*
Nº total de notas atribuídas na T3	*9533*	*9997*	*10087*	*8417*
Nº total de notas atribuídas na T4	*9718*	*9986*	*10169*	*8526*
Nº total de notas atribuídas	*38400*	*39704*	*40534*	*33837*

Elaborado pelas autoras. FONTE: Banco de dados do Celpe-Bras cedido pelo Inep

Conforme mostra o quadro acima (Quadro 1), os dados analisados em cada edição são as notas finais da Parte Escrita e as notas finais de

cada tarefa da Parte Escrita da totalidade dos examinandos por edição: 4471 (2015-2), 4603 (2016-1), 4729 (2016-2) e 3968 (2017-1), somando um conjunto de 17771 dados computados para este estudo. A nota final da Parte Escrita[20] é a média aritmética das notas finais de cada tarefa. Conforme explicado anteriormente, as notas finais de cada tarefa resultam da média das notas atribuídas (duas avaliações independentes) e, em caso de discrepância, a média da terceira nota e a nota original mais próxima a essa (ou a terceira nota no caso desta ser equidistante às originais) (Brasil, 2020, p. 73). Para ilustrar esse processo, incluímos no quadro o número total de notas em cada tarefa por edição e o total de notas atribuídas naquela edição. Nas quatro edições analisadas, o número de provas reavaliadas na Parte Escrita correspondeu a uma média de 15% por tarefa. As notas finais da Parte Escrita foram utilizadas para comparar o comportamento das quatro edições, e as notas finais por tarefa, para analisar cada tarefa em comparação com as demais na mesma edição.

O processamento dos dados foi realizado pelo Núcleo de Assessoria Estatística (NAE)[21] do Departamento de Estatística da Universidade Federal do Rio Grande do Sul (UFRGS), que utilizou os softwares *RStudio* (versão 1.0.143) e *Statistical Package for the Social Sciences* - SPSS (versão 18.0). O estudo do comportamento das edições da Parte Escrita e das tarefas que as compõem foi feito a partir da distribuição das notas dos examinandos em cada edição e em cada tarefa, usando a análise descritiva, através de gráficos e medidas resumo, para visualizar o desempenho dos examinandos em relação à tendência central e dispersão (Divardin, 2011).

Para avaliar a suposição de que os dados seguem distribuição normal, foi utilizado o teste não paramétrico de Shapiro Wilk. Para a comparação entre as quatro edições em relação à variável nota final da Parte Escrita, foi utilizado o teste de comparações múltiplas, após a aplicação

20 Para a decisão sobre o nível de certificação obtido pelo examinando, as notas finais da Parte Escrita são cotejadas com as notas da Parte Oral, que não foram consideradas para a análise neste estudo.

21 Agradecemos aos estudantes Aline Foerster Grande, Filipe Renan Maracci, Juliana Sena de Souza, Renan Baiocco Pereira, sob a coordenação de Luciana Neves Nunes, coautora deste artigo, pelo tratamento dos dados e pela elaboração dos gráficos.

da Análise de Variância (ANOVA) não paramétrica de Kruskal-Wallis, indicada quando existe heterocedasticidade entre os grupos e suposição de normalidade violada. Em cada uma das edições, foi realizada a comparação entre as quatro tarefas através da Análise de Variância (ANOVA) não paramétrica de Friedman, indicada para amostras pareadas e complementada pelos testes de comparações múltiplas. Para todos os testes, o critério de decisão adotado foi nível de significância de 5%.

De forma descritiva, através de histogramas e da plotagem da curva normal no caso das distribuições da nota final da Parte Escrita, por edição, foram avaliadas as distribuições dos dados das notas quanto à sua forma, esperando-se que sigam uma distribuição aproximadamente normal, ou seja, que as distribuições tenham forma de sino e, portanto, haja maior concentração da frequência na parte central da distribuição. Considerando que um teste, como qualquer outro instrumento usado para medir, deve gerar comportamentos equivalentes a partir dos instrumentos utilizados em diferentes edições (se usados nas mesmas condições, com os mesmos objetivos e com grupos de participantes com perfis semelhantes) (Brown, 1996, p. 185), espera-se que os resultados sejam regulares e equivalentes entre as edições como também entre cada uma das tarefas de uma edição para outra caso haja intenção de distingui-las quanto às suas características e/ou nível de dificuldade. Se esse for o caso, os resultados podem indicar uma estabilidade desejada em exames de proficiência. Uma distribuição assimétrica indicaria que há maior frequência, ou seja, maior concentração de examinandos, em alguma das extremidades da distribuição e mereceria um estudo mais aprofundado[22].

Para a comparação das tarefas em cada uma das edições, foram feitos histogramas com o objetivo de se avaliar descritivamente o comportamento

22 Considerando as faixas de certificação do exame, uma distribuição aproximadamente normal coincidiria com os níveis Intermediário (2,0 a 2,75) e Intermediário Superior (2,76 a 3,50), ocupando a parte central, diminuindo gradativamente para as extremidades da curva: à esquerda, coincidindo com o nível Sem Certificação (0,0 a 1,99), e à direita, com os níveis Avançado (3,51 a 4,25) e Avançado Superior (4,26 a 5). Nesse sentido, por um lado, pressupondo-se instrumentos equivalentes, assimetrias na distribuição normal das notas poderiam indicar uma edição com examinandos mais (ou menos) proficientes. Análises mais detalhadas poderiam incluir outras variáveis, tais como tempo, estratégias e local de estudo da língua pelos examinandos, outras línguas de socialização, usos da língua, entre outros. Por outro lado, assimetrias também podem sinalizar a necessidade de estudos mais aprofundados sobre o próprio instrumento.

das tarefas. Nesse caso, avaliou-se a forma da distribuição de cada uma das variáveis com objetivo de se verificar, mesmo que descritivamente, o "grau de dificuldade" de cada uma das tarefas. A assimetria do gráfico remete à ideia de que se pode classificar a tarefa como "fácil" ou "difícil", de acordo com o lado da assimetria. Quando o gráfico mostra maior concentração de valores à esquerda, apresenta-se simetria positiva, ou seja, a tarefa parece ter sido difícil, pois há mais concentração em valores baixos. Já no caso de assimetria negativa, com maior concentração à direita, há indicação de que a tarefa parece ter sido fácil[23].

4. Análise descritiva dos resultados

Em primeiro lugar, apresentamos os resultados da distribuição das notas finais da Parte Escrita nas quatro edições do Celpe-Bras e, em seguida, a distribuição das notas nas quatro tarefas de cada edição separadamente, para analisar o comportamento da Parte Escrita e das tarefas quanto à forma e à dispersão. Conforme dito acima, foram realizados os testes de normalidade através do teste não paramétrico de Shapiro Wilk, que é indicado para testar aderência de um conjunto de dados a uma distribuição de probabilidade, neste caso, à distribuição normal. Entretanto, cabe observar que, neste estudo, em função do tamanho grande da amostra, o teste de aderência não teve um desempenho satisfatório, rejeitando a hipótese nula (distribuição aproximadamente normal) em todas as situações. Isso se justifica pela alta sensibilidade que esse teste tem com o tamanho da amostra. Mesmo que haja pequenos desvios da normalidade, o teste tende a rejeitar a hipótese de aderência. Portanto, para a análise de simetria das distribuições, optou-se pela análise descritiva, apresentada a seguir.

23 Usamos aspas em "grau de dificuldade", "fácil" e "difícil" para assinalar que os dados serão interpretados como indicadores de tais conclusões. A compreensão aprofundada do grau de dificuldade das tarefas demanda estudos futuros que considerem uma análise qualitativa dos enunciados das tarefas, dos textos de insumo utilizados e das produções dos examinandos, como também dos descritores construídos no ajuste das grades, que estabelecem o grau de exigência das respostas esperadas em cada uma das tarefas.

4.1 DISTRIBUIÇÃO DAS NOTAS FINAIS DA PARTE ESCRITA

Para se comparar as quatro edições quanto à nota final da Parte Escrita, foi realizada a ANOVA não paramétrica de Kruskal-Wallis, e se verificou que parece existir diferença entre as edições (p<0,001). Os resultados estão apresentados na tabela a seguir.

TABELA 1 - RESULTADO DA ANOVA DAS NOTAS FINAIS DA PARTE ESCRITA

Edição	N	Média (DP)
2015-2	4471	2,88 (0,85)
2016-1	4603	2,81 (0,88)
2016-2	4729	2,82 (0,77)
2017-1	3968	3,19 (0,78)

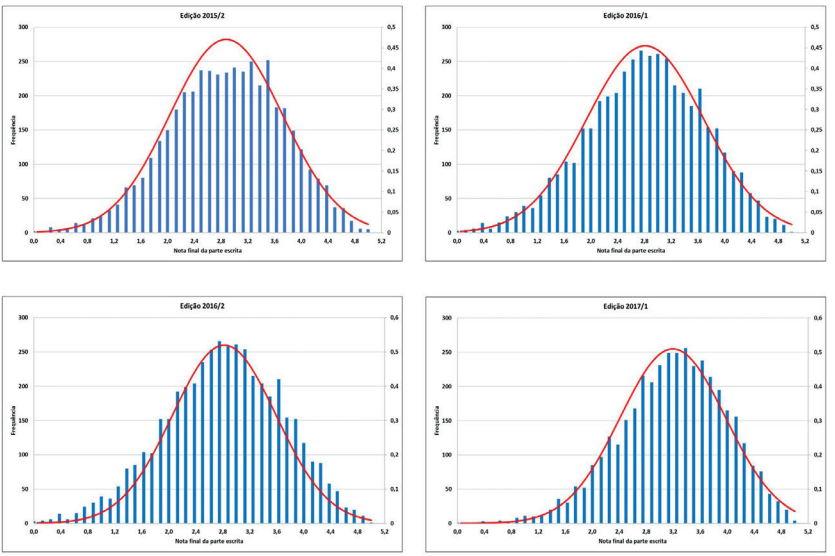
N 17771; p<0,001 | Elaborado pelas autoras

Conforme mostra a tabela 1, há diferença entre as edições (p<0,001). A partir da análise de comparações múltiplas, verificou-se que as edições 2016-1 e 2016-2, que tiveram, respectivamente, médias 2,81 (DP=0,88) e 2,82 (DP=0,77), não diferem entre si (p=1,000). Já a edição 2015-2, com média 2,88 (DP=0,85), e a edição 2017-1, com a maior média, 3,19 (DP=0,78), apresentaram-se significativamente diferentes entre si (p<0,01) e das demais edições (p<0,01). Cabe observar que as diferenças, embora sejam estatisticamente significativas, não são diferenças numéricas muito relevantes: todas as médias se encontram em uma mesma faixa de certificação (Intermediário Superior: 2,76 a 3,50), o que pode indicar que as faixas são bem delineadas. O resultado estatístico deve ser explicado pelo tamanho grande da amostra. Entretanto, foi possível observar que realmente a edição 2017-1 apresenta

valores significativamente mais altos que as demais edições, o que pode ser visualizado nos gráficos a seguir

A figura 2 apresenta a distribuição das notas finais da Parte Escrita em cada edição.

FIGURA 2 - DISTRIBUIÇÃO DAS NOTAS FINAIS DA PARTE ESCRITA POR EDIÇÃO. Valores de p referentes ao teste de Shapiro Wilk.



2015-2: N 4471; M 2,88; p<0,05; **2016-1:** N 4603; M 2,81; p<0,05;
2016-2: N 4603; M 2,82; p<0,05; **2017-1:** N 3968; M 3,19; p<0,05

Elaborado por: Núcleo de Assessoria Estatística (NAE) - Departamento de Matemática - UFRGS

A partir do teste de Shapiro Wilk, verificou-se que, para as quatro edições, a nota final da Parte Escrita parece não seguir distribuição aproximadamente normal (p<0,05). Entretanto, observando os gráficos, com exceção da edição 2017-1, vemos que as outras edições parecem ter uma distribuição regular da frequência das notas, formando uma distribuição em forma de sino, em que os dados se concentram próximo ao centro da escala e se distribuem para ambos os lados de forma regular, conforme proposto por Brown (1996). Na edição de 2015-2, por exemplo, podemos observar que o centro da curva

é mais largo, e a concentração das notas acontece em um espectro mais amplo. Isso mostra uma frequência de notas maior entre 2,3 e 3,5[24], que coincidem com os níveis Intermediário (2,0 a 2,75) e Intermediário Superior (2,76 a 3,50). As edições de 2016-1 e 2016-2 apresentam o ápice da curva mais acentuado e bem próximo ao centro da escala. Essas edições tiveram maior concentração de notas entre 2,4 e 3,4, equivalente aos níveis Intermediário (2,0 a 2,75) e Intermediário Superior (2,76 a 3,50). Já a edição de 2017-1 revela uma pequena assimetria negativa, ou seja, uma tendência para notas mais altas, entre 2,8 e 3,8, o que resultou em uma maior concentração na faixa do Intermediário Superior (2,76 a 3,50) e Avançado (3,51 a 4,25).

4.2 DISTRIBUIÇÃO DAS NOTAS POR TAREFA POR EDIÇÃO

Para se comparar as quatro tarefas quanto à nota final em cada uma das edições, foi realizada a ANOVA não paramétrica de Friedman e se verificou que parece existir diferença entre as tarefas (p<0,001) em todas as edições, conforme mostra a tabela 2, a seguir.

TABELA 2. ANOVA DAS NOTAS FINAIS DAS TAREFAS DA PARTE ESCRITA POR EDIÇÃO

		*Comparações múltiplas**				
		T1		T2	T3	T4
Edição	N	p	Média (DP)	Média (DP)	Média (DP)	Média (DP)
2015-2	*4471*	*<0,001*	*3,01 (0,95)a*	*2,77 (1,22)b*	*2,79 (1,03)b*	*2,93 (1,11)c*
2016-1	*4603*	*<0,001*	*2,91 (1,20)a*	*2,92 (1,07)a*	*2,85 (1,18)a*	*2,56 (1,04)b*
2016-2	*4729*	*<0,001*	*2,77 (0,99)a*	*2,88 (0,87)b*	*3,06 (1,05)c*	*2,56 (1,05)d*
2017-1	*3968*	*<0,001*	*3,61 (0,96)a*	*3,05 (1,02)b*	*3,10 (0,90)c*	*3,01 (1,06)b*

* Letras diferentes na linha indicam diferença significativa entre as tarefas em cada edição.

Elaborado pelas autoras

24 Considerou-se a faixa de notas que apresentaram frequência maior de 200 como parâmetro para todas as edições.

Quando utilizado o teste de comparações múltiplas das tarefas em cada uma das edições, viu-se que, na edição de 2015-2, não houve diferença entre as tarefas 2 e 3, que tiveram, respectivamente, médias 2,77 (DP=1,22) e 2,79 (DP=1,03) (p=1,000). A tarefa 4 teve média 2,93 (DP=1,11), a tarefa 1 teve a maior média, que foi 3,01 (DP=0,95), e elas se apresentaram significativamente diferentes entre si (p=0,001) e das demais edições (p<0,001). Cabe comentar que as diferenças, embora sejam estatisticamente significativas, não são diferenças numéricas muito relevantes. O resultado estatístico deve ser explicado pelo tamanho grande da amostra. Entretanto, foi possível observar que, aparentemente, a tarefa 1 apresenta valores significativamente mais altos que as demais tarefas na edição de 2015-2.

Na edição de 2016-1, viu-se que somente a tarefa 4 se diferenciou significativamente das demais (p<0,001), sendo que a sua média foi a menor entre todas as tarefas: 2,56 (DP=1,18). Não houve diferença entre as tarefas 1, 2 e 3, que tiveram, respectivamente, médias 2,91 (DP=1,04), 2,92 (DP=1,20) e 2,85 (DP=1,07). Já na edição de 2016-2, todas as tarefas se distinguiram significativamente entre si (p<0,001). As tarefas 1, 2, 3 e 4 tiveram, respectivamente, médias 2,77 (DP=0,99), 2,88 (DP=0,87), 3,06 (DP=1,05) e 2,56 (DP=1,05).

Na edição de 2017-1, parece não existir diferença entre as tarefas 2 e 4, que tiveram, respectivamente, médias 3,05 (DP=1,02) e 3,01 (DP=1,06) (p=1,000). A tarefa 1 teve a maior média, que foi 3,61 (DP=0,96), sendo significativamente diferente de todas as demais tarefas (p<0,001). A tarefa 3 teve média 3,10 (DP=0,90) e foi significativamente diferente das tarefas 2 e 4 (p<0,05).

4.2.1 DISTRIBUIÇÃO DAS NOTAS POR TAREFA - EDIÇÃO 2015-2

Para visualizar a distribuição dos resultados nas quatro tarefas por edição e as tendências apontadas na análise anterior, apresentamos a seguir os histogramas de cada edição que representam as distribuições que foram comparadas através da ANOVA não paramétrica de Friedman.

A figura 3, a seguir, apresenta os histogramas das quatro tarefas da edição 2015-2 do Celpe-Bras.

FIGURA 3 - HISTOGRAMAS DAS NOTAS FINAIS POR TAREFA DE 2015-2

N 4471; p<0,001

Médias: T1 = 3,01 (DP=0,95); T2 = 2,77 (DP=1,22); T3 = 2,79 (DP=1,03); T4 = 2,93 (DP=1,11) .

Fonte: Núcleo de Assessoria Estatística (NAE) - Departamento de Matemática - UFRGS

Conforme mostra o conjunto de histogramas, as tarefas 2 e 3 apresentam maior simetria se comparadas às demais. Dentre todas, a tarefa 3 apresenta a distribuição normal esperada, com uma frequência de notas maior na parte central da curva (entre as notas 2,0, 2,5 e 3,0), decrescendo em direção a cada uma das extremidades, com uma média de 2,79. Na tarefa 2, a distribuição das notas é mais dispersa, ou seja, as notas não apresentam um pico tão acentuado na parte central da curva, também distribuindo-se de modo similar em ambas as extremidades. Nesse caso,

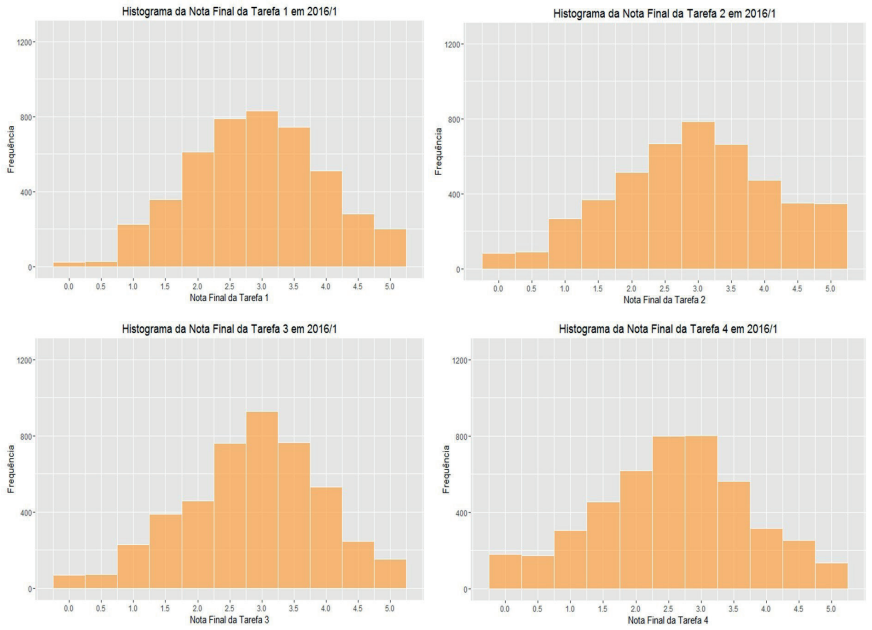
por um lado, ao mesmo tempo em que a frequência das notas 2,5, 3,0 e 3,5 (parte mais achatada no centro da curva) é mais próxima (n=691, 645 e 643, respectivamente), as notas da extremidade esquerda, de zero até 1,5, apresentam maior número de ocorrências, se comparadas às demais tarefas (n(0,0)=108, n(0,50)=124, n(1,0)=300, n(1,5)=471, total= 1003), o que poderia indicar que essa tarefa foi a mais difícil nessa edição, com uma média de 2,77.

A tarefa 1, por outro lado, parece ter sido a mais fácil se considerarmos que teve a maior média entre as tarefas desta edição (3,01) e uma baixa frequência de notas entre 0 e 1,5, com o ápice da curva mais acentuado na parte central da curva (notas entre 2,5, 3,0 e 3,5) e decrescendo para a extremidade direita (notas 4,51 e 5,0). Na tarefa 4, há uma leve assimetria negativa, com um número maior de ocorrências nas notas 4,5 e 5,0, se comparada às demais.

4.2.2 DISTRIBUIÇÃO DAS NOTAS POR TAREFA - EDIÇÃO 2016-1

Na edição de 2016-1, o teste de comparações múltiplas mostrou que as tarefas 1, 2 e 3 apresentaram um comportamento parecido, com médias muito próximas, 2,91, 2,92 e 2,85, respectivamente. Os histogramas dessas três tarefas mostram distribuições simétricas da frequência de notas, como podemos observar na figura a seguir. Já a tarefa 4, embora apresente uma distribuição razoavelmente simétrica, também se mostra mais dispersa. O gráfico apresenta uma frequência de notas mais alta para as notas 0,0 e 0,5 na extremidade esquerda, indicando que a tarefa 4 teve uma tendência para notas mais baixas, com a média mais baixa entre as quatro tarefas (2,56).

FIGURA 4 - HISTOGRAMAS DAS NOTAS FINAIS POR TAREFA DE 2016-1



N 4603; p<0,001

Médias: T1 = 2,91 (DP=1,20); T2 = 2,92 (DP=1,07); T3 = 2,85 (DP=1,18); T4 = 2,56 (DP=1,04)

Fonte: Núcleo de Assessoria Estatística (NAE) - Departamento de Matemática - UFRGS

Destaca-se ainda, nesta edição, a distribuição com assimetria negativa na tarefa 2, com um índice mais alto de ocorrências na extremidade direita (notas 4,5 e 5,0) do que as demais tarefas, tendência que resultou na média 2,92.

4.2.3 DISTRIBUIÇÃO DAS NOTAS POR TAREFA - EDIÇÃO 2016-2

Na edição de 2016-2, todas as tarefas se apresentaram distintas entre si, conforme mostram os histogramas a seguir.

FIGURA 5 - HISTOGRAMAS DAS NOTAS FINAIS POR TAREFA DE 2016-2

N 4603; p<0,001

Médias: : T1 = 2,77 (DP=0,99); T2 = 2,88 (DP=0,87); T3 = 3,06 (DP=1,05); T4 = 2,56 (DP=1,05)

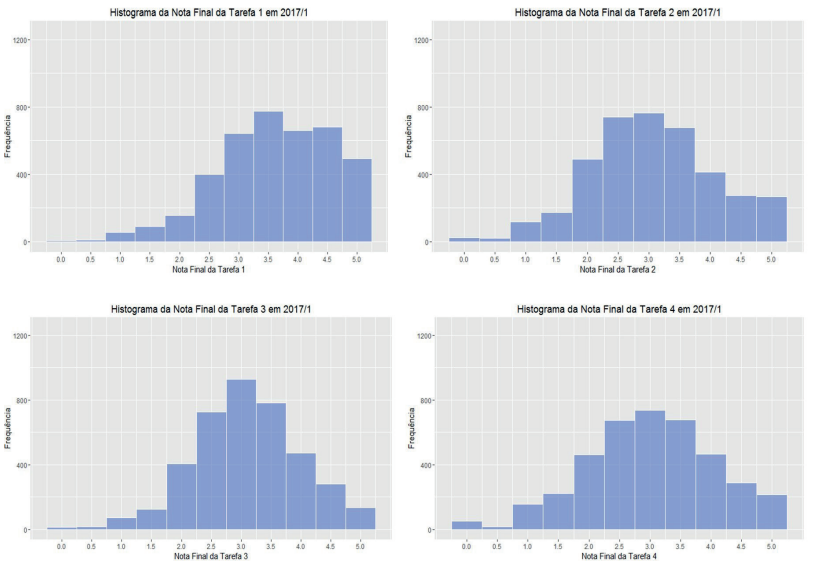
Fonte: Núcleo de Assessoria Estatística (NAE) - Departamento de Matemática - UFRGS

Observa-se, na tarefa 1, uma leve assimetria positiva em que a maior frequência de notas se dá entre 2,0, 2,5 e 3,0, apresentando, assim, uma leve tendência a notas mais baixas se comparada às tarefas 2 e 3. Os histogramas das tarefas 2 e 3 mostram uma leve assimetria negativa com uma concentração de notas mais acentuada para a extremidade direita. Em ambas, a maior frequência se dá entre as notas 2,5, 3,0 e 3,5, na parte central da curva. A tarefa 2, por um lado, com média de 2,88, destaca-se pela concentração atípica de notas 3,0, indicando uma dispersão menor do que as outras tarefas. A tarefa 3, por outro lado, destaca-se por uma maior frequência de notas 4,5 e 5,0, elevando a média para 3,06, a mais alta entre as quatro tarefas, indicando, assim, um nível de dificuldade menor nesta edição. A tarefa 4, por sua vez, apresentou uma distribuição dispersa das notas, sem um padrão a ser referido. Observa-se uma alta frequência das notas 1,0 e 1,5 e uma tendência para as notas mais baixas, deslocando a nota de maior frequência para a nota 2,5. Esse resultado indica que a tarefa 4 foi a mais difícil, com uma média de 2,56.

4.2.4 DISTRIBUIÇÃO DAS NOTAS POR TAREFA - EDIÇÃO 2017-1

As distribuições das notas de todas as tarefas da edição de 2017-1 tiveram médias mais altas se comparadas às outras edições analisadas, apresentando maior frequência entre as notas 3,0 e 3,5, como podemos observar nos histogramas a seguir.

FIGURA 6 - HISTOGRAMAS DAS NOTAS FINAIS POR TAREFA DE 2017-1



N 3968; p<0,001

Médias: T1 = 3,61 (DP=0,96); T2 = 3,05 (DP=1,02); T3 = 3,10 (DP=0,90); T4 = 3,01 (DP=1,06)

Fonte: Núcleo de Assessoria Estatística (NAE) -Departamento de Matemática- UFRGS

Destaca-se, nesta edição, a acentuada assimetria negativa da tarefa 1, a qual tende fortemente para a direita, apresentando um alto índice de ocorrências nas notas 4,0, 4,5 e 5,0, com a média mais alta entre as tarefas (3,61). O histograma da tarefa 3 evidencia uma distribuição simétrica com concentração de notas no centro da curva, mais especificamente na nota 3,0, indicando uma dispersão menor para as extremidades. A média desta

tarefa foi de 3,10. As tarefas 2 e 4 apresentam uma leve tendência para notas mais altas, elevando as frequências das notas nas extremidades direitas e resultando nas médias de 3,05 e 3,01, respectivamente. Essa tendência a valores mais altos em todas as tarefas da edição 2017-1 fez com que a média da nota final da edição se elevasse significativamente (3,19, conforme tabela 1), diferentemente das outras edições que vinham apresentando média em torno de 2,83.

5. Discussão dos resultados

Conforme vimos, de acordo com Brown (1996), uma análise descritiva dos resultados de uma avaliação permite mostrar o comportamento típico e a variação de desempenho de um grupo de examinandos, além de fornecer dados para uma avaliação do próprio instrumento. Considerando que as quatro edições do Celpe-Bras analisadas envolvem um grupo grande de participantes, uma distribuição normal dos resultados poderia indicar edições equivalentes. Assimetrias (positivas ou negativas), por sua vez, podem indicar aspectos do exame a serem estudados, explicados e possivelmente ajustados. Como apresentado na seção anterior, há alguns comportamentos das edições analisadas que recomendariam algum estudo. Apresentamos, a seguir, uma síntese desses resultados, levantando algumas questões para reflexão.

Considerando os resultados referentes às notas finais nas quatro edições da Parte Escrita do Celpe-Bras, observou-se que, com exceção da edição 2017-1, a distribuição da frequência de notas se concentrou próximo ao centro da escala e se distribuiu para ambos os lados de forma regular. Mesmo que a edição de 2017-1 tenha apresentando uma assimetria negativa em relação às demais (o que poderia indicar uma edição menos exigente[25]), na análise do conjunto das tarefas nas quatro edições, as eventuais assimetrias de algumas tarefas em edições específicas (como veremos a seguir) se dispersaram e não influenciaram no desempenho das edições como um todo. Esses resultados indicam que a avaliação da

25 Novamente chamamos a atenção que os resultados com base unicamente na distribuição das notas dos examinandos são indicadores de níveis de dificuldade. Conforme já dito, estudos qualitativos seriam necessários para confirmar e explicar diferentes níveis de dificuldade nas edições do exame e nas tarefas.

Parte Escrita do Celpe-Bras, nas quatro edições analisadas, seguiu o padrão de desempenho esperado, mostrando-se bastante regular e equivalente.

Também pode-se concluir que, em todas as edições, a Parte Escrita distinguiu o desempenho de examinandos menos e mais proficientes. Em todas as edições, o ponto mais alto da curva foi ocupado pelas notas 2,8 (2016-1 e 2016-2), 3,4 (2017-1) e 3,5 (2015-1), o que coincide com a faixa referente ao nível Intermediário Superior (2,76 a 3,50), desempenho no centro da escala dos níveis avaliados pelo exame. A distribuição em direção às extremidades, do mesmo modo, foi bastante regular, tanto para a esquerda, nível Sem Certificação (0 a 1,99) e Intermediário (2,0 a 2,75), como para a direita, Avançado (2,76 a 4,25) e Avançado Superior (4,26 a 5,0). A classificação da maioria dos examinandos nos níveis Intermediário e Intermediário Superior, que variou de 63% a 70%, nas quatro edições, pode ser considerada como desejável para os usos desse exame, por exemplo para obtenção de cidadania e ingresso na vida acadêmica e profissional (ver notas de rodapé 1 a 3).

Na edição de 2017-1, como vimos, houve uma assimetria à direita e uma média significativamente mais alta (3,19 (DP=0,78), comparada a 2,88 (DP=0,85), 2,81 (DP=0,88) e 2,82 (DP=0,77) nas edições 2015-2, 2016-1 e 2016-2, respectivamente). Se, por um lado, isso poderia significar uma inconsistência em relação ao nível de dificuldade do exame entre as edições, por outro lado, poderia indicar um grupo mais preparado naquele ano ou, ainda, um comportamento avaliativo diferenciado em relação às edições anteriores. Caberia, pois, uma análise qualitativa das características dessa edição (especificações das tarefas e ajustes dos parâmetros da grade de avaliação) em relação às demais e um estudo comparativo das tarefas dessa edição com as de uma outra, todas respondidas por um mesmo grupo de participantes, para compreender ou descartar possíveis diferenças nos instrumentos.

Na comparação das tarefas em cada uma das edições, o objetivo foi analisar descritivamente o comportamento de cada uma delas no conjunto das quatro tarefas que compõem a edição. Considerou-se a distribuição das notas, as médias e os desvios padrão (ANOVA não paramétrica de Friedman) para indicar possíveis diferenças relacionadas aos graus de

dificuldade. Ao passo que, na edição 2016-1, as quatro tarefas tiveram comportamentos muito semelhantes (com uma indicação de a tarefa 4 ser levemente mais difícil), na edição 2016-2, todas as tarefas se comportaram de modo distinto entre si (com uma indicação de a tarefa 3 ser mais fácil e a 4, mais difícil). Se comparadas as quatro edições, não foi possível estabelecer um padrão em relação a uma tarefa que tenha sido sempre a mais (ou menos) difícil, mas foi possível levantar algumas tendências. Se considerarmos as assimetrias e médias de cada tarefa por edição, a tarefa mais fácil variou entre a tarefa 1 (2015-2 e 2017-1) e a tarefa 3 (2016-1 e 2016-2); e a mais difícil foi a tarefa 4 em três edições (2016-1, 2016-2, 2017-1) e a tarefa 2, em uma edição (2015-2).

Os comportamentos acima sugerem algumas questões para reflexão e possíveis estudos. Uma das questões refere-se à adoção de diferentes graus de dificuldade entre as tarefas do exame: é objetivo do sistema Celpe-Bras propor tarefas de diferentes dificuldades? Caso afirmativo, quais características da tarefa podem estar influenciando diferentes comportamentos? A modalidade dos textos de insumo (áudio, vídeo, impresso) é relevante? O gênero discursivo a ser produzido é relevante? Em um amplo estudo de descrição dos elementos que compõem as tarefas da Parte Escrita do exame Celpe-Bras (edições de 1998 a 2017), Schoffen *et al.* (2018, p. 72), foi mostrado, por exemplo, que, em relação ao que é proposto como produção textual, a tarefa 4 possui um perfil mais estável e características diferentes das outras três tarefas. O relatório de pesquisa mostrou que os gêneros do discurso mais recorrentes na tarefa 4 (do texto a ser produzido) são *carta/e-mail, carta do leitor* e *artigo de opinião*, e que *posicionar-se* é o propósito de maior ocorrência, enquanto que, nas outras tarefas, mesmo tendo como gênero e propósito mais recorrentes *carta/e-mail* e *divulgar,* há uma ampla variedade de opções[26]. Os resultados também apontaram que

26 Por exemplo, identificou-se que 72,5% das tarefas 4 tiveram como propósito posicionar-se, já divulgar ocorreu em 33% das tarefas 1, em 26% das tarefas 2 e em 17,7% das tarefas 3 (Schoffen et al, 2018, p. 48 e 65 a 70).

> [...] a relação entre compreensão (vídeo, áudio e leitura) e produção escrita pode estar sendo proposta de diferentes maneiras: ao passo que, nas tarefas de vídeo, áudio e na tarefa III, o examinando é, em geral, solicitado a retextualizar as informações do texto de insumo para reorganizá-las de modo a produzir outro gênero do discurso, com outros propósitos e outra relação de interlocução; na tarefa IV, as informações do texto de insumo devem ser usadas como base para a construção de argumentos para sustentar uma tomada de posição sobre determinado assunto. (Schoffen et al, 2018, p. 72)

Nesse sentido, caberia uma análise qualitativa de todas as tarefas deste estudo, estabelecendo relações entre a distribuição da frequência de notas e as características das tarefas no intuito de contribuir para construir justificativas fundamentadas para diferentes graus de dificuldade das tarefas que integram compreensão (áudio, vídeo, leitura) e produção escrita.

Em um estudo qualitativo de 50 textos de examinandos classificados como Intermediário Superior e Avançado Superior na edição de 2016-1, Kunrath (2019) distinguiu aspectos relacionados à recontextualização de informações do texto de insumo e ao uso de recursos linguístico-discursivos que foram produtivos para a distinção dos níveis. Entre os fatores levantados está a organização do conteúdo em um texto argumentativo:

> De acordo com Crowhurst e Piche (1979) e Beauvais et al (2011), no texto argumentativo, como o artigo de opinião, é o autor que organiza o conteúdo do seu texto de forma a convencer o interlocutor de sua opinião. A produção desse texto requer uma organização elaborada do conteúdo através do uso de estratégias de transformação do conhecimento complexo e sofisticado a fim de tornar o texto relevante e convincente. Desse modo, a escolha das informações do texto de insumo dependerá totalmente do autor/examinando, que selecionará as informações que considerar mais relevantes para a construção da sua argumentação e para refutar os argumentos levantados por possíveis opositores. E será a adequação e a relevância das informações a favor e contra mobilizadas no texto de insumo que determinarão o nível de compreensão e de produção escrita do examinando no momento da avaliação. (Kunrath, 2019 p. 121)

As conclusões acima combinadas com as indicações de maior dificuldade da tarefa 4 em três edições e da coincidência de nível de dificuldade da tarefa 3 (a outra tarefa que integra leitura e escrita) com as demais apontam para a possibilidade de diminuir o número de tarefas da Parte Escrita para três (em vez de quatro), contribuindo para uma avaliação mais enxuta e de igual consistência em termos de distinção dos níveis de proficiência dos examinandos. Por um lado, entendendo que, para distinguir níveis de proficiência no uso da língua portuguesa, seria desejável ter tarefas de níveis de dificuldade distintos e também com diferentes combinações de habilidades, no caso de reduzir para três tarefas, a implicação dos resultados aqui descritos seria a de retirar a tarefa 3 da Parte Escrita. O exame manteria, assim, três tarefas de produção de texto integradas com compreensão multimodal (texto em vídeo), compreensão oral (texto em áudio) e leitura (texto impresso).

Por outro lado, uma possível retirada dessa tarefa teria que ponderar também as diferenças de níveis de dificuldade encontradas entre as tarefas 3 e 4 em todas as edições analisadas aqui, pois elas podem estar indicando distinções entre tarefas integradas de leitura e escrita relevantes de serem testadas e que efetivamente distinguem níveis de proficiência. Com base na análise de 44 produções de texto em resposta a quatro tarefas de leitura escrita, relacionando o desempenho dos participantes com as características dos enunciados das tarefas e os descritores das grades de avaliação, Gomes (2009) corroborou estudos anteriores que entendem a complexidade como um atributo da relação entre tarefa, participante e avaliação. Os resultados apontaram para alguns fatores no enunciado da tarefa e nos parâmetros de avaliação que impactaram nas notas dos examinandos: os papéis interlocutivos menos/mais antagônicos, a exigência de defesa de ideias menos/mais abstratas, materializadas no uso de linguagem menos/mais complexa (p. 98), e a maior/menor clareza na orientação quanto aos propósitos e à interlocução do texto a ser escrito (p. 99). Nesse sentido, poderia-se argumentar que manter duas tarefas de leitura e escrita que explicitamente propusessem uma distinção de complexidade na sua elaboração poderia ser interessante para permitir inferências sobre níveis de proficiência em diferentes aspectos de leitura e de escrita. Nesse caso, no entanto, seria necessário sistematizar e consolidar as características das tarefas 3 e 4 para que suas especificações ficassem mais claras.

Também caberia analisar mais detalhadamente os resultados referentes às assimetrias mostradas pelos histogramas. Do total de 16 tarefas analisadas, mesmo que em intensidades variadas, 5 apresentaram assimetrias negativas (tarefas 1 e 4, 2015-2; tarefa 2, 2016-1; tarefa 3, 2016-2 e tarefa 1, 2017-1), e somente uma resultou em uma assimetria levemente positiva (tarefa 1 da edição 2016-2). Esse resultado indica que, quando a curva se apresenta assimétrica, há uma tendência de os examinandos terem notas mais altas. Por um lado, isso poderia ser um indicador desejável de que muitos examinandos que se candidataram ao exame estavam preparados para a obtenção de certificação, principalmente nos níveis Intermediário (2,0 a 2,75) e Intermediário Superior (2,76 a 3,50). Por outro lado, como afirma Brown (1996, p. 50), entre os estudos desejáveis de exames de alta relevância, está a análise da qualidade das tarefas, o que inclui uma verificação de conteúdo e formato, nível de dificuldade e de seu potencial de discriminar diferentes desempenhos. Considerando a importância da padronização desses aspectos para informar tanto os elaboradores do exame como também os examinadores, cabe novamente a recomendação de uma análise das características das diferentes tarefas, para verificar se seria possível sistematizar índices que pudessem prever resultados semelhantes em futuras edições do exame.

In'nami e Koizumi (2016) destacam que nem sempre é fácil distinguir a causa da variação no desempenho dos examinandos. Segundo os autores, isso pode ocorrer devido a algum aspecto da tarefa em si, bem como do avaliador (ou da avaliação). Os resultados das análises descritivas apresentadas aqui dão indícios sobre questões que poderiam ser estudadas de modo mais aprofundado no intuito de justificar tanto a normalidade dos resultados como as assimetrias observadas com vistas a sistematizar índices preditivos para a construção de tarefas de níveis diferentes que integram compreensão oral ou leitura com produção escrita, se essa for a intenção do sistema avaliativo Celpe-Bras. Um objetivo alternativo poderia ser a busca por resultados uniformes entre as tarefas de uma mesma edição. Nesse caso, seria importante ampliar os estudos para analisar correlações entre níveis de dificuldade e desempenhos de examinandos, buscando compreender até que ponto tarefas indicadas como mais difíceis podem pressupor notas mais altas dos mesmos examinandos nas mais fáceis. Antes disso, no entanto, é preciso que os comportamentos das tarefas apresentados aqui sejam descritos qualitativamente e então

testados de modo sistemático para se chegar a índices mais precisos quanto aos níveis de dificuldade.

6. Considerações finais

Considerando que os resultados de uma avaliação de alta relevância, como é o caso do exame Celpe-Bras, são usados para tomar decisões sobre a vida pessoal, acadêmica e profissional dos indivíduos, entendemos como fundamentais não só a publicação de informações sobre o construto, o formato e os procedimentos adotados no sistema de avaliação, mas também o desenvolvimento de pesquisas que analisem resultados do exame no intuito de contribuir para a construção do argumento de validade e a confiabilidade do exame para os usos a que se destina. Como vimos, o processo de validação de um teste "envolve uma avaliação da plausibilidade e da apropriação das interpretações e dos usos propostos para os resultados do teste" (Kane, 2012, p. 34). A análise descritiva dos resultados da Parte Escrita do Celpe-Bras de quatro edições do exame que apresentamos neste estudo pode contribuir com dados empíricos sobre o instrumento de avaliação, na medida em que explicita o comportamento geral de cada edição e em cada uma das quatro tarefas que a compõem.

Mostramos, nas análises e na discussão dos dados, que há evidências de um comportamento bastante estável entre as edições, o que é um resultado positivo para a confiabilidade do exame. No entanto, também há alguns aspectos que merecem estudos mais detalhados quanto às características das tarefas e de que modo podem ser combinadas em cada edição para desempenhar funções específicas de, por um lado, testar habilidades diferenciadas e, por outro, de apresentar níveis de dificuldade diferentes. Entendemos que um construto que privilegia o uso da língua, ações coordenadas de compreensão e produção visando a participações em determinadas práticas sociais propostas no exame, e os níveis de proficiência graduados a partir de um único instrumento foram aspectos que caracterizaram o Celpe-Bras como um exame inovador na época em que foi criado. Foram também esses aspectos que tornaram o exame um fator motivador de mudanças em propostas de ensino e maneiras de aprender a língua portuguesa que priorizassem a interação com textos autênticos e em contextos diversos.

Após uma trajetória de quase de 30 anos, várias pesquisas já mostraram a consistência teórico-prática das tarefas e dos procedimentos de correção e descreveram impactos do exame no ensino e na preparação dos examinandos (cf. http://www.ufrgs.br/acervocelpebras/pesquisas). De acordo com Brown (1996, p. 69), tanto o desenvolvimento quanto o aperfeiçoamento de sistemas de avaliação de alta relevância demandam procedimentos-chave como pilotar as tarefas em um grupo semelhante ao grupo-alvo de examinandos, analisar as tarefas qualitativamente (em relação às especificações do exame) e quantitativamente (com testes estatísticos), e selecionar os melhores exemplares para compor uma edição revisada, enxuta e eficaz do instrumento. Nesse sentido, entendemos como fundamental fortalecer o campo de pesquisa envolvendo procedimentos estatísticos para produção de resultados e de possíveis relações dessas análises com características do exame (por exemplo, para estabelecer índices preditivos quanto ao desempenho esperado e ao nível de dificuldade da tarefa), desempenhos e perfis de examinandos, e usos do exame.

Para Chapelle (2012, p. 30), o desafio da validação de um exame é a explicitação continuada e robusta "dos valores subjacentes ao seu design e aos seus usos", o que requer a construção de uma cadeia argumentativa que "explicite as conexões dos desempenhos do examinando aos resultados do processo avaliativo, às interpretações desses resultados, às decisões que serão feitas a partir dessas interpretações e, finalmente, às consequências que resultarão dessas decisões". Esperamos que a análise apresentada aqui possa inspirar trabalhos futuros sobre essas relações. Tais pesquisas podem trazer subsídios para tomadas de decisão mais fundamentadas tanto acerca da calibragem entre construto e especificações do exame como da elaboração e do aprimoramento do próprio instrumento de avaliação, e, consequentemente, sobre a plausibilidade e a apropriação dos usos dos resultados do exame propostos e atualmente em vigor.

Referências

BACHMAN, Lyle F.; PALMER, Adrian S. *Language Testing in Practice*: Designing and Developing Useful Language Tests. Oxford: Oxford University Press, 1996.

BEAUVAIS, Caroline; OLIVE, Thierry; PASSERAULT, Jean-Michel. Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology,* 103 (2): 415-428, 2011.

BRASIL. *Documento base do exame Celpe-Bras* [recurso eletrônico]. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2020.

BROWN. James Dean. *Testing in Language Programs*. New Jersey: Prentice Hall Regents, 1996.

CHAPELLE, Carol A. Conceptions of validity. In: FULCHER, Glenn; DAVIDSON, Fred (eds.). *The Routledge Handbook of Language Testing*. Routledge: London; New York, 2012. p. 21-33.

CROWHURST, Marion; PICHE, Gene L. audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English*, 13 (2): 101-109, 1979.

DAVIDSON, Fred. The specifications and criterion referenced assessment, In: FULCHER, Glenn; DAVIDSON, Fred (eds.). *The Routledge Handbook of Language Testing*. Routledge: London; New York, 2012. p. 197-207.

DIVARDIN, Gisele W. *Elaboração e validação de um modelo padrão de avaliação para exames de proficiência de leitura em inglês para ingressantes em programas de pós-graduação na UTFPR - Campus de Ponta Grossa*. Tese de Doutorado, UFPR, 2011.

GOMES, Maíra S. *A complexidade de tarefas de leitura e de produção escrita no exame Celpe-Bras*. Dissertação de Mestrado, UFRGS, 2009.

HUSSAIN, Shafaat; TADESSE, Tessema; SAJID, Sumaiya. Norm-Referenced and Criterion-Referenced Test in EFL Classroom. *International Journal of Humanities and Social Science Invention*, 4 (10): 24-30, 2015.

IN'NAMI, Yo; KOIZUMI, Rie. Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, 33 (3): 341-366, 2016.

KANE, Michael. Articulating a validity argument. In: FULCHER, Glenn; DAVIDSON, Fred (eds.). *The Routledge Handbook of Language Testing*. Routledge: London; New York, 2012. p. 34-47.

KUNRATH, Simone P. *Os descritores gerais e a progressão dos níveis de proficiência do exame Celpe-Bras*. Tese de Doutorado, UFRGS, 2019.

SCARAMUCCI, Matilde V. R. O professor avaliador: sobre a importância da avaliação na formação do professor de língua estrangeira. In: ROTTAVA, L.; SANTOS, S.R. (orgs.) *Ensino-aprendizagem de Línguas*: Língua Estrangeira. Coleção Linguagens, Ijuí: Editora da UNIJUI, 2006. p. 49-161.

SCHLATTER, Margarete; SCARAMUCCI, Matilde V. R.; PRATI, Silvia; ACUNA, Leonor. Celpe-Bras e CELU: Impactos da construção de parâmetros comuns de avaliação de proficiência em português e espanhol. In: ZOPPI FONTANA, Mónica. (Org.). *O português do Brasil como língua transnacional*. Campinas, SP: Editora RG, 2009, p. 95-122.

SCHOFFEN, Juliana Roquele; SCHLATTER, Margarete; KUNRATH, Simone Paula; NAGASAWA, Ellen Yurika; SIRIANNI, Gabrielle Rodrigues; MENDEL, Kaiane; TRUYLLIO, Luana Ramos; DIVINO, Luiza Sarmento. *Estudo descritivo das tarefas da Parte Escrita do exame Celpe-Bras:* edições de 1998 a 2017. [recurso eletrônico] Porto Alegre: Instituto de Letras - UFRGS, 2018.

Artigo submetido em 04/08/2020

Aprovado em 08/03/2021