Letras

Revista do Programa de Pós-Graduação em Letras
Universidade Federal de Santa Maria

n. 1

Artigo Original

Letras, v. 34, n. 1, p. 01-21, e89188, Edição Especial 2025 - DOI:10.5902/2176148589188 - ISSN 2176-1485
Submissão: 30/09/2024 Aprovação: 30/09/2024 Publicação: 04/07/2025

Linguística Aplicada à análise de testes e exames

Confiabilidade entre avaliadores em foco: o processo de atribuição de notas e a reavaliação na Parte Escrita do Celpe-Bras

Inter-rater reliability in focus: the rating process and re-assessment in the written part of Celpe-Bras

Giovana Lazzaretti Segat

Universidade Federal do Rio Grande do Sul

Juliana Roquele Schoffen

Universidade Federal do Rio Grande do Sul

Ana Beatriz Âreas da Luz Fontes

Universidade Federal do Rio Grande do Sul

Resumo: Este artigo apresenta discussões sobre a confiabilidade entre avaliadores na Parte Escrita do Exame Celpe-Bras. A partir de dados de três edições do exame (2016.1, 2016.2 e 2017.1), o estudo pauta o consenso e a discrepância entre avaliadores durante o processo de avaliação, bem como o impacto da reavaliação de textos com notas discrepantes na definição da nota final da Parte Escrita e do nível de certificação. Os dados e análises apresentados fornecem evidências para atestar a confiabilidade entre avaliadores no processo de avaliação da Parte Escrita do Celpe-Bras.
Palavras-chave: Celpe-Bras; Parte Escrita; Confiabilidade

Abstract: This article discusses the inter-rater reliability in the Written Part of the Celpe-Bras Exam. Based on data from three editions of the exam (2016.1, 2016.2, and 2017.1), the study addresses both consensus and discrepancies among raters during the rating process, as well as the impact of re-assessing texts with discrepant scores on the final grade for the Written Part and the certification level. The data and analyses presented provide evidence to attest to the inter-rater reliability in the assessment process of the Written Part of the Celpe-Bras.
Keywords: Celpe-Bras; Written Part; Reliability

Introdução

A avaliação ocupa papel central nas práticas educacionais, uma vez que referencia o currículo, relaciona-se com a elaboração de materiais didáticos, com as práticas de ensino e de aprendizagem, e possibilita (ou não) o acesso a algumas esferas sociais, como cursos de graduação ou de pós-graduação em universidades. Em razão dessa teia de conexões, McNamara (2004, p. 763) afirma que a avaliação de línguas é uma das áreas centrais da Linguística Aplicada. Para além dos impactos dentro do campo de estudos, a avaliação de línguas nomeadas tem um papel muito importante em questões políticas e sociais. McNamara (2004, p. 764) afirma que essas avaliações "têm marcado sua relevância social no mundo contemporâneo, pois desempenham um papel em processos institucionais e políticos que são socialmente significativos"; por essa razão, as avaliações precisam ser levadas a sério e praticadas com responsabilidade e ética por todos os seus participantes.

Neste artigo, tomamos como objeto de estudo o exame Celpe-Bras. O Certificado de Proficiência em Língua Portuguesa para Estrangeiros, Celpe-Bras, é um exame brasileiro que busca aferir proficiência em língua portuguesa a partir da avaliação integrada de habilidades de uso da língua. Dada a relevância do exame e seus contextos de uso, é fundamental que o teste tenha comprometimento ético e seriedade frente aos processos de avaliação e aos resultados gerados por ele. Ter acesso a informações e dados sobre o exame, bem como a análises quantitativas e qualitativas sobre a confiabilidade, a validade, os impactos, a equidade e as demais características importantes dos testes é, portanto, fundamental e do interesse de todos os envolvidos.

Almejando contribuir com uma pequena etapa para que esse fim seja alcançado, neste artigo, apresentamos alguns resultados de uma pesquisa que analisou processos e dados que demonstram a confiabilidade entre avaliadores do Celpe-Bras, com foco na Parte Escrita do exame. Consideramos que os processos de avaliação e de reavaliação de textos com notas discrepantes, como feitos atualmente, são indicadores de confiabilidade entre avaliadores. Dentro do escopo da Linguística Aplicada, descrevemos o processo de avaliação e reavaliação da Parte Escrita do Exame Celpe-Bras, discutimos a noção de confiabilidade no campo de estudos da avaliação e, mais especificamente, descrevemos o consenso e a discrepância – e, portanto, a reavaliação – entre avaliadores nas tarefas da Parte Escrita do exame Celpe-Bras nas duas edições aplicadas em 2016 (2016.1 e 2016.2) e na primeira edição de 2017 (2017.1). Para tanto, organizamos o artigo da seguinte maneira: inicialmente, apresentamos o Celpe-Bras e damos especial atenção às características da Parte Escrita do exame. Na sequência, topicalizamos a confiabilidade e a confiabilidade entre avaliadores, central para a discussão conduzida neste artigo. Destinamos a seção Atribuição de nota, consenso e discrepância entre avaliadores para a apresentação dos dados e a análise. Esta se desdobra em subseções destinadas a cada uma das

edições analisadas e ao cruzamento dos dados descritos. Após esse cruzamento, procedemos à etapa de considerações finais, propondo reflexões sobre o percurso e encaminhamentos para outras análises possíveis.

O Exame Celpe-Bras

Criado com o objetivo de avaliar a proficiência de estudantes candidatos ao Programa de Estudantes-Convênio de Graduação (PEC-G), o Celpe-Bras foi aplicado pela primeira vez em 1998 e vem sendo realizado duas vezes ao ano desde então, em postos aplicadores no Brasil e no exterior. Atualmente, ele é utilizado para fins de ingresso em cursos de graduação e pós-graduação em universidades brasileiras, para a revalidação de diplomas no Brasil e como uma das opções para o processo de naturalização de estrangeiros junto ao Ministério da Justiça.

O Celpe-Bras avalia as habilidades de compreensão e de produção oral e escrita de maneira integrada, utilizando um único instrumento para aferir diferentes níveis de proficiência, uma vez que se considera que todos os examinandos são capazes de desempenhar ações em língua portuguesa, cada um no seu nível de proficiência, da mesma forma como acontece nas situações de uso da língua no cotidiano (Brasil, 2013). O exame é, portanto, um teste de desempenho (Brasil, 2020) em que o examinando é solicitado a produzir textos por escrito e oralmente a partir da leitura e da compreensão oral em duas etapas: Parte Escrita e Parte Oral.

As principais características do Celpe-Bras são a ênfase no uso da língua, o uso de textos autênticos e a avaliação integrada da compreensão e da produção (oral e escrita) (Brasil, 2020). Nesse sentido, é possível dizer que a avaliação 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). Assim, a proficiência é sempre definida localmente, por ser situada em contextos de uso, em determinada prática social (Brasil, 2020, p. 27).

A Parte Escrita do Celpe-Bras é composta por quatro tarefas diferentes que avaliam as habilidades de compreensão oral, leitura e produção escrita de modo integrado. Já a Parte Oral do Celpe-Bras acontece nos turnos e/ou dias subsequentes à aplicação da Parte Escrita. Segundo o Documento Base do Celpe-Bras, "a Parte Oral do Exame procura refletir um fenômeno constitutivo das relações interpessoais: a interatividade" (Brasil, 2020, p. 41). Essa parte busca aferir o desempenho do examinando quanto a sua compreensão e produção oral em uma interação de 20 minutos com o avaliador. A primeira etapa, que utiliza os 5 minutos iniciais, é desenvolvida com base em informações pessoais do examinando, geralmente disponibilizadas a partir do formulário de inscrição. Já a segunda etapa se vale de Elementos Provocadores (EPs), como são chamados os materiais utilizados para apoiar a interação na Parte Oral, para gerar oportunidades de diálogo durante os 15 minutos subsequentes.

É importante ressaltar que o exame atribui uma nota final para a Parte Escrita e uma nota final para a Parte Oral. A partir disso, a nota final do examinando no Celpe-Bras é estabelecida pela nota mais baixa entre as duas partes do exame, isto é, essa nota determinará a certificação atribuída ao examinando (Brasil, 2020, p. 79). São 4 os níveis de proficiência certificados no exame: Intermediário, Intermediário Superior, Avançado e Avançado Superior.

Figura 1 – Faixa de notas dos níveis de certificação do Celpe-Bras

Nível	Faixa de notas
Sem Certificação	De 0,00 a 1,99
Intermediário	De 2,00 a 2,75
Intermediário Superior	De 2,76 a 3,50
Avançado	De 3,51 a 4,25
Avançado Superior	De 4,26 a 5,00

Fonte: Brasil (2020, p. 79)

Para receber a certificação, portanto, o examinando precisa alcançar pelo menos o nível intermediário nas duas partes do exame. Essa escolha do Celpe-Bras é baseada no entendimento de que, em virtude de se certificar as quatro habilidades conjuntamente, o nível final deve ser conferido contemplando o desempenho do examinando como um todo, de acordo com o que ele demonstra ser capaz de fazer tanto na Parte Escrita quanto na Parte Oral.

A Parte Escrita do Celpe-Bras

A Parte Escrita do Celpe-Bras tem papel importante na definição do nível de proficiência dos examinandos. Conforme Segat (2023), nas edições de 2016.1, 2016.2 e 2017.1, respectivamente 84,4%, 79,1% e 71,5% dos examinandos tiveram suas notas finais e o nível de certificação determinado pela nota obtida na PE. Composta por quatro tarefas diferentes que avaliam as habilidades de compreensão oral, leitura e produção escrita de modo integrado, a parte escrita organiza-se conforme a figura abaixo, alternando a integração entre compreensão oral/imagética e produção escrita e a integração entre leitura e produção escrita. O examinando dispõe de 3 horas para realizá-la.

Em consonância com o seu construto teórico, o Celpe-Bras optou, desde sua primeira edição, por realizar uma avaliação holística na Parte Escrita. Fazer uma avaliação holística significa atribuir uma única nota para um texto baseado em uma leitura mais ampla da produção. A decisão por esses parâmetros se sustenta no entendimento de que o texto produzido pelo examinando é um enunciado único e irrepetível, que precisa ser avaliado em sua singularidade (Schoffen, 2009).

Figura 2 – Tarefas da Parte Escrita do Celpe-Bras

Tarefas	Habilidades envolvidas	Tempo total
1	Compreensão oral e imagética (vídeo) + produção escrita	3h
2	Compreensão oral (áudio) + produção escrita	
3	Leitura + produção escrita	
4	Leitura + produção escrita	

Fonte: Brasil (2020, p. 35)

Como os construtos teóricos que subjazem aos exames são abstratos, é importante explicitar os conceitos que os sustentam para que, posteriormente, possa ser averiguada a sua operacionalização. A proficiência em língua portuguesa, no Celpe-Bras, é avaliada pelo desempenho dos examinandos em tarefas que "(...) pressupõe a familiaridade com diferentes práticas de letramento de que participam cidadãos escolarizados." (Brasil, 2020, p. 30). As tarefas visam criar oportunidades de ação no mundo em diferentes situações sociais (Schlatter et al., 2009), com base no conceito de uso da linguagem como uma ação conjunta dos participantes com um propósito social (Clark, 1996) e na noção de gênero do discurso (Bakhtin, 2003), que fornecem os critérios para a avaliação, a partir das situações de comunicação propostas nas tarefas. Entendendo o uso da linguagem como uma prática social situada, cada tarefa pressupõe uma resposta que construa adequadamente uma determinada relação de interlocução em um determinado gênero discursivo, implicando que sejam considerados, no momento da avaliação, os seguintes aspectos: enunciador, interlocutor, propósito, informações, organização do texto, recursos linguísticos (gramática e vocabulário).

Outro conceito fundamental é o de tarefa integrada, que busca proporcionar um contexto mais autêntico de uso da língua durante a avaliação e alinha-se ao construto teórico do exame, em especial à ideia do que é proficiência para o Celpe-Bras. Segundo Knoch e Sitajalabhorn (2013),

> tarefas de escrita integrada são tarefas nas quais os examinandos são apresentados a um ou mais materiais de insumo linguisticamente robustos e são solicitados a produzir textos escritos que exigem (1) localizar ideias nos textos de insumo, (2) selecionar ideias, (3) sintetizar ideias de um ou mais textos de insumo, (4) transformar a língua usada no insumo, (5) organizar ideias e (6) usar convenções estilísticas como conectar ideias e referenciar os insumos (Knoch; Sitajalabhorn, 2013, p. 306).

Mendel (2019) aponta que a integração de habilidades corresponde a um construto teórico orientado ao letramento acadêmico, uma vez que, para a realização de tarefas escolarizadas/acadêmicas, muitas vezes mais de uma habilidade é, de fato, mobilizada. Nesse sentido, as tarefas integradas em avaliações de proficiência têm sido apontadas como mais representativas do tipo de escrita esperada nas práticas de determinados contextos e, portanto, mais autênticas, bem como dotadas de potencial para exercer efeitos retroativos positivos no ensino.

Compreender a integração de tarefas não é apenas fundamental para os examinandos, que necessitam ter noções mínimas de como o exame é proposto para apresentar um bom desempenho na prova, mas também para os avaliadores, que precisam considerar essa especificidade no momento da atribuição de notas. Ao propor tarefas integradas, "o Celpe-Bras se configura como um dos poucos [exames] cujo instrumento consiste, em sua totalidade, em uma avaliação de desempenho integrada das habilidades" (Mendel, 2019, p. 72). Para que o construto teórico seja devidamente operacionalizado, contudo, é importante que, para além das outras etapas do exame (elaboração de tarefas, discussão sobre diretrizes do exame, contexto de aplicação), ele também seja considerado ao longo de todo o processo de avaliação da Parte Escrita.

As características descritas anteriormente são fundamentais pois representam uma operacionalização de um construto teórico. Exames que avaliam a língua em uso apresentando tarefas integradas de leitura e escrita, que consideram uma série de fatores para além dos linguísticos, terão, conforme Bachman e Palmer (1996, p. 135), mais dificuldades de manter altos níveis de confiabilidade. Wang et al. (2017) confirmam que os desafios são grandes para os exames que utilizam tarefas integradas, como o Celpe-Bras, já que "os avaliadores devem avaliar tanto a expressão escrita dos alunos, quanto suas habilidades de leitura, a fim de incorporar informações relevantes dos textos de insumo em seus textos" (Wang et al., 2017, p. 36). Nesse sentido, Bachman e Palmer (1996) apontam a importância de operacionalizar os construtos teóricos de forma que sejam coerentes ao que está sendo aferido, para que as qualidades do teste, como validade e confiabilidade, por exemplo, possam ser discutidas.

Confiabilidade

Ainda que historicamente muitos conceitos e práticas da avaliação linguística sejam provenientes da psicometria e do campo da educação de maneira mais geral (*Standards for Educational and Psychological Testing/AERA/APA/NCME*, 2014), na década de 90 houve uma mudança de paradigma e os estudos de avaliação começaram a ocupar um espaço na Linguística Aplicada. Isso foi essencial para que houvesse uma teorização mais coerente com as práticas de avaliação linguística. Dois autores fundamentais nessa virada foram Bachman e Palmer, que propuseram um framework de avaliação composto por seis qualidades que, juntas, caracterizam a utilidade do teste; são elas: confiabilidade, validade de construto, autenticidade, interatividade, impacto e praticidade. Mais recentemente, pesquisadores apontam a necessidade de se considerar, também, a equidade (Yan; Fan, 2022).

Neste trabalho, damos especial atenção à confiabilidade, ainda que haja a compreensão de que essas seis características operam de maneira conjunta e se influenciam. A confiabilidade, conforme descrita pelos autores, diz respeito à constância e à acurácia dos processos do teste

que se refletem no resultado da avaliação. Ainda que não seja inerente ao teste, a confiabilidade é fundamental para a interpretação de resultados e para a realização de inferências sobre eles. Em resumo, a pergunta central para a confiabilidade é o quão estáveis e precisos os processos e os resultados de uma avaliação podem ser. A confiabilidade preocupa-se, portanto, em identificar o(s) erro(s) na avaliação ou a(s) fonte(s) de variância, contrastando-os com a constância e a acurácia dos processos do teste.

São muitas as possíveis fontes de variância (ou erros de mensuração) que afetam a confiabilidade de uma avaliação. Como aponta Chapelle (2013),

> esse erro pode vir de fatores como condições difíceis para ouvir durante o teste de compreensão oral, fadiga por parte dos examinandos, instruções pouco claras em uma seção de um teste, tarefas mal escritas, rubricas de pontuação inadequadas em um teste de escrita ou simplesmente uma avaliação inadequada do desempenho da amostra. (Chapelle, 2013, p. 4919).

McKay e Plonsky (2021, p. 468) apontam que há erro(s) em todas as mensurações feitas, e certos tipos (e quantidades) de erro são esperados. Por essa razão, é esperado que uma avaliação apresente variância indesejada e inesperada, desde que esta não seja suficientemente relevante para interferir na confiabilidade dos resultados. Há que se cuidar, portanto, dos erros sistemáticos que podem ser identificados e mitigados, visando a manutenção da confiabilidade e das outras qualidades do teste.

Em exames de proficiência, especialmente os aplicados em larga escala, quatro fontes de erro sistemáticos são elencadas como principais por Yan e Fan (2022), quais sejam: a situação (local, preparação para/do estudante), a variação individual (saúde, motivação, atenção), o avaliador (leniência, treinamento) e o instrumento de avaliação (conteúdo do teste, elaboração de tarefas, questões técnicas). Conforme os autores, enquanto as duas primeiras fontes de erro são mais difíceis de controlar ou explicar sistematicamente, a terceira e a quarta fontes podem ser vistas afetando, potencialmente, a própria natureza da habilidade que está sendo testada, impactando tanto na confiabilidade quanto na validade do teste (Yan; Fan, 2022, p. 483). McKay e Plonsky (2021) afirmam que os linguistas aplicados devem fazer uma distinção entre os tipos de confiabilidade e as diferentes formas de aferi-la. Nesse sentido, discorremos, a seguir, sobre a confiabilidade entre avaliadores, que é o foco desta análise.

Confiabilidade entre avaliadores

Por considerar a atenção dada à temática nas pesquisas da área e por entender que os examinadores têm papel fundamental no processo de avaliação e de manutenção da confiabilidade dos resultados do exame, optamos por enfocar neste estudo a confiabilidade entre avaliadores. Conforme Davis (2016, p.117), "a qualidade das decisões da atribuição de notas do avaliador tem consequências importantes para a confiabilidade e a validade do teste" (Davis, 2016, p. 117).

Nesse sentido, a confiabilidade entre avaliadores indica até que ponto os pares de avaliadores (ou mais) podem chegar a um consenso e atribuir notas consistentes em suas avaliações sobre os mesmos examinandos (Yan; Fan, 2022, p. 481). Para mensurar a confiabilidade entre avaliadores, utilizamos a estimativa de consenso, que verifica a porcentagem de notas iguais que dois avaliadores diferentes atribuem para a mesma produção. De acordo com Yan e Fan (2022), "às vezes, se for difícil alcançar um alto nível de concordância exata entre os avaliadores, as pontuações adjacentes também serão contadas como consenso (por exemplo, em uma escala de cinco pontos, as pontuações atribuídas pelos dois avaliadores estão separadas por apenas um ponto)" (Yan; Fan, 2022, p. 482). Esse é o caso do Celpe-Bras: para o exame, os avaliadores não chegam a um consenso apenas quando uma mesma nota é atribuída por dois avaliadores diferentes a uma mesma produção, mas também quando atribuem notas adjacentes, com diferença de apenas um ponto entre elas.

Tendo em vista que a confiabilidade tem, como interesse central, identificar o(s) erro(s) na avaliação ou a(s) fonte(s) de variância, contrastando-os com a constância e a acurácia dos processos do teste, algumas medidas podem ser adotadas visando mitigar os efeitos desses possíveis erros nos resultados da avaliação. Para evitar maiores problemas associados, principalmente, a diferenças de interpretação ou de demais influências não desejadas em decorrência da prática dos avaliadores, McNamara (1996) propõe que, a fim de minimizar os efeitos da discordância que possam levar a erros de mensuração, sejam implementados: (i) o uso de descritores formulados para cada nível de avaliação, incluindo exemplos das características do desempenho de cada nível; (ii) treinamento dos avaliadores no uso das rubricas e na execução dos procedimentos de avaliação; e (iii) avaliação de cada examinando mais de uma vez e adoção de procedimentos para lidar com as possíveis discrepâncias de notas (McNamara, 1996, p. 117). Como aponta Neves (2018), esses procedimentos "podem auxiliar na diminuição da subjetividade, para que esta não seja uma forte fonte de erro de mensuração" (Neves, 2018, p. 70).

No Celpe-Bras, a formação dos avaliadores, que inclui instrução sobre o uso de descritores formulados para cada nível de avaliação e exemplos de textos avaliados em cada nível, e a prática de reavaliação dos textos que receberam notas discrepantes são procedimentos adotados a fim de garantir a confiabilidade dos resultados do exame. Além deles, conforme consta no Documento Base (2020), há também o "monitoramento do desempenho dos avaliadores com o auxílio de relatórios diários" (Brasil, 2020, p. 72). Dessa forma, no contexto do Celpe-Bras, cabe ao coordenador da avaliação de cada tarefa, responsável por orientar e conduzir o grupo de examinadores alocados em seus grupos, acompanhar a atribuição de notas e trabalhar a fim de lidar com situações diversas ao longo da correção, auxiliando, assim, na manutenção da acurácia do processo.

Atribuição de nota, consenso e discrepância entre avaliadores

Para entender como a formação dos avaliadores e o processo de atribuição de notas, materializados no consenso e na discrepância que podem encaminhar (ou não) à reavaliação, fomentam a discussão sobre a confiabilidade, destinamos as próximas seções para a descrição desses processos e para a análise de dados provenientes de alguns processos de correção. Inicialmente, explicamos como funciona a atribuição de notas na PE do exame; na sequência, caracterizamos o consenso e a discrepância e passamos para a análise dos dados de três edições do Celpe-Bras. Destacamos que alguns aspectos serão explicitados na descrição da primeira edição analisada e que, nas subsequentes, será feita apenas rápida menção, considerando que os procedimentos, tanto do exame quanto da análise, foram os mesmos. Por fim, apresentaremos algumas considerações provenientes do cruzamento dos dados descritos e das análises realizadas.

Atribuição de nota

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão governamental responsável pelo Celpe-Bras, contrata, por meio de licitação, uma empresa para executar e implementar os procedimentos que perpassam o planejamento, a aplicação e a avaliação do Celpe-Bras. Essa empresa é denominada "empresa aplicadora". Após a aplicação da prova, cada Posto Aplicador envia as produções dos examinandos para a empresa aplicadora, para que esta dê sequência à avaliação dos textos da Parte Escrita e à reavaliação das interações dos examinandos que tiveram notas discrepantes na Parte Oral.

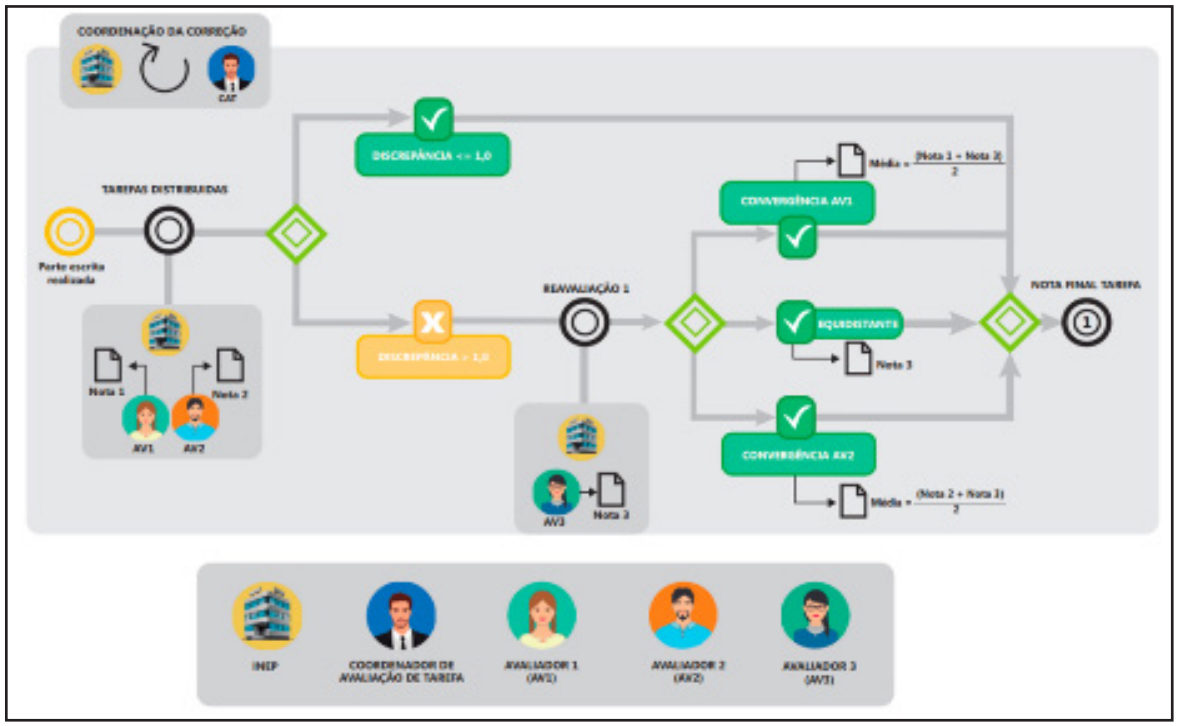
Destacamos que, após a aplicação do exame e envio dos textos à empresa responsável, e antes de ter início a avaliação propriamente dita, acontece a formação dos avaliadores da Parte Escrita. Conforme o Documento Base,

> a avaliação é realizada por professores e pesquisadores da área de PLE, selecionados pela empresa contratada para operacionalizar o Exame. Esses professores devem ter formação em Letras, em nível de graduação e de pós graduação, experiência no ensino de português para estrangeiros e experiência em pesquisa na área de PLE e/ou de avaliação. (Brasil, 2020, p. 71).

Após selecionados, os avaliadores passam por um curso autoformativo individual online e por uma formação síncrona, na qual se reúnem os coordenadores das tarefas da edição do exame com os examinadores designados para cada grupo, geralmente compostos por 4 a 8 pessoas. O último passo antes do início efetivo da avaliação das produções escritas é a avaliação de uma amostra composta por 10 textos, que funciona como um teste da correção a fim de verificar se o avaliador está seguindo os critérios estipulados e se é capaz de operacionalizá-los em conformidade com a formação recebida. Finalizadas todas as etapas prévias, inicia-se a avaliação dos textos dos

examinandos. Cada texto produzido é avaliado, de forma independente, por dois avaliadores, que atribuem ao texto uma nota de 0 a 5. Essa nota é proveniente do cotejamento entre o texto avaliado e os parâmetros de avaliação holísticos (Brasil, 2020, p. 39) que serão utilizados igualmente nas quatro tarefas, considerando as especificidades de cada uma, abordadas durante a formação. Além disso, são tomadas como balizadoras as orientações apresentadas no documento de Resposta Esperada. O processo de avaliação está demonstrado na Figura 3.

Figura 3 – Fluxograma da Parte Escrita do Exame Celpe-Bras



Fonte: Brasil (2020, p. 73)

Destacamos que os avaliadores não sabem quem será o outro avaliador daquele mesmo texto, nem conhecem a nota atribuída, tampouco se o texto recebeu anteriormente notas discrepantes e se foi disponibilizado para uma 3ª avaliação. A nota final da tarefa resulta da média aritmética das notas atribuídas pelos dois avaliadores, desde que não haja diferença maior que 1 ponto entre elas. Se houver diferença de dois pontos ou mais entre as notas, o texto é avaliado por um terceiro examinador. A nota final da Parte Escrita é calculada a partir da média aritmética entre as notas finais das quatro tarefas.

Consenso e discrepância

Os resultados referentes à análise de consenso e discrepância e, por conseguinte, à quantidade de textos e examinandos reavaliados, apresentados neste artigo, são provenientes de dados gerados

no processo de avaliação da Parte Escrita do Celpe-Bras. Esses dados são compilados a cada edição pela empresa aplicadora e compõem o banco de dados do exame Celpe-Bras mantido pelo Inep. Para verificar o consenso entre os avaliadores que atribuem notas na Parte Escrita do Celpe-Bras, recorremos ao teste de consenso entre as notas atribuídas, calculado a partir da porcentagem de concordância. Utilizamos, para tanto, os dados das edições de 2016.1, 2016.2 e 2017.1. Ainda que seja possível inferir uma porcentagem de consenso geral, ressaltamos que os avaliadores, no exame Celpe-Bras, são designados para avaliar uma dentre as quatro tarefas apresentadas e que as notas atribuídas por estes são correspondentes àquela tarefa específica. Isso exige que as análises sejam conduzidas por tarefa e não pela totalidade da edição.

O consenso se refere ao nível de concordância entre as notas atribuídas pelos avaliadores. No Celpe-Bras, há consenso entre os avaliadores não apenas quando estes atribuem exatamente a mesma nota, mas também quando a nota atribuída é adjacente, isto é, varia em apenas um ponto (o Avaliador 1 dá uma nota 3 e o Avaliador 2 dá uma nota 4, por exemplo). Podemos pensar, portanto, que o oposto imediato do consenso será a discrepância. Para o cálculo do consenso, foram consideradas apenas as notas atribuídas pelos Avaliadores 1 e 2 – na falta de consenso entre eles, constata-se a discrepância, que encaminhará o texto à terceira avaliação. De acordo com Stemler (2004), o valor de 75% é considerado o mínimo de consenso aceitável, enquanto valores a partir de 90% são considerados altos em avaliações de larga escala.

Cabe lembrar que, para o cálculo do consenso, consideramos os pares de avaliadores e como estes atribuem notas a um mesmo texto. Para isso, realizamos o cálculo de porcentagem de casos que respondiam ao consenso e à discrepância. Essa análise foi feita para todos os examinandos, nas quatro tarefas apresentadas na Parte Escrita. A partir do consenso entre examinadores para um mesmo texto, é possível calcular o consenso obtido por tarefa e, também, na totalidade da Parte Escrita. Na sequência, apresentaremos os dados referentes a cada uma das três edições.

Edição 2016.1

Nesta edição, 4.603 examinandos realizaram a Parte Escrita, o que caracterizou a avaliação de 18.412 textos considerando-se as quatro tarefas propostas. Chamamos a atenção para o fato de a quantidade de examinandos e de textos avaliados serem diferentes, pois um mesmo examinando produz quatro textos na Parte Escrita. O índice geral de consenso entre as notas atribuídas na PE, na edição de 2016.1, foi de 84,36%; de maneira complementar, a discrepância foi de 15,64%. Apresentamos, na sequência, os dados de consenso e discrepância para cada uma das tarefas da edição:

Tabela 1 – Consenso e discrepância por tarefas da Parte Escrita do Exame Celpe-Bras (2016.1)

Tarefa	Consenso entre avaliadores	Discrepância
Tarefa 1	87,14%	12,86%
Tarefa 2	84,42%	15,58%
Tarefa 3	82,82%	17,18%
Tarefa 4	83,05%	16,95%
Total da PE	84,36%	15,64%

Fonte: elaborada pelas autoras

Como os dados demonstram, a tarefa que apresentou maior índice de consenso foi a Tarefa 1, com 87,14% de concordância entre as notas atribuídas. Por outro lado, a que teve menor consenso foi a Tarefa 3, com 82,82%. A partir dos resultados de cada tarefa e das notas atribuídas por todos os avaliadores desta edição, é possível constatar que o consenso entre avaliadores da edição ficou em 84,36%, porcentagem bem superior à aceitável referenciada pela literatura da área (Stemler, 2004).

Ao observar essa discrepância a partir dos examinandos reavaliados, constatamos que 2.267 sujeitos, ou seja, 49,25% dos examinandos da edição, tiveram pelo menos um de seus textos reavaliados. Nesse universo de examinandos reavaliados, contudo, é possível que um mesmo examinando tenha tido um, dois, três ou até mesmo os quatro textos reavaliados - a maioria dos examinandos que passaram por reavaliação tiveram apenas 1 texto reavaliado (76,13%).

Para além dos dados já apresentados, uma de nossas hipóteses iniciais era que a reavaliação de textos poderia alterar a nota final dos examinandos. Para verificar, então, se a reavaliação dos textos que receberam notas discrepantes na Parte Escrita do Celpe-Bras afeta a nota final da Parte Escrita, a nota final do exame e o nível de certificação atribuído ao examinando, comparamos o valor original da nota final da Parte Escrita com o cálculo de como seria a nota final da Parte Escrita sem a reavaliação dos textos que tiveram notas discrepantes (fazendo-se a média entre as notas do Avaliador 1 e Avaliador 2). Para calcular a porcentagem de examinandos que tiveram alteração nas notas ou no nível de certificação, realizamos a comparação entre as notas e o nível de proficiência originais e as notas e o nível de proficiência resultantes do cálculo descrito anteriormente.

Como recorte para essa análise, utilizamos apenas os examinandos que tiveram textos reavaliados e que, ao mesmo tempo, tiveram sua nota de proficiência proveniente da nota final da Parte Escrita. Relembramos que, no Celpe-Bras, o certificado é atribuído com base na nota mais baixa recebida pelo participante entre as duas partes do exame (Brasil, 2020, p. 79). Nessa edição, 3.886 examinandos (84,42%) tiveram sua nota final de proficiência proveniente da nota final da Parte Escrita do exame. Destacamos, ainda, que desses examinandos que tiveram a nota final de proficiência proveniente da nota final da PE, 41,17% (1.895) tiveram algum texto reavaliado,

o que confirma o potencial de impacto que a parte escrita e a reavaliação de textos com notas discrepantes podem ter na certificação dos examinandos.

Tabela 2 – Alteração de nota final da Parte Escrita, da nota final de proficiência e do nível de certificação (2016.1)

	Examinandos reavaliados	**Alteração de nota final da PE**	**Alteração de nota final de proficiência**	**Alteração de nível de certificação**	**Porcentagem de examinandos com alteração no nível de certificação**
Total	1.184	847	844	297	25,08%

Fonte: elaborada pelas autoras

Em resumo, dos 1.895 examinandos que tiveram textos reavaliados, 1.397 tiveram alteração na nota final da Parte Escrita. Destes, 1.391 tiveram alteração na nota final de proficiência e, destes, 504 tiveram alteração no nível de certificação. Destacamos que não há total correspondência entre a quantidade de examinandos que tiveram alteração na nota final da Parte Escrita e na nota final de proficiência. Isso ocorre porque, no recorte analisado, havia alguns poucos casos em que as notas finais da PE e da PO eram originalmente iguais, o que nos levou a incluir esses examinandos como participantes que tiveram a nota final proveniente da PE. A partir do cálculo de como seria a nota final da Parte Escrita sem a reavaliação dos textos que tiveram notas discrepantes (calculada a partir da média entre as notas do Avaliador 1 e Avaliador 2), obtivemos uma nova nota final para a PE, em todos esses casos, superior à original. Como não realizamos modificações na nota final da PO, esta seguiu igual e menor que a nova nota calculada para a PE, indicando que, embora a reavaliação indicasse mudança na nota final da PE, ela não teria impactos na nota final de certificação, que passaria a ser definida pela nota final da PO. Outro ponto que merece destaque é a porcentagem de alteração no nível de proficiência. Ainda que represente uma alteração menor se comparada às alterações na nota final da Parte Escrita (73,72%) e na nota final de proficiência (73,40%), consideramos 504 (26,60%) um número bem expressivo de examinandos que tiveram a mudança de nível ocasionada pela reavaliação de textos na Parte Escrita.

Ao identificar a mudança de nível, buscamos mapear se essa mudança aumentou ou diminuiu o nível de certificação dos examinandos. Os dados indicaram que para 336 examinandos, 66,67% dos 504 que tiveram algum tipo de alteração, a reavaliação de texto(s) na Parte Escrita do exame impactou de maneira a aumentar o nível de certificação. Esse aumento ocorre, na maior parte dos casos, para o nível imediatamente superior (Intermediário para Intermediário Superior, por exemplo). Para os demais 168 examinandos, 33,33% da amostra, houve diminuição do nível de certificação, igualmente para o nível imediatamente inferior (do Intermediário Superior para

o Intermediário, por exemplo). Percebemos então que, na maior parte dos casos de reavaliação ocorridos nesta edição, a reavaliação foi benéfica ao examinando.

Edição 2016.2

Na segunda edição de 2016, 4.729 examinandos realizaram a Parte Escrita, o que caracterizou a avaliação de 18.916 textos dentre as quatro tarefas propostas. Identificamos que o índice geral de consenso entre as notas atribuídas na PE, na edição de 2016.2, foi de 85,33% e a discrepância foi de 14,67%. A tabela a seguir esquematiza os dados de consenso e discrepância para cada uma das tarefas da edição:

Tabela 3 – Consenso e discrepância por tarefas da Parte Escrita do Exame Celpe-Bras (2016.2)

Tarefa	Consenso entre avaliadores	Discrepância
Tarefa 1	83,97%	16,03%
Tarefa 2	87,23%	12,77%
Tarefa 3	86,7%	13,3%
Tarefa 4	84,97%	15,03%
Total da PE	85,33%	14,67%

Fonte: elaborada pelas autoras

Como os dados demonstram, a tarefa que apresentou maior índice de consenso foi a Tarefa 2, com 87,23% de concordância entre as notas atribuídas. Por outro lado, a que teve menor consenso foi a Tarefa 1, com 83,97%. A partir dos resultados de cada tarefa e das notas atribuídas por todos os avaliadores desta edição, é possível projetar que o consenso total entre avaliadores da edição ficou em 85,33%, porcentagem aceitável referenciada pela literatura da área.

Ressaltamos que, ao observar essa discrepância a partir dos examinandos reavaliados, constatamos que 2.163 sujeitos, ou seja, 45,74% dos candidatos da edição, tiveram pelo menos um de seus textos reavaliados. Como na edição anterior, a maioria desses examinandos, 1.678 (77,58%), teve apenas 1 texto reavaliado.

Na edição 2016.2, 79,06% dos examinandos (3.739) tiveram sua nota final de proficiência proveniente da nota final da Parte Escrita do exame. Desses, 45,76% (1.711) tiveram algum texto reavaliado. Quanto ao impacto da reavaliação de textos na definição da nota final da parte escrita, na nota final do exame e no nível de certificação, apresentamos os dados na tabela:

Tabela 4 – Alteração de nota final da Parte Escrita, da nota final de proficiência e do nível de certificação (2016.2)

	Examinandos reavaliados	**Alteração de nota final da PE**	**Alteração de nota final de proficiência**	**Alteração de nível de certificação**	**Porcentagem de examinandos com alteração no nível de certificação**
Total	1.711	1.275	1.270	447	26,12%

Fonte: elaborada pelas autoras

Em resumo, dos 1.711 examinandos resultantes do recorte escolhido para essa análise (que considera apenas o número de examinandos que tiveram textos reavaliados e que tiveram a sua nota final da Parte Escrita dando origem à nota de proficiência e ao nível de certificação), 1.275 tiveram alteração na nota final da Parte Escrita. Destes, 1.270 tiveram alteração na nota final de proficiência e, destes, 447 tiveram alteração no nível de certificação. Ainda que represente uma alteração menor se comparada às alterações na nota final da Parte Escrita (74,52%) e na nota final de proficiência (74,22%), consideramos 447 (26,12%) um número bem expressivo de examinandos que tiveram mudança de nível ocasionada pela reavaliação de textos na Parte Escrita.

Constatada a mudança de nível para 447 examinandos, identificamos que para 197 destes examinandos (44,07%), a reavaliação de texto(s) na Parte Escrita impactou de maneira a aumentar o nível de certificação - novamente, esse aumento é para o nível imediatamente superior (Intermediário para Intermediário Superior, por exemplo). Os outros 250 examinandos (55,93%) tiveram diminuição do nível, também para o imediatamente inferior. Cabe destacar que, na edição de 2016.2, a reavaliação de textos na Parte Escrita do exame ocasionou mais a diminuição do que o aumento no nível de certificação.

Edição 2017.1

Nesta edição, 3.968 examinandos realizaram a Parte Escrita, o que caracterizou a avaliação de 15.872 textos dentre as quatro tarefas propostas. O índice geral de consenso entre as notas atribuídas na Parte Escrita foi de 86,81%; já a discrepância foi de 13,19%. A tabela apresentada na sequência apresenta os dados relativos ao consenso e à discrepância por tarefas:

Tabela 5 – Consenso e discrepância por tarefas da Parte Escrita do Exame Celpe-Bras (2017.1)

Tarefa	Consenso entre avaliadores	Discrepância
Tarefa 1	85,36%	14,64%
Tarefa 2	88,89%	11,11%
Tarefa 3	87,88%	12,12%
Tarefa 4	85,13%	14,87%
Total da PE	86,81%	13,19%

Fonte: elaborada pelas autoras

Bem como na edição anterior, a tarefa que apresentou maior índice de consenso foi a Tarefa 2, com 88,89% de concordância entre as notas atribuídas. A que teve menor consenso foi a Tarefa 4, com 85,13%. É possível projetar, a partir dos dados das demais tarefas, que o consenso entre avaliadores da edição é de 86,81%, porcentagem aceitável referenciada pela literatura da área.

Considerando os examinandos reavaliados, verificamos que 1.710 sujeitos (43,09%) tiveram pelo menos um de seus textos submetidos a uma terceira avaliação, sendo que a maioria, 1.366 (79,9%), teve apenas 1 texto reavaliado.

Na edição de 2017.1, 2.838 examinandos (71,52%) tiveram sua nota final de proficiência proveniente da nota final da Parte Escrita do exame. Desses examinandos, 1.184 (41,72%) tiveram algum texto reavaliado. Partimos do grupo de examinandos reavaliados em alguma das quatro tarefas do exame para verificar o possível impacto da reavaliação dos textos na nota final da Parte Escrita e do exame e do nível de certificação, conforme dados esquematizados na sequência:

Tabela 6 – Alteração de nota final da Parte Escrita, da nota final de proficiência e do nível de certificação (2017.1)

	Examinandos reavaliados	Alteração de nota final da PE	Alteração de nota final de proficiência	Alteração de nível de certificação	Porcentagem de examinandos com alteração no nível de certificação
Total	1.184	847	844	297	25,08%

Fonte: elaborada pelas autoras

Em resumo, dos 1184 examinandos resultantes do recorte escolhido para essa análise, 847 tiveram alteração na nota final da Parte Escrita, destes, 844 tiveram alteração na nota final de proficiência e, destes, 297 tiveram alteração no nível de certificação. Ainda que represente uma alteração menor se comparada às alterações na nota final da Parte Escrita (71,54%) e na nota final

de proficiência (71,28%), consideramos 297 (25,08%) um número bem expressivo de examinandos que tiveram a mudança de nível ocasionada pela reavaliação de textos na Parte Escrita. Quanto à mudança de nível, para 189 examinandos, 63,64% dos 297 que tiveram algum tipo de alteração, a reavaliação de texto(s) na Parte Escrita do exame impactou de maneira a aumentar o nível de certificação. Para os demais 108 examinandos (36,36%), houve diminuição do nível. Em ambos os casos, a mudança é para um nível imediatamente superior ou inferior.

Cruzamento dos dados

Buscando compreender o que o processo de atribuição de notas na Parte Escrita e a reavaliação de produções com notas discrepantes poderiam informar sobre a confiabilidade da avaliação da Parte Escrita do Celpe-Bras, discorremos sobre os resultados do cálculo de consenso/discrepância, sobre a quantidade de textos reavaliados e sobre o impacto da reavaliação na nota final da PE e na nota e no nível de certificação dos examinandos. Wheelan (2016) defende que uma das funções principais da estatística é utilizar os dados que temos disponíveis para fazer "conjecturas informadas sobre perguntas mais amplas" (Wheelan, 2016, p. 21). Nesse sentido, realizaremos, nesta seção, o cruzamento dos dados estatísticos anteriormente apresentados para fazer conjecturas informadas sobre a confiabilidade entre avaliadores da Parte Escrita do Exame Celpe-Bras.

Ressaltamos, inicialmente, a relevância da nota final da Parte Escrita para a definição do nível de certificação dos examinandos. Reforçamos que, conforme Segat (2023), nas edições de 2016.1, 2016.2 e 2017.1, respectivamente 84,4%, 79,1% e 71,5% dos examinandos tiveram suas notas finais e o nível de certificação determinado pela nota obtida na parte escrita. A partir desse dado, referenciamos o artigo de Schlatter *et al.* (2021) que analisa estatisticamente as notas de quatro edições do exame. As autoras apontam que, na edição de 2017.1, a Tarefa 1 teve notas médias maiores do que as outras tarefas da mesma edição e de edições anteriores, o que acabou aumentando, também, a nota final da Parte Escrita. Essa é, provavelmente, a justificativa para que um número maior de examinandos tenha a nota de proficiência proveniente da Parte Oral, baixando, por conseguinte, a porcentagem de examinandos que tiveram a nota de proficiência originada pela nota da PE, conforme referido anteriormente (71,5%).

Em relação ao consenso, não é possível identificar nenhum padrão de diferença entre as tarefas e as edições, isto é, não é possível afirmar que alguma das edições ou das tarefas tenha recorrentemente menor ou maior consenso (e por conseguinte, maior ou menor discrepância). De acordo com Stemler (2004), 75% é considerado o valor mínimo para se afirmar que o consenso entre os avaliadores é aceitável, enquanto valores a partir de 90% são considerados altos. As porcentagens que indicam o consenso entre os avaliadores nas tarefas e nas edições de 2016.1, 2016.2 e 2017.1 ficam entre 83% e 88%, valores considerados aceitáveis e, inclusive, mais próximos do que seria classificado como um consenso alto entre os avaliadores. Nesse sentido, é possível

afirmar que o nível de consenso é aceitável entre avaliadores de todas as tarefas nas edições de 2016.1, 2016.2 e 2017.1 do exame Celpe-Bras. Entendemos que esse dado também ajuda a demonstrar que a reavaliação é um recurso muito importante para verificação da nota atribuída, no sentido de garantir um novo olhar que garanta que o examinando receba uma nota adequada e que o resultado seja, portanto, mais confiável.

Outro aspecto que consideramos muito relevante nas análises realizadas é o fato de confirmarmos que a reavaliação afeta o nível de certificação de boa parte dos examinandos. Se a própria reavaliação é um mecanismo que ajuda a garantir a confiabilidade entre avaliadores, pareceu-nos fundamental explorar como as notas atribuídas por essa reavaliação poderiam impactar o nível de certificação dos examinandos. Como os resultados indicaram, é possível inferir que a reavaliação afeta o nível de certificação de pelo menos 25% dos examinandos reavaliados, o que consideramos um valor muito expressivo. Ao pensarmos nesses valores em comparação ao número total de examinandos de cada edição, é evidente que a porcentagem diminui (e muito). Nesse cenário, poderíamos dizer que, para a edição de 2016.1, do total de 4.603 examinandos, 504 (10,95%) obtiveram mudança no nível de certificação. Na edição de 2016.2, do total de 4.729 examinandos, 447 (9,45%) obtiveram mudança no nível de certificação e na edição de 2017.1, do total de 3.968 examinandos, 297 (7,48%) obtiveram mudança após a reavaliação. De qualquer maneira, essa análise é importante pois, para todos os fins práticos, o nível de certificação será a informação que importa para o examinando, capaz de incluí-lo ou excluí-lo de determinadas práticas sociais.

Quanto a mudança no nível de certificação, isto é, se houve aumento ou diminuição após a reavaliação, nas edições de 2016.1 (66,67%) e 2017.1 (63,64%) houve tendência de aumento, enquanto na edição de 2016.2 predominou a diminuição (55,93%). Destacamos que os níveis de certificação que concentraram as maiores alterações foram o Intermediário e o Intermediário Superior, seja como resultado de aumento ou de diminuição de nível. Esse resultado é bastante coerente com as notas médias e as notas modas observadas e descritas no estudo de Schlatter et al. (2021), que discorre sobre as notas das tarefas da Parte Escrita, sublinhando a concentração de examinandos nessas duas faixas de notas e certificação.

Considerações finais

Este artigo sistematizou a discussão sobre confiabilidade entre avaliadores na Parte Escrita do Exame Celpe-Bras, considerando os processos de avaliação e de reavaliação de textos com notas discrepantes e seus impactos na definição da nota e nível de certificação dos examinandos. Dentro do escopo da Linguística Aplicada, descrevemos o processo de avaliação e reavaliação da Parte Escrita do Exame Celpe-Bras, discutimos a noção de confiabilidade no campo de estudos da avaliação e, mais especificamente, descrevemos o consenso e a discrepância – e, portanto, a

reavaliação – entre avaliadores. Além disso, analisamos o impacto da reavaliação na definição da nota final da Parte Escrita, na definição da nota de certificação e do nível de certificação, partindo de dados provenientes das tarefas da Parte Escrita das duas edições do exame aplicadas em 2016 e da primeira edição de 2017.

Entendemos que os resultados apresentados fornecem evidências para atestar a confiabilidade entre avaliadores no processo de avaliação da Parte Escrita, a partir do resultado do cálculo de consenso e da constatação do impacto da reavaliação na definição da nota final da parte escrita e da nota final do exame e, por conseguinte, do nível de certificação. Atribuímos esse resultado à formação dos avaliadores e propomos que apostar nesta formação e no acompanhamento deles é a solução para diminuir ainda mais a discrepância (aumentando o consenso) e garantir uma avaliação ainda mais confiável e válida. Além disso, entendemos que essa formação proporciona reflexões no campo do letramento em avaliação que podem ocasionar impactos positivos não só nas práticas avaliativas do Celpe-Bras como também em outras que têm a participação desses sujeitos. Dito isso, indicamos que novas pesquisas como esta sejam feitas a partir dos resultados de outras edições do exame, junto a análises qualitativas do processo de formação e atribuição de notas, tanto na Parte Escrita quanto na Parte Oral.

Ressaltamos que as evidências encontradas por este estudo possibilitam fazer inferências somente sobre a confiabilidade entre avaliadores e não sobre os demais tipos de confiabilidade, visto que esta é uma característica que é influenciada por muitos fatores diferentes (perfil do avaliador, perfil do examinando, dificuldade da tarefa, por exemplo). Nesse sentido, entendemos que há alguns aspectos que merecem estudos mais detalhados e desde outras perspectivas metodológicas. Defendemos que é fundamental, se há o intuito de ter uma prática avaliativa acompanhada por uma reflexão crítica e ética, que os responsáveis pelo exame se empenhem para que pesquisas como esta possam ser feitas a cada edição de aplicação das provas, a fim de que haja um acompanhamento longitudinal da confiabilidade entre avaliadores – e dos demais tipos de confiabilidade. O fato de termos acesso apenas a essas edições do exame limita as discussões feitas a um período específico, ainda que as análises nos permitam projetar os mesmos resultados para outras edições. Além disso, outros dados, que sejam úteis para discutir e verificar se o argumento de validade e de confiabilidade do exame estão sendo construídos para os usos a que se destinam, devem ser reportados para a comunidade interessada.

Referências

BAKHTIN, Mikhail Mikhailovich. **Estética da criação verbal**. São Paulo: Martins Fontes, 2003.

BACHMAN, Lyle F.; PALMER, Adrian S. **Language testing in practice: Designing and developing useful language tests**. Oxford: Oxford University Press, 1996.

BRASIL. **Guia do Participante: tarefas comentadas que compõem a edição de abril de 2013 do exame**. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Ministério da Educação, 2013.

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

CHAPELLE, Carol A. **Reliability in language assessment**. Iowa State University, 2013.

CLARK, Herbert. **Using Language**. Cambridge: Cambridge University Press, 1996.

DAVIS, Larry. **The influence of training and experience on rater performance in scoring spoken language**. Language Testing, v. 33, n. 1, p. 117–135, 2016.

KNOCH, Ute; SITAJALABHORN, Woranon. **A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes**. Assessing Writing, v. 18, n. 4, p. 300-308, 2013.

MCNAMARA, Timothy Francis. **Measuring Second Language Performance**. Londres: Longman, 1996.

MCNAMARA, Timothy Francis. Language Testing. In: DAVIES, Alan; ELDER, Catherine. **The Handbook of Applied Linguistics**. Londres: Blackwell Publishing, 2004.

MCKAY, Todd; PLONSKY, Luke. Reliability Analyses: Estimating Error. In: WINKE, Paula; BRUNFAUT, Tineke. **The Routledge Handbook of Second Language Acquisition and Language Testing**. London: Routledge, 2021.

MENDEL, Kaiane. **Proficiência e autoria na avaliação integrada de leitura e escrita do exame Celpe-Bras**. Dissertação (Mestrado em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

NEVES, Liliane de Oliveira. **Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras: um estudo longitudinal**. 2018. Tese (Doutorado em Estudos da Linguagem) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2018.

SCHLATTER, Margarete; SCARAMUCCI, Matilde; PRATI, Silvia; ACUÑA, Leonor. Celpe-Bras e Celu: impactos da construção de parâmetros comuns de avaliação de proficiência em portugês e em espanhol. In: FONTANA, Mónica Zoppi (Org.) **O português do Brasil como língua transnacional**. Campinas: RG Editora, 2009.

SCHLATTER, Margarete; NUNES, Luciana Neves; KUNRATH, Simone Paula. **Análise descritiva da parte escrita do exame CELPE-BRAS**. Brasília, 2021.

SCHOFFEN, Juliana Roquele. **Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras**. Tese (Doutorado em Linguística Aplicada) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

SEGAT, Giovana Lazzaretti. **Estudos sobre confiabilidade em exames de proficiência: o processo de atribuição de notas e a reavaliação na parte escrita do Celpe-Bras**. Dissertação (Mestrado em Letras) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

STEMLER, Steven. **A comparison of consensus consistency and measurement approaches to estimating interrater reliability**. Practical Assessment Research & Evaluation, v.9, n. 4, 2004.

WANG, Jue; ENGELHARD, George; RACZYNSKI, Kevin; SONG, Tian; WOLFE; Edward. **Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach**. Assessing Writing, v. 33, 2017.

WHEELAN, Charles. **Estatística: o que é, para que serve, como funciona**. Rio de Janeiro: Zahar, 2016.

YAN, Xun; FAN, Jason. Reliability and dependability. In: FULCHER, Glenn; HARDING, Luke. **The Routledge Handbook of Language Testing**. 2. ed. London: Routledge, 2022. pp. 477-494.