CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS DE LINGUAGENS

Liliane de Oliveira Neves

Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras: um estudo longitudinal

Tese apresentada ao Programa de Pós-Graduação em Estudos de Linguagens do CEFET-MG, como requisito parcial para obtenção do título de doutora.

Área de concentração: Tecnologia e processos discursivos.

Linha de Pesquisa III: Linguagem, ensino, aprendizagem e tecnologia.

Orientador: Prof. Dr. Jerônimo Coura-Sobrinho (CEFET-MG).
Coorientador: Prof. Dr. Felipe Dias Paiva (CEFET-MG).

Belo Horizonte, agosto de 2018.

Neves, Liliane de Oliveira.
N518c Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras : um estudo longitudinal / Liliane de Oliveira Neves. - 2018.
240 f. : il., grafs., tabs.
Orientador: Jerônimo Coura Sobrinho
Coorientador: Felipe Dias Paiva

Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação em Estudos de Linguagens, Belo Horizonte, 2018.
Bibliografia.

1. Certificado de Proficiência em Língua Portuguesa para Estrangeiros. 2. Confiabilidade. 3. Avaliação educacional – Metodologia - Análise. I. Coura Sobrinho, Jerônimo. II. Paiva, Felipe Dias. III. Título.

CDD: 469.824

Ficha elaborada pela Biblioteca - Campus I – CEFET-MG
Bibliotecário: Wagner Oliveira Braga CRB6 - 3261

Liliane de Oliveira Neves

Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras: um estudo longitudinal

Tese apresentada ao Programa de Pós-Graduação em Estudos de Linguagens do CEFET-MG, como requisito parcial para obtenção do título de doutora.
Área de concentração: Tecnologia e processos discursivos.
Linha de Pesquisa III: Linguagem, ensino, aprendizagem e tecnologia.

Aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Jerônimo Coura-Sobrinho (orientador)
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Prof. Dr. Felipe Dias Paiva (coorientador)
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Prof. Dr. Rui Brites Correia da Silva
Instituto Superior de Economia e Gestão da Universidade de Lisboa (ISEG)

Prof. Dr. Ronaldo Amorim Ozório da Matta Lima
Universidade Federal Fluminense (UFF)

Prof[a]. Dr[a]. Elizabeth do Nascimento
Universidade Federal de Minas Gerais (UFMG)

Prof. Dr. Renato Caixeta da Silva
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Belo Horizonte, 10 de agosto de 2018.

Agradecimentos

A Deus, por iluminar sempre os meus caminhos.

A minha família, por sempre me apoiar em minhas decisões, inclusive as acadêmicas.

Ao CEFET-MG, pela minha liberação para a realização deste estudo.

À CAPES, pela concessão da bolsa de estudos de doutorado sanduíche, em Portugal.

Ao INEP, pela presteza e liberação dos dados desta pesquisa.

Ao prof. Jerônimo Coura-Sobrinho, por aceitar, mais uma vez, ser meu orientador e pelas ricas contribuições em meu trabalho. A você, serei eternamente grata!

Ao prof. Felipe Dias Paiva, por ter aceito o desafio de me coorientar e por viabilizar minha inserção na área dos estudos quantitativos.

Ao prof. Rui Brites, por ter aceito o convite de ser meu coorientador no estágio de doutorado sanduíche, no ISEG / Universidade de Lisboa – Portugal, e pelos diálogos produtivos.

Aos professores da minha banca de qualificação e defesa, pelas ricas contribuições.

Ao prof. João Marôco (ISPA), pelo auxílio teórico-metodológico.

À profa. Ana Nápoles, pela força e incentivo de sempre.

À Mariana Valentin, ex-orientanda de PIBIC, pelo auxílio na fase de levantamento de dados.

Aos servidores e estagiários da Secretaria do Posling, em especial à Sandra, pela gentileza e atenção.

Ao prof. Renato Caixeta, pelo apoio para a concretização de meu estágio de doutorado sanduíche.

Aos meus colegas de trabalho: prof. Márcio, prof. Irlen, prof[a]. Heloísa, Fátima, Luciana, Kenny, Imaculada, Flávia, Wesley e Marquinhos, por terem me incentivado a seguir nessa caminhada acadêmica.

A minha amiga Maria Luiza, pelo carinho e atenção de sempre.

À Natália Tosatti, pela amizade, torcidas positivas e contribuições ao longo desta caminhada.

À Laura Ferreira, colega de curso e orientação, pelos diálogos metodológicos.

Aos amigos que fiz no ISEG, em especial: Filomena Ferreira, Gicele Martins, Cláudia Melatti, Jorge Lima e Edilson Araújo.

À Carla Mirelle, pela partilha de tempo, casa e experiências em Lisboa.

Aos amigos Arcade, Brice, Gérard, Isabela, Júnia, Naveen, Rafaela e Rose, pelas motivações ao longo do processo.

E até ao Hachiko, pelas distrações.

RESUMO

As avaliações em larga escala desempenham papel importante na sociedade, pois servem para identificação de saberes de determinados grupos, (re)direcionamento de políticas públicas e tomada de decisões. Devido a isso, é necessário que apresentem resultados consistentes e que reflitam o construto que objetivam avaliar. Nesse cenário de avaliação, esta tese trata do exame que confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras), que é composto por duas provas, uma escrita e outra oral. A prova oral, foco deste trabalho, é uma interação face a face em que participam o examinando e dois avaliadores: o avaliador-interlocutor (AI) e o avaliador-observador (AO), sendo que ambos fazem a avaliação do desempenho oral do examinando, a partir de descritores constantes de duas grades distintas. A avaliação é feita em primeira instância (imediatamente após a aplicação da prova) e, havendo discrepância significativa entre as notas atribuídas pelos dois avaliadores, a interação é reavaliada em segunda e/ou terceira instância. O objetivo geral desta tese é analisar de que maneira a confiabilidade dos resultados do exame tem relação com o comportamento avaliativo de AI e AO. A confiabilidade é uma das qualidades desejáveis de todo teste e diz respeito à consistência da avaliação, ou seja, quanto mais os resultados forem livres de erro, mais confiáveis eles serão. Já o comportamento avaliativo é entendido na pesquisa como a maneira como os avaliadores atribuem notas ao desempenho oral dos examinandos, nas diferentes instâncias. Foi empregada uma metodologia quantitativa de análise de dados, que levou em conta dados de sete edições consecutivas do exame Celpe-Bras, envolvendo notas de 29.831 examinandos, sendo que o marco teórico considerou estudos da Psicometria (como Murphy e Davidshofer, 2005; Urbina, 2007), da Estatística (como Marôco e Garcia Marques, 2006; Marôco, 2014) e da Linguística Aplicada (como Bachman, 1990; 2004). Descrições e análise dos níveis de proficiência atribuídos aos examinandos e de informações estatísticas das notas, como medidas de tendência central e de dispersão, serviram de base para constatar a existência de variabilidade de comportamento avaliativo. A pergunta de pesquisa: *o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste?*, foi respondida com base em três técnicas. São elas: (i) uma análise preliminar ao estudo da confiabilidade, via Análise dos Componentes Principais, para verificar a dimensionalidade da escala de avaliação; (ii) cálculo do coeficiente *Alfa de Cronbach*, para verificar a consistência interna dos itens da escala e (iii) cálculo do coeficiente *Kappa*, para identificar o nível de concordância entre os avaliadores. Os resultados permitem responder positivamente à pergunta de pesquisa, na medida em que: (i) a escala de avaliação apresenta-se unidimensional, ou seja, avalia um único construto, na avaliação realizada em primeira instância; na segunda instância, ela é bidimensional; (ii) as sete edições apresentam valores altos do coeficiente de confiabilidade na avaliação feita em primeira instância, o que significa que os itens da escala possuem elevada consistência interna; já na avaliação realizada em segunda instância, a confiabilidade revela-se moderada e (iii) as sete edições, na avaliação em primeira instância, apresentam valores *satisfatórios* de concordância entre os avaliadores, ainda que baixos; a avaliação realizada em segunda instância apresenta valor *pobre* de concordância. Isso significa que a segunda instância, que é a responsável por dirimir os problemas avaliativos que surgem na primeira, é marcada por comportamento diferenciado dos sujeitos avaliadores, diminuindo, portanto, a confiabilidade dos resultados. Os resultados desta tese sinalizam para a necessidade de algumas ações, das quais destacamos: 1) revisão dos descritores da grade avaliativa, de forma que seja possível diminuir os níveis de subjetividade inerente à própria atividade de avaliar; 2) intensificar as capacitações dos envolvidos no processo avaliativo. Essas ações são necessárias para melhorar o grau de confiabilidade dos resultados do Celpe-Bras.

Palavras-chave: exame Celpe-Bras; confiabilidade; comportamento avaliativo.

ABSTRACT

Large-scale assessments play an important role in society, since they aid the identification of the knowledges of particular groups, the (re)directing of public policy and the process of decision-making. Therefore, they must present consistent results that reflect the construct to be evaluated. In this scenario, this thesis focuses on the test to Certificate of Proficiency in Portuguese for Foreigners (Celpe-Bras), which is composed of two parts, one written and the other oral. The oral part of the test, focus of this thesis, is a face-to-face interaction between the examinee and two evaluators: the evaluator-interlocutor (AI), who conducts the interaction, and the evaluator-observer (AO), both responsible to rate the oral performance of the examinee, based on descriptors of two distinct grids. The evaluation is done in the first instance (immediately after the test has been applied) and, if there is a significant discrepancy between the scores assigned by the two evaluators, the interaction is re-evaluated in the second and / or third instances. The general objective of this thesis is to analyze how the reliability of the test results is related to the rater's behavior of AI and AO. Reliability is one of the desirable qualities of tests and it is related to the consistency of evaluation, i.e., the more results are error-free, the more reliable they will be. Raters' behavior is considered in this research as the way in which the evaluators attribute grades to the oral performance of the examinees, in different instances. A quantitative methodology was used, which took into account data from seven consecutive editions of the Celpe-Bras exam, involving 29,831 examinees, and the theoretical framework was based on studies of Psychometrics (such as Murphy and Davidshofer, 2005), Statistics (such as Marôco and Garcia Marques, 2006; Marôco, 2014) and Applied Linguistics (such as Bachman, 1990, 2004). Descriptions and analyses of the levels of proficiency attributed to the examinees and statistical information of the grades, such as measures of central tendency and dispersion, served as basis to verify the existence of variability of raters' behavior. The research question: *can evaluative behavior be considered a source of measurement error that interferes with the reliability of the test results?*, was answered based on three techniques. They are: (i) a preliminary analysis to the reliability's study, via Principal Component Analysis, to verify the dimensionality of the evaluation scale; (ii) calculation of Cronbach's alpha coefficient to verify the internal consistency of the scale items and (iii) calculation of the Kappa coefficient to identify the level of agreement among the raters. The results allow us to respond positively to the research question, since: (i) the scale of evaluation is unidimensional, i. e., it evaluates a single construct, in the evaluation performed in the first instance; in the second instance, it is two-dimensional; (ii) the seven editions present high values of reliability coefficient in the first instance of evaluation, which means that the scale items have high internal consistency; in the evaluation carried out in the second instance, the reliability is moderate and (iii) the seven editions, in the first instance of evaluation, present *satisfactory* values of agreement among the evaluators, albeit low; the evaluation carried out in the second instance presents a *poor* value. This means that the second instance, which is responsible for solving the evaluative problems that arise in the first one, is marked by a different behavior of the raters, thus reducing the reliability of the results. The results of this thesis point to the need to take actions, which are worth highlighting: 1) review of the descriptors of the evaluation grid, so that to possibly reduce the levels of subjectivity inherent to the evaluation activity itself; 2) to intensify the training of those involved in the evaluation process. These actions are necessary to improve the reliability of Celpe-Bras's results.

Keywords: Celpe-Bras exam; reliability; raters' behavior.

RÉSUMÉ

Les tests à grande échelle jouent un rôle important dans la société, en ce sens qu'ils servent à mesurer des savoirs de groupes d'individus déterminés, à apporter des changements dans des politiques publiques et à prendre des décisions. Pour cela, il est crucial que ces tests aboutissent à des résultats de qualité, autrement dit, consistants et qui reflètent le construit que l'on se propose de mesurer. C'est dans ce cadre que cette thèse traite du test qui octroie le Certificat d'Aptitude Linguistique en Portugais Langue Etrangère (Celpe-Bras), composé d'une épreuve écrite et une autre orale. L'épreuve orale est une interaction face à face où l'on a un candidat et deux examinateurs qui sont l'examinateur interlocuteur (AI) et l'examinateur observateur (AO). Ces deux examinateurs évaluent la performance orale du candidat sur la base de descripteurs présents dans deux grilles distinctes. Les notes de l'épreuve orale sont attribuées sur place, juste après l'interview, ce que nous appelons « évaluation de première instance ». En cas de discrépance entre les notes attribuées par les examinateurs, la performance du candidat est réévaluée en deuxième instance et, en cas d'une nouvelle discrépance, une autre révision en troisième instance est faite. L'objectif général de cette thèse est d'analyser dans quelle mesure la fiabilité des résultats du test est liée à ce que nous avons appelé le « comportement évaluatif » des examinateurs AI et AO. La fiabilité est une des qualités les plus désirables en matière de tests et elle se rapporte à la consistance des résultats, en d'autres termes, moins il y a d'erreurs dans les résultats, plus ils sont fiables. Le comportement évaluatif est défini ici comme la façon dont les examinateurs attribuent des notes à la performance orale des candidats, au niveau des différentes instances. La méthodologie utilisée est quantitative, qui a pris en compte des données de sept éditions consécutives du test Celpe-Bras, soit un échantillon de 29.831 notes de candidats. En ce qui concerne le cadre théorique, des études des domaines de la psychométrie (à savoir Murphy et Davidshofer, 2005 ; Urbina, 2007), de la statistique (comme Marôco e Garcia Marques, 2006 : Marôco, 2014) et de la linguistique appliquée (comme Bachman, 1990 ; 2004) ont été utilisées. La description et l'analyse des niveaux de compétence attribués aux candidats et des informations statistiques des notes comme des mesures de tendance centrale et de dispersion nous ont servi de base pour constater l'existence de variabilité au niveau du comportement évaluatif. La question de recherche « le comportement évaluatif peut-il être considéré une source d'erreur d'évaluation qui interfère dans la fiabilité des résultats du test ? » a été répondue sur la base de trois techniques, à savoir : (i) une analyse préliminaire à l'étude de la fiabilité, via l'Analyse en Composantes Principales, pour vérifier la dimensionnalité de l'échelle d'évaluation ; (ii) le calcul du coefficient Alpha de Cronbach, pour vérifier la consistance interne des items de l'échelle ; (iii) le calcul du coefficient Kappa, pour vérifier le degré de concordance entre examinateurs. Les résultats permettent de répondre positivement à la question de recherche étant donné que : (i) l'échelle d'évaluation se présente unidimensionnelle, autrement dit, elle n'évalue qu'un construit pendant l'évaluation réalisée en première instance ; en deuxième instance, elle est bidimensionnelle ; (ii) les sept éditions présentent des valeurs hautement fiables quant au coefficient Alfa de Cronbach pendant l'évaluation en première instance, ce qui signifie que les items de chaque échelle présentent une haute consistance interne ; quant aux évaluations réalisées en deuxième instance, cependant, le degré de fiabilité se révèle modéré ; (iii) les sept éditions présentent, en première instance, des valeurs satisfaisantes de concordance entre examinateurs, bien que basses ; les évaluations réalisées en deuxième instance présentent des valeurs faibles de concordance. Cela signifie que la deuxième instance, où les problèmes observés au niveau de l'évaluation de la première instance sont censés être réglés, est marquée par des comportements différents de la part des évaluateurs, réduisant ainsi la fiabilité des résultats. Les résultats obtenus de cette thèse indiquent la nécessité de mener certaines actions. Parmi ces actions, nous en citions: 1) une révision des descripteurs de la grille d'évaluation, dans le sens de diminuer le niveau de subjectivité inhérente à la pratique de l'évaluation ; 2) intensifier la formation des parties prenantes du système d'évaluation. Ces actions sont désirables pour augmenter le degré de fiabilité des résultats du Celpe-Bras.

Mots-clés : Test Celpe-Bras ; Fiabilité ; Comportement évaluatif.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACP	Análise dos Componentes Principais
AI	Avaliador interlocutor
AMPPLIE	Associação Mineira dos Professores de Português como Língua Estrangeira
AO	Avaliador observador
APLE-RJ	Associação dos Professores de Português Língua Estrangeira do Estado do Rio de Janeiro
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CAPLE	Centro de Avaliação de Português Língua Estrangeira
CEFET-MG	Centro Federal de Educação Tecnológica de Minas Gerais
CELPE-BRAS	Certificado de Proficiência em Língua Portuguesa para Estrangeiros
CELU	*Certificado de Español Lengua y Uso*
CPLP	Comunidade dos Países de Língua Portuguesa
D.O.U.	Diário Oficial da União
DAEB	Diretoria da Educação Básica
DELE	*Diploma de Español como Lengua Extranjera*
DELF	*Diplôme d'Études en Langue Française*
EC	Estratégias Comunicativas
EP	Elemento(s) provocador(es) da conversa
EPLIS	Exame de Proficiência em Inglês Aeronáutico do SISCEAB
IELTS	*International English Language Testing System*
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KMO	Kaiser-Meyer-Olkin
LP	Língua Portuguesa
MEC	Ministério da Educação
PE	Prova Escrita / Parte Escrita
PEC-G	Programa de Estudantes-Convênio de Graduação
PEC-PG	Programa de Estudantes-Convênio de Pós-Graduação
PIBIC	Programa Institucional de Iniciação Científica
PLE	Português como Língua Estrangeira
PO	Prova Oral / Parte Oral
QECR	Quadro Europeu Comum de Referência para as Línguas
SDEA	*Santos Dumont English Assessment*
SISCEAB	Sistema de Controle do Espaço Aéreo Brasileiro
SPSS	*Statistical Package for the Social Sciences*
TOEFL	*Test of English as a Foreign Language*
TOEIC	*Test of English for International Communication*
TRI	Teoria de Resposta ao Item

SUMÁRIO

CONSIDERAÇÕES INICIAIS

As avaliações em larga escala podem ser consideradas instrumentos importantes para o (re)direcionamento de políticas públicas, identificação de saberes de determinados grupos ou instituições e tomada de decisões. No cenário da avaliação linguística, a aferição da competência dos sujeitos é feita por meio de testes, estando essa competência atrelada ao construto que subjaz a eles.

Como reflexo de um movimento de internacionalização de instituições de ensino, existem vários testes de proficiência linguística que são exigidos como um dos requisitos para a concretização de mobilidades acadêmicas e contratações profissionais. Dentre eles, há, no Brasil, o exame que confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros, doravante Celpe-Bras, que instaura-se como um instrumento de alta relevância na divulgação da Língua Portuguesa. Diante disso, esta pesquisa insere-se na área de avaliação linguística e objetiva discutir a confiabilidade dos resultados da prova oral do Celpe-Bras, uma característica importante que todo exame dessa natureza deve buscar.

Scaramucci (2009), ao tratar sobre a avaliação da leitura em inglês como língua estrangeira, demonstrou preocupação com a escassez de estudos brasileiros na área da avaliação quando comparados com outras temáticas.

Schoffen (2009), por sua vez, especificamente no que diz respeito ao Celpe-Bras, também trata da necessidade de se intensificar estudos científicos sobre o exame, devido à crescente procura e aos usos que dele são feitos no Brasil e no exterior. Segundo a pesquisadora, são necessárias pesquisas sobre validade e confiabilidade, de forma a contemplar os diversos aspectos envolvidos no sistema de avaliação, como os níveis de proficiência certificados pelo exame, as condições de aplicação, a elaboração das tarefas, o sistema de correção e as grades de avaliação, entre outros, bem como dos efeitos retroativos que o exame vem gerando no ensino (SCHOFFEN, 2009, p. 10).

Coura-Sobrinho (2014), ao tratar dos contextos de aplicação do Celpe-Bras, também reforça o posicionamento de que são necessárias novas investigações no sentido de ampliar discussões acerca dos fatores que podem interferir no desempenho dos examinandos e, consequentemente, na confiabilidade dos resultados do exame.

Diante disso e considerando-se a importância que os processos avaliativos exercem na sociedade, é relevante desenvolver estudos para esse acompanhamento dos instrumentos utilizados na avaliação, com vistas a validá-los, ratificá-los, aprimorá-los e promover

discussões acadêmicas em torno dos temas que deles emergem. Portanto, é isso que se propõe com esta pesquisa, dada a relevância do Celpe-Bras no cenário internacional: tratar da confiabilidade dos seus resultados e do comportamento avaliativo dos profissionais responsáveis pela atribuição de notas aos examinandos.

Alguns pesquisadores já se dedicaram à análise da prova oral do Celpe-Bras e fizeram considerações relevantes sobre confiabilidade, a exemplo de Sakamori (2006), Furtoso (2011a), Bottura (2014), Coura-Sobrinho (2014) e Costa (2015). Em todas as pesquisas, são citados fatores que podem interferir na confiabilidade dos resultados do teste, motivo pelo qual entendemos ser pertinente que haja discussões acerca dessa qualidade.

O interesse da pesquisadora na investigação sobre temas atinentes à avaliação de proficiência linguística surgiu em 2010, desde quando vem se dedicando a estudar o exame Celpe-Bras, além de se envolver nos processos de aplicação e avaliação. Em pesquisa de mestrado defendida em 2012[1], no Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), foi dada ênfase à prova escrita, cuja metodologia foi alicerçada na Análise do Discurso. Dando sequência à pesquisa realizada em 2012, a proposta desta tese é dar foco à prova oral do exame, em especial à confiabilidade dos seus resultados e ao comportamento avaliativo dos atribuidores de nota.

Tratar da confiabilidade, para Bachman (1990), significa responder a seguinte pergunta: "o quanto do desempenho individual em um teste está relacionado ao erro de mensuração ou a outros fatores, além da habilidade linguística que se quer medir?" (BACHMAN, 1990, p. 160). Diante desse questionamento, podemos inferir que, para se garantir a confiabilidade em um instrumento de avaliação, é preciso avaliar quantas e quais variáveis[2] podem exercer influência sobre ele. Dito de outra maneira, é preciso minimizar os

---

[1] DAMAZO, Liliane Oliveira. *A modalização na produção de textos em português como língua estrangeira.* 2012. 220 f. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2012.

[2] A palavra *variáveis* carrega conotações diversas a depender da área do conhecimento e, por este estudo estar circunscrito na área da linguística aplicada e baseado em pressupostos das estatística, vale fazer algumas considerações. No processo de análise estatística, o investigador depara-se com "algo" que precisa medir, controlar ou manipular durante o processo de investigação. Este "algo" designa-se por "variável" (MARÔCO, 2014, p. 7 – grifos do autor). Dito de outra maneira, variável é qualquer característica que varia de um indivíduo para outro. As hipóteses, em geral, contêm uma variável independente (causa) e uma variável dependente (efeito) (LEVIN; FOX; FORDE, 2012, p. 440). Ou seja, do ponto de vista da estatística, *variável* refere-se a algo que varia de um indivíduo para o outro, que faz parte de um banco de dados e que precisa ser analisado no processo de investigação. No caso desta pesquisa, as variáveis que fazem parte do banco de dados são os critérios avaliados para a mensuração do desempenho oral do examinando (compreensão, competência interacional, fluência, adequações lexical e gramatical e pronúncia), além das outras criadas para melhor exploração dos dados, como os níveis de proficiência, por exemplo. Do ponto de vista da linguística aplicada, as

efeitos dessas variáveis no resultado do teste, para que ele reflita diretamente a habilidade linguística que se quer medir e, consequentemente, que seja confiável.

Por tratar-se de um processo subjetivo, a avaliação da proficiência oral precisa ser mensurada com base em descritores explicitados em grades de avaliação. Nesse processo de mensuração, encontram-se algumas variáveis que interferem no resultado do desempenho oral dos examinandos. Discorremos sobre essas variáveis no capítulo 2.

Frente às lacunas e desafios que se instauram no cenário da avaliação linguística e considerando a relevância do Celpe-Bras como instrumento de promoção e difusão da língua portuguesa, parafraseamos o questionamento apresentado por Bachman (1990) para sustentar o nosso posicionamento de que é preciso investigar sobre a confiabilidade dos resultados do exame. Dessa forma, explicitamos o **problema de pesquisa**: *o comportamento avaliativo dos atribuidores de nota pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste?*

Na tentativa de responder a essa indagação, o **objetivo geral** da pesquisa é analisar de que maneira a confiabilidade tem relação com o comportamento avaliativo.

Os **objetivos específicos** são:

1. definir o que é *comportamento avaliativo*;
2. descrever o processo de aplicação e avaliação das provas orais do exame, abordando os critérios que compõem a grade;
3. analisar o comportamento avaliativo a partir das diferentes instâncias avaliativas;
4. estimar a confiabilidade dos resultados da prova oral de sete edições consecutivas do exame.

Trata-se de uma pesquisa que dialoga com a Linguística Aplicada e a Estatística, na medida em que aborda um instrumento de avaliação linguística sob a ótica de uma **abordagem quantitativa**. Assim, esta pesquisa pode contribuir com as discussões que permeiam os processos de aplicação e de avaliação da prova oral do Celpe-Bras, ao buscar auxílio nesse diálogo que, em última análise, permite comparar o comportamento de avaliação e os resultados do teste.

---

variáveis que podem exercer influência no processo avaliativo dizem respeito ao instrumento em si, às formas de aplicação e de avaliação, às atitudes dos avaliadores, ao candidato, ou seja, aos diversos fatores que podem interferir no processo.

Para o cumprimento dos objetivos, o **percurso metodológico** inclui:

1. solicitação, ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), dos dados da aplicação da prova oral de sete edições consecutivas do exame;
2. apresentação e caracterização do universo que compõe a população de estudo, por meio da explicitação das três instâncias de avaliação das provas orais, a saber: 1ª instância: nota atribuída no posto aplicador; 2ª instância: nota atribuída no processo de compatibilização, quando da existência de notas discrepantes; 3ª instância: nota de consenso, quando da existência de notas discrepantes na 2ª instância;
3. descrição e análise do comportamento avaliativo dos atribuidores de nota por meio de métodos estatísticos;
4. análise preliminar da dimensionalidade da escala avaliativa, por meio da Análise dos Componentes Principais (ACP);
5. estimativa da confiabilidade dos resultados do exame, a partir da verificação:
   (i) da consistência interna dos itens da escala, por meio do cálculo do coeficiente *Alfa de Cronbach*, e
   (ii) do nível de concordância entre os avaliadores, por meio do cálculo do coeficiente *Kappa*.

A **justificativa** da pesquisa encontra-se ancorada na necessidade de preencher lacunas na área da avaliação, contribuindo, assim, para o aprimoramento dos processos que envolvem o exame. Essas lacunas dizem respeito, especialmente, à necessidade de se atrelar os aspectos teóricos relativos às qualidades desejáveis dos testes aos dados empíricos do exame.

A **organização da tese** encontra-se configurada em cinco capítulos, além das considerações iniciais e finais, sendo que, a cada início de capítulo, é apresentado um sumário esquemático dos seus temas mais relevantes.

No capítulo 1, apresentamos uma discussão sobre avaliação em larga escala, dando ênfase aos testes utilizados para mensurar habilidades linguísticas.

No capítulo 2, tratamos sobre o exame Celpe-Bras, dando maior foco em sua Prova Oral, explicitando o seu formato, suas características e critérios de avaliação, além de apresentarmos algumas pesquisas que dialogam entre si pelo objeto da interação face a face.

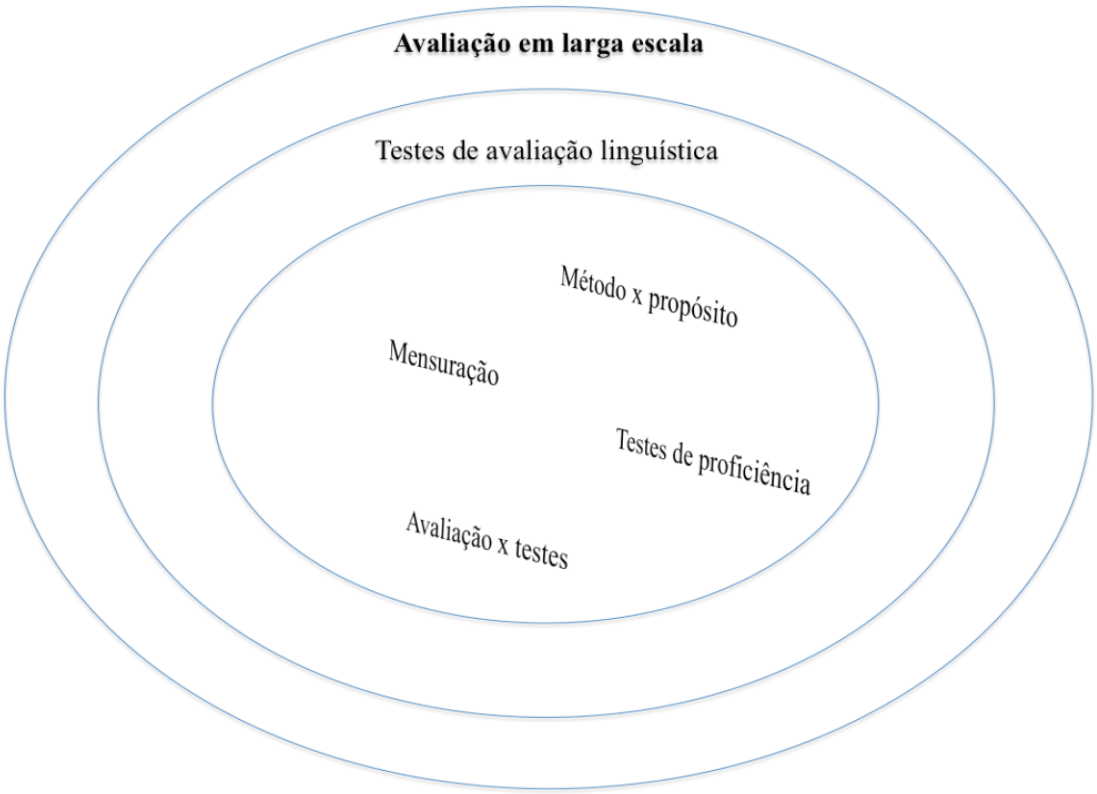
No capítulo 3, discutimos sobre as qualidades desejáveis dos testes, focalizando a confiabilidade, o que representa o referencial teórico da pesquisa.

O capítulo 4, por sua vez, é o espaço dedicado às questões metodológicas, em que a pesquisa é descrita quanto a natureza, características e procedimentos, além do detalhamento do percurso traçado para coleta e análise dos dados.

No capítulo 5, dividido em quatro partes, são apresentadas a análise dos dados e a discussão dos resultados.

Ao fim dos capítulos, há as referências, os anexos e os apêndices, estando estes organizados por parte do capítulo 5 e por tema.

CAPÍTULO 1
AVALIAÇÃO EM LARGA ESCALA



1 AVALIAÇÃO EM LARGA ESCALA

Neste capítulo, fazemos uma discussão sobre avaliação em larga escala, dando ênfase aos testes[3] de língua.

A avaliação em larga escala pode ser considerada um dispositivo que revela características macro do ensino e da aprendizagem e deve ser utilizada para a elaboração e implementação de políticas públicas voltadas para a melhoria do sistema educacional.

Para Rosa Becker (2010, p. 3), a avaliação não é um fim em si mesmo, mas um instrumento que deve ser utilizado para corrigir rumos e pensar o futuro e, nesse sentido, juntamente com as avaliações decorrentes do processo avaliativo, devem ser criados e utilizados instrumentos que contribuam para a solução dos problemas sociais que afetam a população em idade escolar.

Bauer, Alavarse e Oliveira (2015), ao tratarem das avaliações em larga escala nas reformas educacionais, consideram que o papel de destaque da avaliação padronizada nas políticas públicas educacionais, geralmente, aparece justificado pela necessidade de mudanças nas concepções de gestão na educação *pari passu* à mudança nas organizações em geral (BAUER; ALAVARSE; OLIVEIRA, 2015, p. 1370).

No Brasil, o Ministério da Educação (MEC) delega ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), uma autarquia federal, a competência de realização de avaliações em larga escala, como as listadas no Quadro 1, a seguir.

[3] Neste trabalho, os termos *teste* e *exame* são intercambiáveis.

Quadro 1 - Algumas avaliações realizadas pelo INEP

<table>
<tr><th colspan="2">Avaliações no Brasil</th></tr>
<tr><th>Educação Básica</th><th>Educação Superior</th></tr>
<tr><td>- Encceja (Exame Nacional para Certificação de Competências de Jovens e Adultos) - Ofertado a jovens e adultos residentes no Brasil como oportunidade de concluir seus estudos e buscar certificação do Ensino Fundamental e do Ensino Médio.
- Enem (Exame Nacional do Ensino Médio) - Utilizado para avaliar o desempenho do estudante ao final da Educação Básica e como mecanismo de seleção para o ingresso no Ensino Superior e a outros programas do MEC.
- Prova Docente - A Prova Nacional de Concurso tem o objetivo principal de subsidiar os Estados, o Distrito Federal e os Municípios na realização de concursos públicos para a contratação de docentes para a educação básica. Trata-se de uma prova anual, aplicada de forma descentralizada em todo o país para os candidatos ao ingresso na carreira docente das redes de educação básica.
- Provinha Brasil - É uma avaliação diagnóstica que visa investigar as habilidades desenvolvidas, em Língua Portuguesa e Matemática, pelas crianças matriculadas no 2º ano do Ensino Fundamental das escolas públicas brasileiras.
- Saeb (Sistema Nacional da Educação Básica) - A cada dois anos, investiga, por meio de testes cognitivos e questionários, os principais envolvidos no processo educativo e oferece informações sobre estudantes, professores, dirigentes educacionais e seus respectivos sistemas de ensino e escolas.</td><td>- Revalida (Exame Nacional de Revalidação de Diplomas Médicos) - Orienta o reconhecimento de diplomas de medicina emitidos por instituições de educação superior estrangeiras. Possui duas etapas: uma prova objetiva com questões de múltipla escola e discursivas e uma avaliação de habilidades clínicas.
- Saeg (Sistema de Avaliação de Escolas de Governo) - Avaliação in loco das Escolas de Governo com a finalidade de promover melhoria da qualidade, o aumento permanente de sua eficácia institucional e efetividade acadêmica e social, e a promoção do aprofundamento dos compromissos e responsabilidades sociais.
- Sinaes (Sistema Nacional de Avaliação da Educação Superior) - Tem o objetivo de assegurar o processo nacional de avaliação das Instituições de Educação Superior. Seus instrumentos são:
Anasem (Avaliação Nacional Seriada dos Estudantes de Medicina).
Avaliação externa (in-loco).
Avaliação interna (auto-avaliação).
Enade (Exame Nacional de Desempenho de Estudantes).</td></tr>
<tr><th colspan="2">Avaliações internacionais</th></tr>
<tr><td colspan="2">- ARCU-SUL (Sistema de Acreditação Regional de Cursos de Graduação no Mercosul - Resultado de um acordo entre os Ministérios de Educação de Argentina, Brasil, Bolívia, Chile Paraguai e Uruguai, o sistema executa a avaliação e acreditação de cursos universitários.
- Celpe-Bras (Exame de Certificação de Proficiência em Língua Portuguesa para Estrangeiros) - Único certificado brasileiro de proficiência em português como língua estrangeira reconhecido oficialmente. É aplicado no Brasil e em outros países com o apoio do Ministério das Relações Exteriores.
- Encceja Exterior (Exame Nacional para Certificação de Jovens e Adultos) - Ofertado a jovens e adultos residentes no exterior como oportunidade de concluir seus estudos e buscar certificação do Ensino Fundamental e do Ensino Médio. É aplicado com o apoio do Ministério das Relações Exteriores.
- ERCE/LLCE (Estudos Regionais Comparativos) - Por meio de uma rede de diretores regionais de avaliação educacional na América Latina e no Caribe (Orealc/Unesco), avaliam a qualidade da educação no Ensino Fundamental. O foco é melhorar a igualdade das situações educacionais.
- PISA (Programme for International Student Assessment) ou Programa Internacional de Avaliação de Estudantes - Iniciativa de avaliação comparada coordenada pela Organização para Cooperação e Desenvolvimento Econômico (OCDE), aplicada de forma amostral a estudantes matriculados a partir do 8º ano do Ensino Fundamental na faixa etária dos 15 anos.</td></tr>
</table>

Fonte: adaptado de <www.inep.gov.br>. Acesso em: 3 jan. 2018.

Como se pode notar, há uma gama de avaliações de responsabilidade do governo brasileiro, cada uma com objetivos e características específicos[4]. E é nessa seara que muitos pesquisadores dedicam-se a investigar as avaliações sob perspectivas diversas.

Castro (2009) considera que os sistemas de avaliação em larga escala exercem um papel importante como instrumento de melhoria da qualidade da educação e podem prover informações estratégicas para aprofundar o debate sobre as políticas educacionais de um país e mostrar o que os alunos estão aprendendo, ou o que deveriam ter aprendido, em relação aos conteúdos e habilidades básicas estabelecidos no currículo (CASTRO, 2009, p. 276). A pesquisadora ressalta ainda, assim como também elencado por Rosa Becker (2010), que um dos grandes desafios das políticas educacionais é utilizar de modo eficiente os resultados das avaliações para a melhoria da escola, da sala de aula e da formação de professores.

Sudbrack e Cocco (2014), ao tratarem da educação básica, também entendem a avaliação em larga escala como instrumento de política pública para a melhoria da qualidade da educação, mas questionam, por exemplo, a elaboração de provas homogêneas para todo o país, desconsiderando os conhecimentos culturais e sociais das diferentes regiões.

Uma crítica relacionada à divulgação dos resultados das avaliações em larga escala é apresentada por Franco (2001) e Vianna (2003). Para eles, a apresentação dos resultados, a partir da exploração de técnicas complexas, deveria ser feita com uma linguagem acessível para não especialistas em métodos quantitativos, pois isso possibilitaria, por exemplo, que os professores refletissem sobre a sua própria realidade, com o objetivo de estabelecerem linhas de ação.

Diferentemente dos exames da educação básica e do ensino superior, o Celpe-Bras, foco desta tese, não tem como objetivo principal avaliar a qualidade ou propor melhorias para o ensino brasileiro, pois trata-se de um exame que avalia a proficiência de estrangeiros em língua portuguesa e os resultados são utilizados no meio educacional ou no mercado de trabalho, quando as instituições assim o exigirem. Não obstante, como um exame de larga escala e tendo aplicações periódicas, os seus resultados podem ser úteis para avaliar a sua qualidade, podem gerar um efeito retroativo no ensino de línguas, na formação de professores da área de PLE, além de servirem de base para a realização de estudos científicos.

---

[4] Além desses sistemas de avaliação de responsabilidade do Inep, há, por exemplo, o relativo aos cursos de pós-graduação, sob a responsabilidade da Coordenação de Aperfeiçoamento de Pessoal do Nível Superior (CAPES), que utiliza de indicadores específicos para avaliar o sistema de pós-graduação brasileiro.

Entendemos que os resultados das avaliações em larga escala devam ser capazes de promover mudanças reais e significativas, como a melhora na qualificação e valorização do trabalho docente, atenção especial para os problemas de aprendizagem, promoção de motivação para os estudos, proposição de novos currículos. Independentemente do objetivo das avaliações (avaliar o ensino e/ou a aprendizagem, classificar candidatos, avaliar determinadas habilidades etc.), elas devem servir de instrumento (re)direcionador para a tomada de decisões, sejam pessoais, institucionais ou de políticas públicas. Devido a isso e, dado o impacto que podem exercer na vida dos sujeitos a que elas se submetem, as avaliações necessitam apresentar resultados de qualidade.

Para Toffoli (2015) e Toffoli, Andrade, Bornia e Quevedo-Camargo (2016), são questões fundamentais para elaboração e aplicação de avaliações em larga escala:

(i) **validade**, um conceito que vem sendo discutido e modificado ao longo do tempo e que consiste em saber se as interpretações e ações sobre os resultados dos testes são justificadas, tanto com base nas evidências científicas como nas consequências sociais e éticas da utilização do teste (TOFFOLI, 2015, p. 54).

(ii) **Confiabilidade**, que está relacionada à consistência dos resultados da avaliação, isso significa, por exemplo, que as pontuações atribuídas por dois avaliadores a uma mesma resposta de teste não sejam demasiadamente diferentes (TOFFOLI, 2015, p. 59).

(iii) **Comparabilidade**, que se refere à possibilidade de se estabelecer comparação entre os indivíduos que se submetem aos testes. No Brasil, o SAEB utiliza uma escala única referenciada, e os diversos estados brasileiros mantêm a mesma matriz de referência, viabilizando a comparação entre os desempenhos dos estudantes em todo o território nacional (TOFFOLI, 2015, p. 63).

(iv) **Justiça**, que está ligada à equidade do teste, ou a possibilidade de garantir a todos os participantes oportunidades iguais, e, para isso, é necessário que os testes sejam imparciais e apropriados para os vários grupos que serão testados (TOFFOLI, 2015, p. 77).

A avaliação em larga escala, portanto, caracteriza-se pela complexidade, devido ao seu grau de responsabilidade social e necessidade de gerar resultados consistentes e significativos para os propósitos a que se destina.

1.1 Testes para avaliação linguística

Num cenário em que os intercâmbios acadêmicos e culturais estão cada vez mais frequentes, a língua instaura-se fortemente como um objeto de trocas, pois é por meio dela

que as relações se estabelecem. É, também, a partir desse objeto, a língua, e com base em posicionamentos políticos e econômicos de alguns países, que são criados instrumentos para avaliação do desempenho linguístico dos sujeitos.

Nos testes de língua, tem-se a língua como o instrumento e como o próprio objeto de avaliação. Eles desempenham um papel importante na vida de muitas pessoas, pois agem como porta de entrada em vários momentos da educação, na busca por um emprego e na transição de um país a outro. São também considerados como um dispositivo para o controle institucional dos indivíduos (McNAMARA, 2000, p. 4).

Segundo McNamara (2000), os testes utilizados para a avaliação de língua diferenciam-se quanto ao método e quanto ao propósito. No que se refere ao método, o pesquisador distingue os testes de lápis e papel dos testes de desempenho. O primeiro método avalia habilidades separadas (gramática, vocabulário, compreensão auditiva e leitora etc.). Os testes de desempenho, ao contrário,

> [...] avaliam as habilidades linguísticas em um ato de comunicação. Testes de desempenho são comumente testes de fala e escrita, em que uma amostra mais ou menos extensa ou da fala ou da escrita é provocada por um examinador e julgada por um ou mais avaliadores treinados, a partir de uma grade de avaliação. Essas amostras são extraídas num contexto de simulação de tarefas do mundo real (McNAMARA, 2000, p. 6).[5]

Ou seja, os dois métodos refletem conceitos diferentes de língua: enquanto um tem uma visão fragmentada, focando habilidades separadas, o outro a entende como interação.

No que se refere ao propósito, de acordo com McNamara (2000), os testes podem classificar-se em dois tipos, quais sejam: testes de conclusão de processo ou testes de rendimento e testes de proficiência. De um lado, os testes de conclusão de processo estão associados à avaliação do processo de aprendizagem, portanto voltados para a avaliação do passado. De outro lado, os testes de proficiência focam em uma situação comunicativa futura.

À medida que avançam os estudos da linguagem, também mudam as abordagens de ensino e aprendizagem de línguas e isso reflete diretamente na área da *avaliação*. Em uma visão estruturalista da linguagem, os testes utilizados para avaliação dos sujeitos refletiam a língua como sistema, com certa tendência em descontextualizar o conhecimento e testá-lo de forma isolada, sendo que essa prática é chamada de testes de pontos discretos (McNAMARA, 2000, p. 14), isto é, medem separadamente as habilidades do sujeito, uma a uma, sendo, por

[5] Todas as traduções são de nossa responsabilidade.

isso, mais objetivos na avaliação. Ainda de acordo com esse autor, dada a necessidade de se avaliar habilidades práticas da língua e, com a chegada do movimento comunicativo, houve a demanda de se criar testes que envolvessem a performance do sujeito de maneira integrada. Daí surgiram os testes integrativos que, ao contrário dos de pontos discretos, consideram a língua como um todo, a integração de várias habilidades, como, por exemplo, levar em consideração a escuta de um áudio para a produção de um texto escrito.

No que toca aos testes de proficiência[6], Bachman (1990) entende que são importantes para a tomada de decisões sobre a competência linguística, quer no contexto da avaliação do desempenho do aluno em programas de línguas, quer para certificar a competência profissional dos professores de línguas (BACHMAN, 1990, p. 6).

Brown e Abeywickrama (2010) consideram importante destacar que os termos *teste* e *avaliação* não são sinônimos. Para eles, enquanto a avaliação é considerada, na prática educacional, um processo contínuo que envolve uma série de técnicas metodológicas, os testes são um subconjunto da avaliação, um gênero de técnicas. Em termos científicos, o teste é um método de medir a habilidade de uma pessoa, o conhecimento ou desempenho em um dado domínio (BROWN; ABEYWICKRAMA, 2010, p. 3). Para esses pesquisadores, os componentes que caracterizam um teste são: (i) é um método; (ii) deve medir; (iii) mede habilidades individuais, conhecimento ou desempenho; (iv) mede o desempenho, mas os resultados referem-se à capacidade do examinando, ou seja, à sua competência e (v) mede em um determinado domínio, a exemplo dos testes de proficiência, testes de pronúncia, testes de vocabulário etc.

Assim como fez McNamara (2000), Brown e Abeywickrama (2010, p. 9-12) tratam de tipos de testes, classificados a partir dos seus propósitos. Para esses autores, os tipos mais comuns são os cinco descritos a seguir:

1. testes de conclusão de processo, cujo objetivo principal é verificar se os objetivos de determinado curso foram atingidos, ao fim de um período de ensino.
2. testes de diagnóstico, cuja finalidade é diagnosticar os aspectos da língua que o estudante precisa desenvolver ou que o curso deve incluir. Em um teste de pronúncia, por exemplo, pode-se identificar as características fonológicas que são difíceis para os aprendizes e, assim, serem incluídos no currículo do curso.

---

[6] Bachman (1990, p. 16) afirma que o termo "proficiência" adquiriu uma variedade de significados e conotações em diferentes contextos e, no livro *Fundamental considerations in language testing*, o autor adota o termo "*language ability*" na maioria das vezes para referir-se a "*proficiency*".

3. testes de nivelamento, cujo objetivo é colocar o estudante em determinado nível de aprendizado. Os testes de conclusão de processo e os de proficiência podem ser desse tipo.
4. testes de proficiência tradicionalmente consistem em itens padronizados de múltipla escolha sobre aspectos gramaticais, lexicais, de compreensão leitora e auditiva[7]. Alguns testes incluem tarefas para medirem a habilidade de escrita, bem como o desempenho oral. Esse tipo de teste chega a um resultado sob a forma de uma pontuação simples, o que, para muitos, é o suficiente para servir de instrumento para aceitar ou excluir alguém de determinada situação. Pelo fato de medirem o desempenho do candidato de uma forma geral, via de regra não são utilizados para fornecerem feedback sobre diagnóstico do sujeito avaliado. Uma questão central nesses testes é como são especificados os construtos de habilidade de língua.
5. testes de aptidão são desenvolvidos para medir a capacidade ou habilidade para aprender uma língua estrangeira *a priori* (antes de iniciar um curso). Foram ostensivamente desenvolvidos para ser aplicados em sala de aula de qualquer língua.

Hughes (2003) também apresenta os quatro primeiros tipos de testes tratados por Brown e Abeywickrama (2010) e, especificamente no que toca aos testes de proficiência, o autor afirma que são elaborados para medir a habilidade linguística de determinada pessoa, independentemente da formação que possa ter tido na língua a ser avaliada. Portanto, o conteúdo deles não se baseia nos objetivos de cursos de línguas, mas no que se espera que o candidato seja capaz de fazer, a fim de ser considerado proficiente. E isso, então, levanta a questão do que significa ser 'proficiente' que, no caso de alguns testes, significa dominar suficientemente o idioma para um propósito específico (HUGHES, 2003, p. 11).

Há diversos testes de proficiência no panorama internacional, a exemplos de: *Certificado de Español Lengua y Uso* (CELU), *Diploma de Español como Lengua Extranjera* (DELE), *Diplôme d'Études en Langue Française* (DELF), *International English Language Testing System* (IELTS), *Test of English as a Foreign Language* (TOEFL) e *Test of English for International Communication* (TOEIC).

[7] Ressaltamos que o Celpe-Bras, foco desta pesquisa, é um exame de proficiência que não reflete as características dadas pelos autores. Trataremos delas mais adiante e, com mais detalhes, no Capítulo 3.

Em se tratando da língua portuguesa (LP), Portugal e Brasil possuem seus próprios exames. Em Portugal, há o Centro de Avaliação de Português Língua Estrangeira (CAPLE) com seus seis exames[8] correspondentes aos seis níveis (A1, A2, B1, B2, C1 e C2) do Quadro Europeu Comum de Referência para as Línguas (QECR).

No Brasil, há o exame que confere o Certificado de Língua Portuguesa para Estrangeiros (Celpe-Bras, a ser tratado no capítulo 2), sendo o único teste de desempenho que avalia a proficiência em português como língua estrangeira reconhecido oficialmente pelo governo brasileiro, e está sob a responsabilidade do INEP. Diferentemente dos exames do CAPLE, o Celpe-Bras avalia, por meio de um único instrumento, cinco níveis de proficiência: 1 (básico - sem certificação), 2 (intermediário), 3 (intermediário superior), 4 (avançado) e 5 (avançado superior), emitindo certificado para os últimos quatro, naturalmente. De acordo com o Manual do Examinando,

> [...] a decisão de se elaborar uma única prova para certificar diferentes níveis de proficiência, contrariamente ao que ocorre em outros exames, baseia-se na premissa de que examinandos/as de todos os níveis são capazes de desempenhar ações em Língua Portuguesa. O que pode variar é a qualidade desse desempenho, dependendo do nível de proficiência do/a examinando/a (BRASIL, 2015a, p. 8).

Além dos exames CAPLE e Celpe-Bras, ainda não há similares em língua portuguesa para estrangeiros nos demais países onde essa língua é a oficial. Oliveira (2013), por um lado, questiona a necessidade de haver mais de um exame desta natureza, já que a LP é falada nos países da Comunidade dos Países de Língua Portuguesa (os nove Estados-Membros pertencentes à CPLP são Angola, Brasil, Cabo Verde, Guiné-Bissau, Guiné Equatorial, Moçambique, Portugal, São Tomé e Príncipe e Timor Leste), não *pertencendo* a língua a nenhuma nação específica e, portanto, apresentando nuances que refletem as culturas desses países, o que não impede a intercompreensão entre os povos.

Por outro lado, Diniz (2010), ao abordar o "discurso de brasilidade no Celpe-Bras" e tratar de uma "instauração de um litígio entre Portugal e Brasil", apresenta trechos de uma entrevista realizada com Scaramucci, em que uma das perguntas diz respeito a uma possível

---

[8] Os exames do CAPLE são: (i) Acesso ao Português, (ii) Certificado Inicial de Português Língua Estrangeira (CIPLE), (iii) Diploma Elementar de Português Língua Estrangeira (DEPLE), (iv) Diploma Intermediário de Português Língua Estrangeira (DIPLE), (v) Diploma Avançado de Português Língua Estrangeira (DAPLE) e (vi) Diploma Universitário de Português Língua Estrangeira (DUPLE). Além desses testes, o CAPLE ainda disponibiliza o CIPLE, o DEPLE e o DIPLE na versão escolar, para jovens de 12 a 15 anos de idade. Informações disponíveis em http://caple.letras.ulisboa.pt/pages/view/21 Acesso em 26 set 2016.

unificação dos exames, de forma que envolvesse todos os países lusófonos. A entrevistada afirma que o Instituto Camões propôs uma unificação entre os exames, sendo que o Brasil não aceitou a proposta devido ao fato de os instrumentos serem muito diferentes: os exames CAPLE são "de base estruturalista [...], com uma visão de linguagem tradicional". Além disso, o Brasil considera o Celpe-Bras como "um instrumento de política linguística do Brasil... da variedade brasileira [...] uma questão de identidade da nossa língua... da nossa variedade..." (DINIZ, 2010, p. 126). A entrevistada ainda ressalta que o exame brasileiro, por suas características e pela visão de linguagem subjacente, acaba por influenciar o ensino da língua portuguesa.

O assunto abordado nessa entrevista relatada por Diniz (2010) foi ao encontro do que tratou Coura-Sobrinho, em um artigo publicado em 2006 sobre o sistema de avaliação do Celpe-Bras, sobre a concepção de língua subjacente ao exame:

> [...] uma análise superficial de exames internacionais de proficiência em línguas permite perceber que a visão de língua subjacente à sua concepção difere da visão pragmática e discursiva a que se propõe o exame Celpe-Bras, que procura avaliar a capacidade de uso do português em situações reais do dia-a-dia (COURA-SOBRINHO, 2006, p. 127).

Constata-se, portanto, que a visão de língua adotada por um exame é o ponto central a partir do qual todas as relações se estabelecem: o quê avaliar, de que forma avaliar e a partir de qual construto teórico avaliar. Trata-se de uma relação de mão dupla, em que a *identidade da nossa língua* é refletida no exame e em que o exame reflete a *nossa variedade*, nos termos de Scaramucci ora apresentados.

Ao estabelecerem um diálogo entre os exames CELU e Celpe-Bras, Schlatter *et al* (2009) também consideram a implementação desses dois instrumentos como ações de política linguística, instrumentos esses que:

> [...] nascem de uma concepção compartilhada de língua em que a atenção às necessidades dos falantes se mostra central. Mediante exames diretos, que avaliam o grau de desempenho através de avaliações qualitativas, propõem-se a certificar que um falante pode "atuar no mundo em vários contextos de uso da língua", o que é condizente com uma política educativa que pretende formar cidadãos instruídos, bilíngues, de acordo com suas necessidades, e que sejam capazes de desempenhar-se em distintas práticas sociais com desenvoltura e flexibilidade intercultural [...] (SCHLATER *et al*, 2009, p. 109).

Os exames de proficiência são elaborados a partir de construtos bem definidos e para fins específicos, sendo que cabe às instituições educacionais ou mesmo as empresas que terão estrangeiros em seus quadros estabelecer as normas de aceitação dos certificados. Cada um

dos exames tem seu nível ou níveis de proficiência definidos de acordo com a situação específica para o qual foi proposto, eliminando a possibilidade de que possa ser considerado válido em outros contextos de uso ou com funções outras além daquelas para as quais foi elaborado (SCARAMUCCI, 2000, p. 15).

Além desses testes que proporcionam o intercâmbio linguístico-cultural entre sujeitos, seja no meio acadêmico, seja no mercado de trabalho, há no Brasil, também, dois testes de proficiência em língua inglesa bem específicos relativos à aviação civil. Um deles é o *Santos Dumont English Assessment* (SDEA)[9], cujo objetivo principal é o de viabilizar maior segurança nas comunicações entre pilotos, controladores de tráfego e operadores de estações aeronáuticas. Trata-se de um teste direcionado a um público restrito, os pilotos, e que avalia a habilidade do candidato em falar e compreender a língua inglesa dentro de contextos relacionados ao trabalho. Incluem-se aí situações de rotina, imprevistas e de emergência, todas elas apropriadas ao contexto operacional [...] (BRASIL, 2014a, p. 2).

O outro é o Exame de Proficiência em Inglês Aeronáutico do SISCEAB (EPLIS)[10]. Enquanto o SDEA é voltado para pilotos de aeronaves, o EPLIS tem como público-alvo os controladores de tráfego aéreo, operadores de estações aeronáuticas e coordenadores e operadores de busca e salvamento. De acordo com o manual do candidato, trata-se de um exame de proficiência desenvolvido para avaliar o uso da língua inglesa em contexto aeronáutico por profissionais de determinadas áreas do Sistema de Controle do Espaço Aéreo Brasileiro (SISCEAB), especialmente em situações que não estejam previstas na fraseologia (BRASIL, 2017a, p. 4).

---

[9] De acordo com informações constantes do site da Agência Nacional de Aviação Civil (ANAC), "no Brasil, país-membro da Organização de Aviação Civil Internacional (OACI), compete à ANAC, autoridade de aviação civil, certificar a proficiência linguística de pilotos de avião, helicóptero, aeronave de decolagem vertical e dirigível, em operações aéreas envolvendo aeronave civil brasileira fora da jurisdição do espaço aéreo brasileiro. Para que tais pilotos comprovem sua proficiência linguística, deverão demonstrar as habilidades de falar e compreender a língua inglesa, submetendo-se ao exame de proficiência linguística elaborado pela ANAC, o Santos Dumont English Assessment (SDEA)". Disponível em: <http://www.anac.gov.br/assuntos/setor-regulado/profissionais-da-aviacao-civil/processo-de-licencas-e-habilitacoes/proficiencia-linguistica> Acesso em 30 ago. 2016.

Para conhecer a estrutura do exame SDEA, sugerimos o acesso a http://www.anac.gov.br/assuntos/setor-regulado/profissionais-da-aviacao-civil/paginas-complementares/santos-dumont-english-assessment-sdea

[10] Para mais informações sobre o exame, sugerimos o acesso a http://eplis.icea.gov.br/public_html/EPLIS_wp/

1.2 Conceitos fundamentais sobre avaliação de proficiência linguística

Precisões acerca de terminologias utilizadas na área da avaliação são importantes para as discussões que ora propomos. No universo dos testes, do ponto de vista comunicativo da linguagem, as entrevistas orais com interação com pessoas reais supriram uma necessidade de pôr fim à artificialidade inerente às máquinas.

De acordo com Brown (2005), a entrevista oral é uma simulação da capacidade do aprendiz em interagir em segunda língua num evento comunicativo autêntico, utilizando diferentes componentes da sua competência comunicativa. E essa entrevista oral, segundo a pesquisadora, engloba enorme complexidade e potencial de variabilidade por parte do entrevistador, o que pode comprometer a imparcialidade e a generalização de conclusões a que podemos chegar a respeito dos candidatos.

A entrevista oral, então, é caracterizada por uma interação que ocorre entre um (ou mais) avaliador e um (ou mais) examinando. Por meio de perguntas feitas pelo avaliador, o examinando responde, demonstrando a sua habilidade oral, e o avaliador atribui determinada nota a partir de critérios estabelecidos em uma grade de avaliação. Dito de outra maneira, é um teste usado para avaliar a habilidade de fala de um aprendiz. Consiste em tarefas comunicativas autênticas, por meio das quais ele interage com um avaliador (SEGALOWITZ; FREED, apud LOEWEN; REINDERS, 2011, p. 129). Como tratado anteriormente, cada teste de proficiência adota, quando o caso, seus próprios procedimentos de entrevista oral a partir do construto teórico que o subjaz. As particularidades da entrevista oral do exame Celpe-Bras são apresentadas no Capítulo 2.

Refletir sobre instrumentos de avaliação implica abordar conceitos que permeiam a atividade de avaliar, como, por exemplo, os propostos por Bachman (1990; 2004) e Bachman e Palmer (2010): *measurement*, *test*, *evaluation* e *assessment*.

Measurement, ou mensuração, segundo Bachman (1990, p. 18-20), é, nas ciências sociais, o processo de quantificação das características de determinada pessoa, de acordo com regras e procedimentos explícitos. E essa definição inclui três traços distintos: (i) *quantificação*, que envolve a atribuição de números, o que se diferencia de descrições qualitativas; (ii) *características*, que, na área de testes de língua, normalmente estão relacionadas a atributos mentais e habilidades de um sujeito, como, por exemplo, atitude, inteligência, motivação, fluência etc. e, por fim, (iii) *regras e procedimentos*, fatores que devem nortear o processo de quantificação, ou seja, uma dada habilidade deve ser observada da mesma maneira por outro avaliador, em outros contextos e com outros indivíduos.

No que toca ao conceito de teste, Bachman (1990), baseado na definição de Carroll, 1968), afirma que é um instrumento de mensuração concebido para obter uma amostra específica do comportamento de um indivíduo (BACHMAN, 1990, p. 20), comportamento esse tratado posteriormente pelo autor (2004, p. 9) como desempenho. Como um tipo de mensuração, o teste necessariamente quantifica características de um indivíduo de acordo com procedimentos explícitos.

Evaluation, ou avaliação, pode ser definida como uma escolha sistemática de informações com o objetivo de tomar decisões (WEISS, 1972 apud BACHMAN, 1990, p. 22) e a decisão tem maior probabilidade de estar correta quanto mais for confiável e relevante a informação, que não necessariamente precisa ser quantitativa. Bachman (1990) estabelece uma comparação entre teste e avaliação e afirma que um não necessariamente implica o outro.

> Avaliação [*evaluation*] não necessariamente implica testes. Da mesma forma, os testes em si e por si só não são avaliativos. Testes normalmente são utilizados para fins pedagógicos, quer como uma maneira de incentivar os alunos a estudarem, ou como um meio de rever o conteúdo ensinado, caso em que nenhuma decisão avaliativa é feita com base nos resultados de teste. Somente quando esses resultados são usados como base para se tomar determinada decisão é que a avaliação é envolvida (BACHMAN, 1990, p. 22).

Bachman (1990, p. 50) ressalta ainda que outros termos usados como sinônimos são *assessment* e *appraisal* que, em língua portuguesa, significam *avaliação*. Segundo o autor, parece não haver na literatura sobre mensuração uma distinção cuidadosa e, por causa da generalidade e ambiguidade dos termos, considera-os variações estilísticas de *evaluation* e *test*.

Para o termo *assessment*, Bachman (2004) explica que normalmente tem sido utilizado de tantas maneiras diferentes nos campos de teste de língua e mensuração educacional, que parece não haver um consenso das áreas sobre o que exatamente significa. Além de significados variados, há também o fato de outros termos serem frequentemente utilizados como sinônimos para se referirem a *assessment*. Para o autor, portanto:

> [...] **avaliação** [*assessment*] pode ser entendida amplamente como o processo de coleta de informações sobre um determinado objeto de interesse de acordo com procedimentos sistemáticos e substancialmente fundamentados. Um produto ou resultado desse processo, como um escore de um teste ou uma descrição verbal, também são referidos como **avaliação** [*assessment*] (BACHMAN, 2004 p. 7).

Pelo que se pode extrair das definições dadas por Bachman (1990; 2004), a diferença entre *assessment* e *evaluation* configura-se na medida em que o primeiro refere-se ao

processo de coleta de informações, ou ao produto ou ao resultado desse processo e, o segundo, por sua vez, ao que se faz e às decisões que são tomadas a partir das informações coletadas. Dentro do contexto de avaliação (*assessment*), há o processo de quantificação de características de determinada pessoa (*measurement*) e os instrumentos de mensuração para se obter amostras específicas do seu desempenho (*tests*). E, a partir disso, podem ser feitas descrições e avaliações (*evaluation*) em diferentes contextos e para diferentes usos.

O autor ressalta, ainda, que a palavra *evaluation*, que envolve a tomada de decisões e julgamentos de valor, pode ser entendida como um possível uso de *assessment*, embora julgamentos e decisões são normalmente feitos na ausência de informações de *assessment* (BACHMAN, 2004, p. 9). Essas decisões, segundo Bachman (1990; 2004), podem referir-se a microavaliações, no que tange a indivíduos, e macroavaliações, no que tange a programas. Essas macro-avaliações, no nosso entendimento, são uma das funções das avaliações em larga escala, ou seja, a partir dos resultados de determinada avaliação, são feitos julgamentos de valor para a tomada de decisões e (re)definições de políticas públicas, por exemplo, como discutido no início deste capítulo.

Como se pode notar, os termos *assessment* e *evaluation* são distintos entre si, entretanto, parece não haver, em língua portuguesa, outra palavra melhor que as traduza que não seja *avaliação*.

Retomando alguns conceitos discutidos nesta seção e, levando em consideração o exame Celpe-Bras, foco desta tese, podemos afirmar que a entrevista é o instrumento utilizado (*test*) para se obter uma amostra do desempenho oral dos examinandos. No processo de coleta de informações (*assessment*) sobre a proficiência oral desses sujeitos, a partir de procedimentos sistemáticos e fundamentados no construto do exame, há o processo de quantificação, ou seja, de mensuração (*measurement*), para o qual são utilizados regras e procedimentos padronizados. A partir disso, então, pode-se fazer descrições e avaliações (*evaluation*) para servirem de base para a tomada de decisões, sendo que essas avaliações podem relacionar-se ao desempenho do próprio examinando ou também ao próprio exame e aos programas educacionais que o têm como referência. É nesse momento de avaliação (*evaluation*) que podem surgir discussões acerca do impacto e do efeito retroativo gerados pelo exame.

A partir das discussões apresentadas em torno dos conceitos tratados por Bachman (1990; 2004), Bachman e Palmer (2010) consideram que algumas distinções podem ser desnecessárias para efeitos de desenvolvimento e uso da avaliação linguística e que o importante é que o desenvolvedor do teste especifique, de forma clara e explícita, as

condições sob as quais o desempenho do candidato é obtido e os *procedimentos* seguidos para registrar esse desempenho (BACHMAN; PALMER, 2010, p. 20). Segundo os pesquisadores (2010), a avaliação que fazemos e as decisões tomadas a partir dela estão situadas num contexto educacional e social e trazem consequências para os indivíduos, os programas e as instituições. Há que se considerar a existência dos valores inerentes aos indivíduos, às instituições educacionais, às comunidades e aos vários grupos de *stakeholders*[11], bem como leis, regulamentos e políticas que podem reger a maneira como utilizamos uma avaliação e as decisões que podemos tomar. Das consequências que são geradas, pode haver aquelas que são involuntárias ou desconhecidas. Assim, os autores chamam a atenção para o fato de que os desenvolvedores de testes e os seus usuários precisam levar em consideração as consequências possíveis da utilização de uma avaliação e das decisões a serem tomadas pelos diferentes atores nela envolvidos (BACHMAN; PALMER, 2010, p. 25).

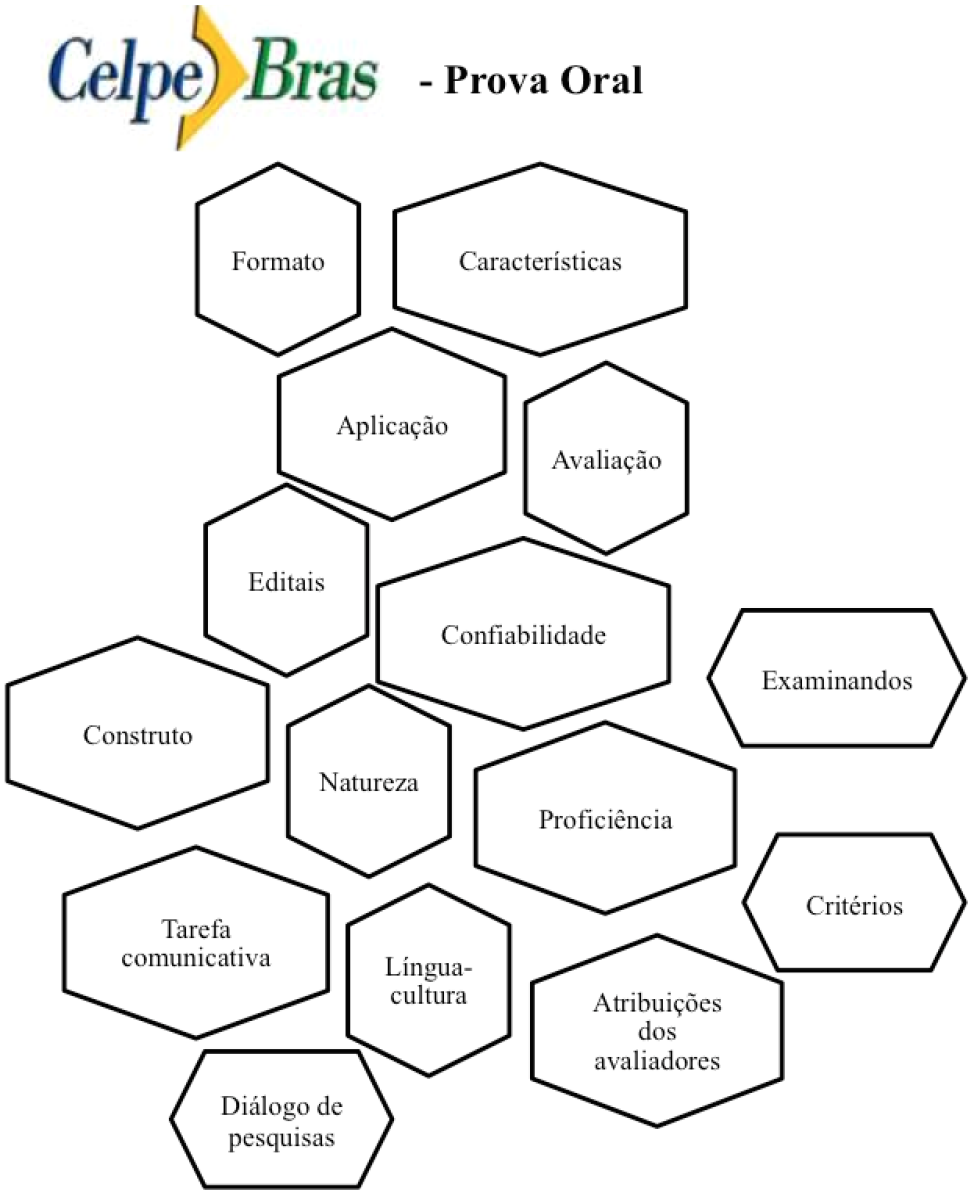
Nesse cenário de avaliação linguística, é relevante destacar a consideração feita por Scaramucci (2014) quando afirma que que os exames, ou testes, são instrumentos potenciais de mudança e de reorientação na sociedade, e chama a atenção para o fato de que a avaliação feita por meio desses instrumentos não nos dá certezas, mas fornecem evidências que devem ser interpretadas. Trata-se, portanto, de um processo passível de erros, os quais devem ser analisados e minimizados, como forma de se manter a justeza com os sujeitos envolvidos.

Neste capítulo, fizemos considerações a respeito das avaliações em larga escala, com ênfase em testes de línguas, exemplificando alguns deles e abordando conceitos fundamentais na área da avaliação. O próximo capítulo dedica-se a apresentar o Celpe-Bras, especialmente a parte oral do exame.

[11] Termo utilizado para referir-se aos diferentes atores envolvidos no processo avaliativo.

CAPÍTULO 2

O EXAME CELPE-BRAS: FOCO NA PARTE ORAL



2 O EXAME CELPE-BRAS: FOCO NA PARTE ORAL

Neste capítulo, apresentamos o exame Celpe-Bras, dando maior ênfase em sua Parte Oral[12], explicitando o seu formato, suas características e critérios de avaliação, além de apresentarmos algumas pesquisas que a têm como foco.

2.1 Informações gerais

O Celpe-Bras foi desenvolvido e outorgado pelo Ministério da Educação do Brasil, é aplicado com o apoio do Ministério das Relações Exteriores e é exigido por instituições de ensino, entidades de classe e empresas como comprovação de proficiência na língua portuguesa. De acordo com o Manual do Examinando (2015),

> (...) podem se inscrever no Exame cidadãos/ãs estrangeiros/as ou brasileiros/as cuja língua materna não seja o português, maiores de 16 anos, com escolaridade equivalente ao ensino fundamental brasileiro, que queiram comprovar, para fins educacionais, profissionais ou outros, a sua proficiência em português nos níveis Intermediário, Intermediário Superior, Avançado e Avançado Superior (BRASIL, 2015a, p. 9).

Dentre esse público que necessita de comprovar sua proficiência linguística por meio do Celpe-Bras, encontram-se os candidatos ao Programa PEC-G (Programa de Estudantes-Convênio de Graduação) e ao PEC-PG (Programa de Estudantes-Convênio de Pós-Graduação), para ingresso em universidades públicas brasileiras. Até 2015[13], outro público que necessitava da comprovação de sua proficiência linguística eram os médicos estrangeiros, para cumprimento de um dos requisitos para revalidação do diploma junto ao Conselho Federal de Medicina.

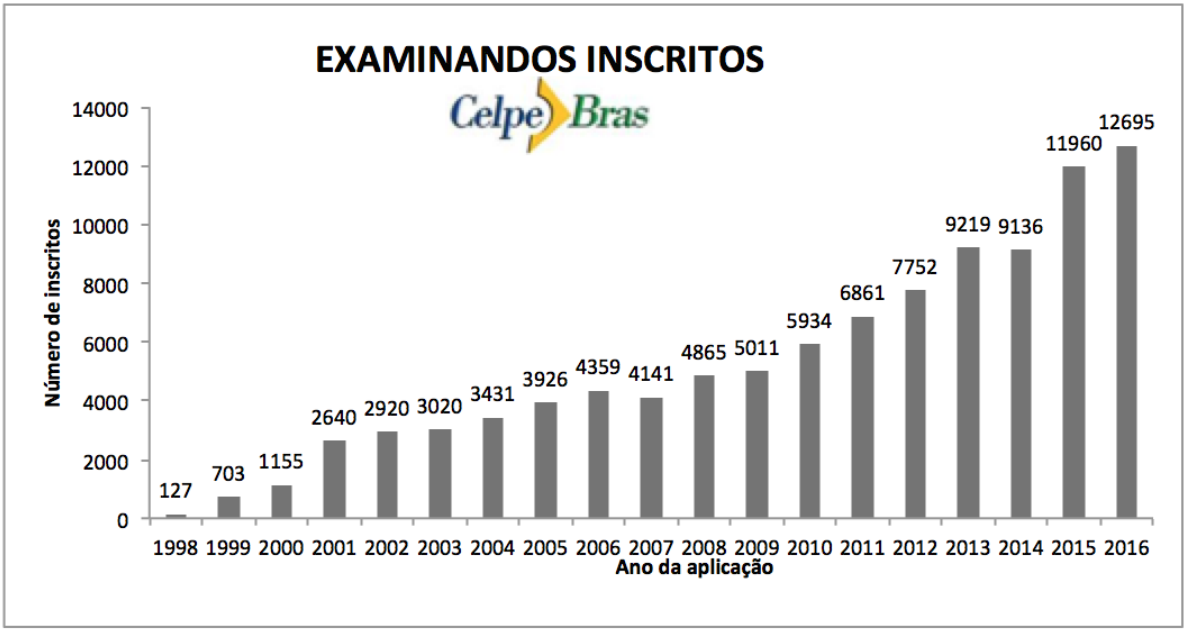
---

[12] Neste trabalho, é dado foco à Parte Oral do exame. Para saber detalhes sobre a Parte Escrita, sugerimos a leitura de Damazo (2012), disponível em <http://www.posgraduacao.cefetmg.br/cefet-mg-posling/index.php/pt/dissertacoes>

[13] Na página eletrônica do Conselho Federal de Medicina (portal.cfm.org.br), consta um Informe Jurídico a respeito da exigência da certificação do Celpe-Bras para a revalidação de diploma médicos estrangeiros. De acordo com a Circular CFM nº 018/2016 – SEJUR, de janeiro de 2016, "o CFM informa aos Conselhos Regionais de Medicina o cumprimento da decisão judicial proferida nos autos do AGRAVO DE INSTRUMENTO Nº *0028271-72.2015.4.03.0000, de modo que está afastada* a exigência da apresentação de certificado de exame de proficiência em língua portuguesa, como condição de inscrição de médicos perante os Conselhos Regionais de Medicina, mantendo-se a r. decisão quanto à suspensão da aplicação da Resolução CFM nº 1831/08 e do art. 2o, parágrafo único, da Resolução CFM nº 1832/08". Informe Jurídico disponível em <http://portal.cfm.org.br>. Acesso em: 16 abr. 2017.

Aplicado desde 1998, o crescimento e a expansão do Celpe-Bras sinalizam sua relevância quando se fala em instrumento de avaliação. O Gráfico 1 e a Figura 1 mostram informações relativas à quantidade de inscritos e aos postos aplicadores credenciados.

Gráfico 1 – Quantidade de inscritos no Exame Celpe-Bras



Nota: os dados do período de 1998 a 2011 foram extraídos de Damazo (2012). Já os de 2012 a 2016, do site do Inep disponível em <http://inep.gov.br/web/guest/acoes-internacionais/celpe-bras?p_p_id=1_WAR_webformportlet_INSTANCE_wOyL7RIB1mAO&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=_118_INSTANCE_iSJV3WO1cuKl__column-1&p_p_col_count=1&_1_WAR_webformportlet_INSTANCE_wOyL7RIB1mAO_javax.portlet.action=saveData>. Acesso em: 10 abr. de 2017.

Figura 1 – Localização geográfica dos postos aplicadores do Celpe-Bras

Fonte: www.sri.cefetmg.br
Notas: 1- mapa elaborado por Mariana Ferreira Valentin da Silva, como uma das atividades da pesquisa de iniciação científica (PIBIC) intitulada *A confiabilidade no exame brasileiro para certificação de proficiência em língua portuguesa para estrangeiros (Celpe-Bras)*. A referida pesquisa foi desenvolvida no CEFET-MG, sob a orientação do Prof. Dr. Jerônimo Coura-Sobrinho e da Ma. Liliane de Oliveira Neves, e contou com o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG);
2- no link a seguir, é possível visualizar o mapa de forma a obter informações de cada posto aplicador: responsável, endereço da localização, telefone e e-mail para contato, endereço do site e capacidade de inscritos. Além disso, por se tratar de um mapa hospedado no *Google Maps*, é possível traçar rotas de deslocamento (https://www.google.com/maps/d/embed?mid=1UyAQVJoyGVm73p_kxDZc4mw7LZo&ll=5.644195492725554%2C-15.31215455624988&z=2).

Tanto em relação ao número de inscritos quanto ao de postos aplicadores credenciados, nota-se um aumento considerável ao longo dos anos. No que se refere aos postos, o crescente interesse de credenciamento de instituições e países faz com que, atualmente, haja 86 postos espalhados no mundo inteiro, sendo 28 no Brasil e 58 no exterior[14].

[14] De acordo com informações do Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP www.inep.gov.br – órgão que administra o exame Celpe-Bras), consulta em 18 de abril de 2017, a quantidade de postos aplicadores no Brasil, por regiões, é: Região Norte (4), Região Centro-Oeste (1), Região Nordeste (3), Região Sudeste (11) e Região Sul (9).

Esse aumento na procura pelo exame reflete (n)a expansão da área de Português como Língua Estrangeira (PLE)[15], o que inclui a elaboração/publicação de materiais didáticos, a crescente necessidade de fortalecer/expandir os cursos de formação de professores da área, o aumento significativo de pesquisas científicas que têm tanto o exame quanto a área do ensino, da aprendizagem e da avaliação como foco, as políticas linguísticas voltadas para a difusão da Língua Portuguesa no mundo. Da mesma forma, instaura-se aí o efeito retroativo do exame, especialmente no que se refere ao ensino da língua para estrangeiros.

Ressaltamos que essa expansão da área de PLE também é evidenciada pela criação, no Brasil, de duas associações de professores[16], quais sejam: a Associação dos Professores de Português Língua Estrangeira do Estado do Rio de Janeiro (APLE-RJ), fundada em 05/11/12, e a Associação Mineira dos Professores de Português como Língua Estrangeira (AMPPLIE), fundada em 16/10/2014, sendo as únicas de caráter regional do país. Essas duas associações têm papel importante na difusão da língua portuguesa, permitindo troca permanente de ideias e experiências, aliando sempre o ensino ao desenvolvimento de pesquisas científicas.

Embora seja relativamente novo, o Celpe-Bras pode ser considerado um exame de proficiência de alta relevância, na medida em que a quantidade de inscritos aumenta a cada ano e decisões importantes são tomadas a partir dos seus resultados. Nesse sentido, é necessário que as ações e os pressupostos teóricos que subjazem ao exame sejam acompanhados por pesquisas que consolidem o seu prestígio e reconhecimento internacional, a partir de discussões sobre os processos de elaboração, aplicação e avaliação, dos seus níveis de proficiência adotados para a certificação de examinandos, por exemplo.

A realização de pesquisas dessa natureza provoca repercussões sobre a prática docente, conforme destacado por Júdice (2000) que, ao tratar da avaliação como "um instrumento de diálogo" e fazer algumas considerações sobre o Celpe-Bras, afirma que a avaliação é entendida como processo episódico que nos oferece um panorama da situação de ensino-aprendizagem de Português-Língua Estrangeira (PLE) em diversos pontos do mundo

---

Postos aplicadores no exterior em: África (7), América Central (4), América do Norte (6), América do Sul (22), Ásia (4), Europa (14), Oriente Médio (1), totalizando 58 postos. Em pesquisa realizada em 2012, Damazo (2012, p. 22) apresentou os seguintes números: 21 postos no Brasil e 46 no exterior.

[15] Sugerimos a leitura de Coelho (2015) para conhecimento de outros termos da área, como, por exemplo, PLA (Português como Língua Adicional), PFOL (Português para Falantes de Outras Línguas), PLH (Português como Língua de Herança), PLAc (Português como Língua de Acolhimento), PL2/PSL (Português Língua Segunda ou Português como Segunda Língua).

[16] Informações sobre a APLE-RJ e a AMPPLIE podem ser obtidas nos seguintes endereços: www.aplerj.com.br e www.ampplie.com.br.

e favorece o diálogo entre os docentes da área, com repercussões sobre sua prática (JÚDICE, 2000, p. 55).

Muitos pesquisadores têm se debruçado sobre o Celpe-Bras, a partir de diversos temas: efeito retroativo e/ou impacto (Silva, 2006; Agossa, 2017), proficiência comunicativa (Scaramucci, 2001), características e implementação do exame (Schlatter, 2006), relações entre livro didático e o exame (Castro, 2006; Jha, 2016), ensino e aprendizagem na avaliação de proficiência (Furtoso, 2011b), política linguística (Diniz e Zoppi-Fontanta, 2006), comparação com o exame CELU (Schlatter *et al.*, 2009), aspectos enunciativos da linguagem (Coura-Sobrinho e Dell´Isola, 2009; Damazo, 2012), descrição de níveis de proficiência de hispano falantes (Schoffen, 2003), gêneros e parâmetros de avaliação (Schoffen, 2009), interculturalidade (Leroy, 2011; Campolina, 2017), competência interacional (Niederauer, 2014), representações do Brasil e enquadramentos temáticos (Lima, 2008; Neves, Agossa e Coura-Sobrinho, 2017), entre outros.

2.2 A aplicação do exame

No que se refere a sua aplicação, todos os examinandos[17] são submetidos a um mesmo teste, que consiste em duas partes, a saber: (i) uma escrita, com duração de 3 horas, composta por duas tarefas que integram a compreensão oral e produção escrita e outras duas que integram leitura e produção escrita; (ii) uma oral, individual, com duração de 20 minutos, em que há a presença de dois avaliadores, um observador (AO) e um interlocutor (AI)[18], sendo que este último interage com o examinando, face a face, a partir das informações constantes da sua ficha de inscrição e de alguns textos utilizados como elementos provocadores[19] da conversa e que tratam sobre assuntos do cotidiano. De acordo com Brasil (2015d), para atuarem como avaliadores, é exigido que os profissionais tenham o português como língua materna ou que tenham alcançado o nível Avançado Superior no exame.

Para obter o certificado, o examinando deve alcançar desempenho satisfatório nas duas partes, sendo que os níveis de proficiência avaliados são: básico (sem certificação),

[17] Embora atualmente o Inep esteja utilizando o termo *participante* nos novos guias/manuais, optamos por utilizar *examinando* para referir-se ao sujeito que presta o exame Celpe-Bras. Com o mesmo sentido, já houve a utilização do termo *candidato*. Tendo em vista a coocorrência desses dois termos nos manuais consultados e nas pesquisas científicas citadas neste trabalho, utilizamo-los como sinônimos.

[18] Além dos termos Avaliador Interlocutor (AI) e Avaliador Observador (AO), também utilizamos os termos entrevistador e observador, respectivamente, para nos referir aos sujeitos avaliadores da parte oral.

[19] Os elementos provocadores são textos curtos utilizados para *provocar* a interação entre o entrevistador e o examinando, de tal forma que seja possível avaliar o desempenho oral do examinando.

intermediário, intermediário superior, avançado e avançado superior, sendo os quatro últimos níveis com certificação.

No que se refere à parte oral do Celpe-Bras, foco deste trabalho, consta do Guia do Participante o seguinte, relativo ao que se entende pela interação face a face e ao que é cobrado do examinando:

> [...] espera-se que o examinando tenha capacidade de conversar, da forma mais natural possível, sobre assuntos do cotidiano e da atualidade veiculados na mídia brasileira. Vale salientar que não se trata de uma entrevista na qual uma pessoa pergunta e a outra responde de forma mecânica, mas sim de uma simulação de conversa em língua portuguesa (BRASIL, 2013a, p. 28).

A interação face a face é dividida em dois momentos: nos primeiros cinco minutos da interação, o examinando deve demonstrar capacidade de conversar sobre questões de natureza pessoal. Nessa etapa, espécie de "quebra-gelo", é criado um ambiente favorável para a interação, de tal forma que o examinando possa revelar a sua capacidade de comunicação, exprimindo entendimento do tema tratado e sua proficiência na língua portuguesa. No segundo momento, com duração de quinze minutos, o examinando tem acesso a três elementos provocadores (EP), textos multimodais, um a cada cinco minutos, cujos temas são discutidos entre ele e o AI, a partir de um roteiro de perguntas previamente elaborado e em função do conteúdo da conversa. Os cinco minutos para cada elemento provocador são destinados à leitura e à interlocução propriamente dita. Os Anexos D e E contêm exemplos de EP e do roteiro de perguntas.

Os EP são escolhidos pelos avaliadores a partir das informações constantes da ficha de inscrição do candidato ao exame e utilizados para desencadear a conversa. Em geral, os elaboradores do exame disponibilizam, a cada edição, um conjunto de vinte elementos provocadores, para que sejam escolhidos três para cada interlocução. Juntamente com o conjunto de EP, é disponibilizado o Roteiro da Interação Face a Face, que contém perguntas pré-definidas para cada elemento, sendo que cabe ao AI adequá-las à conversa. É importante ressaltar que, com exceção da primeira pergunta do Roteiro, obrigatória, as demais podem ser trocadas de ordem, excluídas e/ou ampliadas, em função do desenvolvimento da interlocução.

De acordo com Coura-Sobrinho (2014), os EP, além de serem os motivadores da interação, também promovem compreensões dotadas de aspectos interculturais que podem ser discutidos durante a interlocução. Para isso, então, é preciso, conforme aponta Campolina (2017), que os entrevistadores sejam sujeitos culturalmente sensíveis, evitando julgamentos de valor, para que a entrevista seja bem sucedida.

É na parte oral que o examinando participa de uma interlocução cujo contrato de comunicação, nos termos da Teoria Semiolinguística[20], é bem definido, ou seja, o examinando aceita as condições de aplicação do exame, submetendo-se a uma entrevista, primeiro momento da prova, e a uma conversa sobre tópicos do cotidiano, que caracteriza o segundo momento da prova. É a partir desse contrato preestabelecido, nos dois momentos, que o AI faz perguntas e propõe diálogos e o examinando demonstra aquilo que entendeu do tema, o que sabe da cultura brasileira, da sua própria cultura e, desse modo, revela o seu nível de proficiência oral. Trata-se, portanto, de uma situação de comunicação em que os sujeitos envolvidos, a partir da aceitação tácita de um contrato comunicativo, agem visando a um mesmo objetivo: o da avaliação da proficiência linguística.

O desdobramento do evento comunicativo em *entrevista* e *conversa* constitui uma encenação que tem por finalidade deixar o candidato mais à vontade para interagir, produzindo sua fala de forma espontânea (COURA-SOBRINHO; DELL'ISOLA, 2009, p. 92). De acordo com os pesquisadores, o candidato, nessa situação de avaliação, tem consciência dos riscos gerados por uma performance não satisfatória e, de certa forma, consegue perceber a assimetria existente entre as instâncias enunciativas *avaliador* e *avaliado*, assimetria esta que pode ser salientada pela presença do observador, que também faz a avaliação do desempenho do candidato.

Os autores afirmam, ainda, que as condições físicas da situação de comunicação, ou seja, da situação de avaliação do desempenho oral do candidato, provocam nele uma série de comportamentos que, de certa forma, o impedem de captar a intenção de transformação do gênero entrevista em conversa (COURA-SOBRINHO; DELL'ISOLA, 2009, p. 92) e, a esse fenômeno, chamam de *possível desequilíbrio enunciativo.*

Como se pode notar, o próprio contrato de comunicação estabelecido na parte oral do exame pode ser um fator influenciador no desempenho do examinando, a depender da condução da entrevista, por parte do entrevistador, e da percepção da simetria/assimetria, por

---

[20] De acordo com Damazo (2012), a Teoria Semiolinguística, desenvolvida por Patrick Charaudeau, considera o sujeito como um ser de dupla identidade: um ser social, que tem comportamentos discursivos de acordo com normas estabelecidas pela sociedade e com a situação de comunicação e que é portador de saberes compartilhados; um ser individual, que tem comportamentos estratégicos e que se posiciona a partir dos seus saberes e é, portanto, um sujeito relativamente livre. Há, assim, uma identidade social e uma identidade discursiva que se entrelaçam e determinam uma identidade singular e uma coletiva do sujeito que participa do ato de comunicação.

parte do examinando. Essas duas particularidades, condução da entrevista e simetria/assimetria enunciativa, foram objeto de pesquisa sobre a parte oral do Celpe-Bras, como apresentado ao fim deste capítulo.

2.3 A avaliação do desempenho oral do examinando

Proficiência oral é um construto que não pode ser observado diretamente, sendo, portanto, um construto latente, nos termos de Costa (2011), ou seja, avaliar a proficiência oral implica em considerar evidências empíricas observáveis que possam ser classificadas em notas, conforme sinaliza Fulcher (2003). Diante disso, o exame Celpe-Bras adota seis critérios para indicar determinando nível de proficiência oral do examinando.

A avaliação é feita tanto pelo AI, que conduz a conversa, quanto pelo AO, que observa a interação, com base em duas grades de avaliação distintas (ANEXOS A a C), sendo uma holística e outra analítica, respectivamente, e as notas são atribuídas individualmente, não podendo haver discussão por parte dos avaliadores. Essa *avaliação cega*, feita a partir de dois olhares distintos e individuais, é uma das recomendações que Hughes (2003, p. 50) faz para que fatores externos ao exame sejam minimizados e, assim, seja possível contribuir para a confiabilidade dos seus resultados.

É relevante destacar que esses dois olhares sobre uma mesma interação permitem que o examinando seja avaliado, por um lado, de uma forma global, sistêmica e, por outro, de uma forma analítica, considerando sua autonomia e desenvoltura, o seu grau de contribuição para a manutenção da interação, sua fluência, seu conhecimento de vocabulário e de estruturas da língua, a pronúncia e o nível de compreensão da fala do AI. Esses diferentes olhares podem contribuir para que se chegue a uma atribuição justa de nota.

Os aspectos avaliados na interação face a face são os relacionados a seguir.

> a) Compreensão da fala do entrevistador;
> b) competência para interagir em Língua Portuguesa (o examinando deve apresentar desenvoltura e autonomia durante sua produção oral);
> c) fluência (capacidade de interagir sem interromper o fluxo da conversa);
> d) domínio de vocabulário e de estruturas da Língua Portuguesa (capacidade de usar vocabulário apropriado e estruturas adequadas do português nos diferentes temas abordados);
> e) pronúncia (manter uma pronúncia adequada em relação aos sons, ritmo e entonação da língua portuguesa) (BRASIL, 2013a, p. 28).

Apesar de não compor a lista dos aspectos avaliados, entendemos que a compreensão da leitura dos elementos provocadores ocupa espaço significativo na interação face a face. É

a partir do conteúdo desses elementos que a interação é conduzida e, portanto, faz-se necessária a sua compreensão. As duas grades de avaliação captam os aspectos avaliados e contribuem de igual forma para a composição da nota final do desempenho oral do examinando, que se configura pela média aritmética simples das notas atribuídas pelo AO e pelo AI, numa escala de seis pontos, sendo de 0 a 5.

O AI atribui uma única nota para o desempenho oral global do examinando tomando por base a grade do ANEXO C. Já o AO, atribui uma nota para cada um dos critérios descritos anteriormente, conforme grade dos ANEXOS A e B, sendo que a nota final atribuída por esse avaliador é calculada por meio da ponderação desses critérios, conforme Quadro 2, a seguir.

Quadro 2 - Ponderação dos critérios de avaliação da parte oral (grade analítica)

Critérios de avaliação (grade analítica)	Porcentagem da nota
Compreensão Competência Interacional Fluência	50%
Adequação Lexical Adequação Gramatical	42%
Pronúncia	8%

Fonte: BRASIL (2016, p. 74, adaptado).

Há que se ressaltar que, para uma avaliação com resultados confiáveis, é preciso que os descritores das grades sejam detalhados de forma precisa e que os avaliadores considerem, sobretudo, o construto do exame quando da atribuição das notas. As notas atribuídas são classificadas por nível de proficiência, conforme consta do Quadro 3. O nível do certificado refere-se ao desempenho global do examinando e é determinado pela menor nota entre as partes escrita e oral.

Quadro 3 - Classificação das notas por nível de proficiência

Nível	Pontuação
Avançado Superior	4,26 a 5,00
Avançado	3,51 a 4,25
Intermediário Superior	2,76 a 3,50
Intermediário	2,00 a 2,75
(Básico) Sem certificação	0,00 a 1,99

Fonte: BRASIL, 2016b, p. 73, modificado.

É necessário registrar que Coura-Sobrinho (2006, p. 132), ao dar detalhes sobre o processo de avaliação, apresenta as seguintes faixas para se chegar ao desempenho final do examinando: avançado superior (4,17 a 5,00); avançado (3,34 a 4,16); intermediário superior (2,50 a 3,33); intermediário (1,67 a 2,49) e sem certificação (0 a 1,66). Essa mudança de procedimento, bem como outras apresentadas logo adiante, têm a ver com o que o próprio pesquisador aponta em seu texto:

> [...] desde sua primeira aplicação em abril de 1998, o exame Celpe-Bras tem sofrido alguns ajustes, tanto no processo de elaboração, quanto no da avaliação. Isso tem acontecido em decorrência da experiência acumulada e de pesquisas realizadas em universidades brasileiras, por professores envolvidos no ensino de Português como Língua Estrangeira (COURA-SOBRINHO, 2006, p. 127).

Todas as interações face a face são gravadas em áudio e enviadas para o INEP, via sistema eletrônico, para eventual consulta por parte de outra equipe de avaliadores, quando da existência de notas discrepantes. O Edital 1, de 28/01/2016, apresenta detalhes do que são consideradas notas discrepantes (o que chamamos de *discrepância significativa*) e como ocorre a reavaliação das provas.

> 12.6.6. São consideradas situações de discrepâncias de nota na Parte Oral:
> a) diferença entre a nota obtida por meio da grade holística e a nota obtida por meio da grade analítica for igual ou maior que 1,5 (um vírgula cinco) ponto;
> b) diferença entre a nota da Parte Oral e a nota da Parte Escrita for igual ou maior que 2,0 (dois) pontos;
> c) diferença da nota na Parte Oral e na Parte Escrita implicar em mudança do nível de certificação e a nota final na Parte Escrita for superior à nota da Parte Oral.
> 12.6.7 Nos casos definidos no subitem 12.6.6, a Parte Oral será reavaliada, por meio do áudio gravado por ocasião da Interação Face a Face, por dois corretores de maneira independente. Um desses corretores utilizará a grade de correção analítica e o outro corretor utilizará a grade de correção holística.
> 12.6.8 Nos casos de reavaliações definidos nos subitens 12.6.6, a nota final da Parte Oral será a média entre as notas atribuídas pelos corretores por ocasião da reavaliação definida no subitem 12.6.7.
> 12.6.8.1 Caso seja observada alguma das situações de discrepância, definidas no subitem 12.6.6, entre as notas atribuídas por ocasião da reavaliação de que trata o subitem 12.6.7, o áudio gravado por ocasião da Interação Face a Face será conduzido para uma terceira correção, que atribuirá a nota final da Parte Oral do Celpe-Bras – 2016.1 do examinando, utilizando a grade holística e serão descartadas as notas anteriores (BRASIL, 2016b, p. 74).

As notas discrepantes podem ser geradas por diversas razões, às quais os administradores[21] do exame devem ficar atentos, fazendo reflexões e propondo discussões teóricas e práticas, para que essas diferenças significativas de nota não sejam consideradas interferências negativas na *confiabilidade*. Dentre esses aspectos que podem gerar discrepâncias, podemos inferir alguns, com base no sujeito avaliador: a concepção de língua e de linguagem por ele adotada, o contexto de aplicação/avaliação (Brasil e exterior), o conceito sobre avaliação adotado, o entendimento do construto do exame e da grade de avaliação, o tempo de experiência com a aplicação do exame, a formação e a capacitação e até mesmo o seu cansaço.

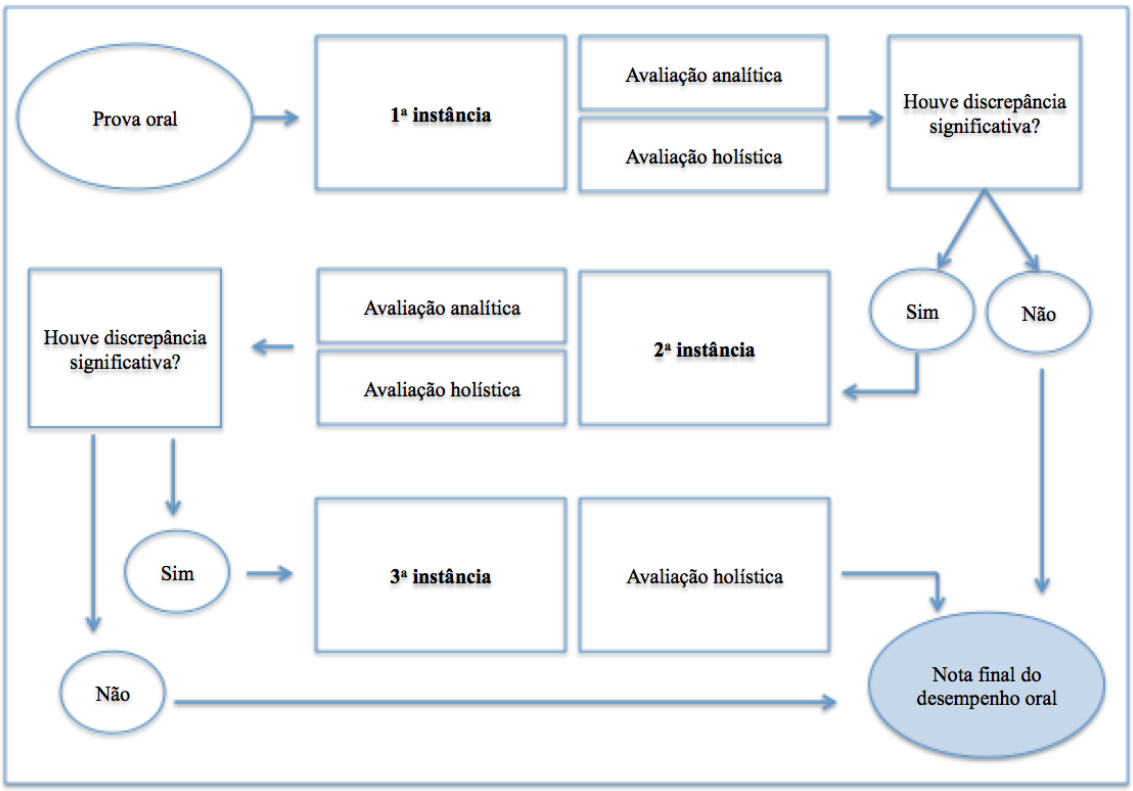
O processo de avaliação da parte oral pode contar com até três instâncias, que são:

- primeira instância: avaliação realizada no posto aplicador por AI e AO;
- segunda instância: avaliação realizada por especialistas colaboradores do INEP, com base nas grades de AI e AO;
- terceira instância: avaliação realizada pelos coordenadores gerais da correção de uma edição específica, com base na grade de AI.

A Figura 2, a seguir, ilustra esse processo.

[21] Utilizamos esse termo para nos referir tanto ao órgão que administra o exame, o Inep, quanto os coordenadores dos postos aplicadores.

Figura 2 - Instâncias do processo de avaliação da parte oral



Fonte: elaborado pela autora, 2018.

Com base na Figura 2, é possível afirmar que trata-se de um processo avaliativo que tenta minimizar as possibilidades de erro de mensuração do desempenho do examinando, tendo em vista que, em se tratando da existência de discrepâncias significativas, uma única interação face a face pode ser avaliada por até cinco sujeitos diferentes, até que se chegue à nota final do desempenho oral. Participa da terceira instância apenas um sujeito, que faz uma avaliação holística do desempenho oral do examinando.[22]

É importante ressaltar que algumas das informações a respeito da atual composição das notas e dos critérios de análise de discrepância foram tornando-se públicas com o passar do tempo. É o que se pode constatar a partir do Quadro 4, a seguir, que apresenta informações referentes às últimas 10 edições do exame[23], sendo de 2013/1 a 2017/2.

[22] No banco de dados disponibilizado pelo INEP para a realização desta pesquisa, as três instâncias são denominadas de posto aplicador, compatibilização e nota de consenso, respectivamente. Considerando-se que não há consenso para atribuição de nota, pois apenas um avaliador atribui nota na terceira instância, sugerimos que as instâncias tenham outras denominações: apenas *primeira*, *segunda* ou *terceira instância*, conforme já deixamos sinalizado na figura 2.

[23] Levantamento feito em março de 2018.

Quadro 4 - Procedimentos para avaliação e reavaliação das interações face a face - edições 2013/1 a 2017/2.

<table>
<tr><th rowspan="2">Edição</th><th rowspan="2">Quem faz a avaliação no posto aplicador?</th><th rowspan="2">Cita os critérios analisados*</th><th rowspan="2">Detalha o cálculo das notas da avaliação analítica?**</th><th colspan="3">Notas discrepantes / valores de referência</th></tr>
<tr><th>Na avaliação da parte oral</th><th>Entre as partes oral (PO) e escrita (PE)</th><th>Quem as analisa?</th></tr>
<tr><td>2013/1</td><td rowspan="3">Dois avaliadores</td><td rowspan="11">Sim</td><td rowspan="7">Não</td><td rowspan="3">Diferença ≥ 2</td><td>Diferença ≥ 2 e implicar mudança de nível.</td><td rowspan="6">Comissão designada pelo Inep. A nota atribuída por essa comissão substitui a dos avaliadores do posto.</td></tr>
<tr><td>2013/2</td><td>a) Diferença ≥ 1,1;
b) implicar mudança de nível nos casos em que a nota da PE for superior à da PO.</td></tr>
<tr><td>2014/1</td></tr>
<tr><td>2014/2</td><td rowspan="4">Entrevistador e observador</td><td rowspan="5">Diferença ≥ 1,5</td><td rowspan="2">Há informações divergentes:
1ª: diferença ≥ 2;
2ª:
a) diferença ≥ 1,1;
b) implicar mudança de nível nos casos em que a nota da PE for superior à da PO.</td></tr>
<tr><td>2015/1</td></tr>
<tr><td>2015/2</td><td rowspan="4">Sim</td><td>a) Diferença ≥ 2;
b) implicar em mudança de nível;
c) quando a nota final da PE for superior à nota da PO.</td><td rowspan="2">Dois outros avaliadores, um utilizando a grade analítica e, o outro, a grade holística. Caso persista a discrepância, há a terceira correção (utilizando-se a grade holística), cuja nota é soberana às anteriores.</td></tr>
<tr><td>2016/1</td></tr>
<tr><td>2016/2</td><td rowspan="3">Avaliador-interlocutor e avaliador-observador</td><td rowspan="2">Quando as situações ocorrerem simultaneamente:
a) Diferença ≥ 2;
b) nota da PO e da PE se enquadrarem em níveis diferentes;
c) quando a nota final da PE for superior à nota da PO.</td></tr>
<tr><td>2017/1</td></tr>
<tr><td>2017/2</td></tr>
</table>

Fonte: elaborado pela autora, 2018.

Nota: este quadro foi elaborado a partir das informações constantes dos editais de abertura de inscrições das edições 2013/1 a 2017/2 (BRASIL, 2013e; 2013f; 2014b; 2014c; 2015b; 2015c; 2016a; 2016b; 2017b; 2017c), publicados no D.O.U. e também disponibilizados em <http://www.ufrgs.br/acervocelpebras/acervo>.

* Compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia.
** 50% para compreensão, competência interacional e fluência; 42% para adequação lexical e adequação gramatical e 8% para pronúncia.

Uma comparação entre os editais citados permite visualizar algumas diferenças de procedimentos adotados na avaliação e reavaliação das provas orais do exame Celpe-Bras, como a denominação dada aos sujeitos envolvidos no processo de avaliação no posto aplicador, o detalhamento do cálculo das notas da avaliação analítica, os valores de referência para se considerar nota com discrepância significativa e os sujeitos que reanalisam as provas com notas discrepantes.

Inferimos que essas alterações possam ter relação com o aprimoramento do processo de avaliação, o que pode refletir, de certa forma, na qualidade de *praticidade* do exame. E, também, que a maior publicidade das normas é em função da necessidade de dar mais transparência ao processo, possibilitando aos examinandos saber mais detalhes dos parâmetros que norteiam a avaliação do seu desempenho, o que, também, tem a ver com a ética do processo.

Esse ato de dar mais publicidade às informações vai ao encontro do que tratam Bachman e Palmer (1996) ao apresentarem, entre os componentes de instrução de um teste, o *método de pontuação*. Segundo os autores,

> [...] a fim de permitir que os candidatos entendam o que deles é esperado e, consequentemente, que eles mostrem o seu melhor desempenho, eles precisam saber como suas respostas serão avaliadas. Se o teste incluir várias partes que serão pontuadas da mesma forma, os critérios para correção podem ser indicados nas instruções gerais. Se diferentes partes forem pontuadas de forma diferente, os critérios devem ser apresentados em instruções específicas de cada parte (BACHMAN; PALMER, 1996, p. 189).

Ao tornarem públicas todas essas informações em destaque, constantes dos Editais, é possível que haja um efeito retroativo bem específico: os candidatos podem se preparar melhor levando-se em conta os pesos atribuídos a cada descritor da grade analítica. Isso, consequentemente, pode ter consequências na reformulação de cursos preparatórios de candidatos ao exame.

Embora as particularidades de composição de nota sejam publicadas para os candidatos, elas não são, ou não foram, até o momento, objeto de discussão nos eventos de capacitação dos aplicadores/avaliadores do exame[24]. Entendemos ser pertinente que, nesses eventos, seja dado foco aos critérios, para que, assim como os candidatos, os avaliadores

[24] Participamos dos eventos para a aplicação/avaliação das provas das edições 2016/1, 2016/2 e 2017/1.

também tenham ciência de como a nota do AO é calculada e como são reavaliadas as provas que apresentam notas discrepantes.

Inferimos, também, que a publicidade para os candidatos, cada vez mais detalhada, das informações do processo de avaliação deva-se ao interesse do Inep em justificar a sua decisão em não prever fase para impetrar recurso aos resultados. Nos editais ora mencionados, há a seguinte informação: "O Inep considera que a metodologia empregada na correção das provas contempla recurso de ofício". Ao fazermos uma busca por editais anteriores[25], vimos que essa informação passou a ser incorporada nos editais de abertura de inscrições a partir da edição 2012/1. Possivelmente, anteriormente a esse período, era aceito que os examinandos contestassem os seus resultados.

O Guia do Examinando (versão simplificada) assim explica o significado de *recurso de ofício*:

> O Celpe-Bras não prevê interposição de recursos, pois adota o sistema de "Recurso de Ofício". O ***recurso de ofício*** é uma espécie de recurso automático, obrigatório, por parte da Administração, não havendo necessidade de o examinando solicitá-lo. Esta modalidade de recurso ocorre ao longo do processo de correção, garantindo que todos os examinandos gozem do direito de terem suas provas avaliadas segundo o mesmo critério (BRASIL, 2013d, p. 8).

Esse "mesmo critério", de que trata o Guia, refere-se aos procedimentos de avaliação e reavaliação das provas (número de avaliadores e procedimentos para análise de notas discrepantes), sendo todos eles acompanhados por supervisores do Inep e da Comissão Técnica do Celpe-Bras[26]. Mas é preciso ressaltar que, embora a Administração Pública recorra de ofício, a regra do edital não afastaria a análise do judiciário, via mandado de segurança, caso haja flagrante ilegalidade.

Considerando-se que o construto do Celpe-Bras é o fator que se encontra no centro de todo o processo de aplicação e avaliação para se garantir resultados confiáveis e é definidor, portanto, de todas as suas ações, apresentamos considerações a respeito da natureza do exame, do conceito de tarefa, do de proficiência e de língua-cultura que o subjazem.

O Celpe-Bras é um exame de natureza comunicativa, que avalia a capacidade que o examinando tem de usar a língua em determinada situação de comunicação. Nesse sentido, a

---

[25] Disponíveis em http://www.ufrgs.br/acervocelpebras/acervo

[26] Para a composição da Comissão Técnica, o INEP faz uma chamada pública, com critérios que levem em conta a formação e a atuação na área de PLE e Celpe-Bras.

sua proficiência é avaliada a partir do seu desempenho em tarefas comunicativas que se assemelham a situações do cotidiano.

O conceito de tarefa comunicativa pressupõe a realização de uma ação mediada pela linguagem por meio de textos (orais e/ou escritos) organizados de forma socialmente construída. Em outras palavras, trata-se de um convite para interagir no mundo, um convite para o uso da linguagem com um propósito social (BRASIL, 2013a, p. 7).

De acordo com o Guia de Capacitação para Examinadores da Parte Oral do Celpe-Bras[27], em referência à concepção teórica do exame, alguns aspectos são relevantes para a avaliação do desempenho do examinando: o propósito da comunicação, o enunciador, os interlocutores e os gêneros do discurso.

> Verifica-se, dessa forma, sua capacidade de produzir textos orais e escritos para agir em sociedade, com propósitos comunicativos precisos e específicos, comunicando-se com os conhecimentos de que dispõe acerca da língua e sobre os rituais sociais que regulam a interlocução. Embora não haja questões explícitas sobre gramática e vocabulário, esses elementos são levados em conta na avaliação do desempenho do examinando, dada sua importância na elaboração de qualquer texto (oral ou escrito) (BRASIL, 2013b, p. 7).

Considerando-se, então, que cabe ao examinando demonstrar essa capacidade de produzir textos para agir em sociedade, levando em conta esses aspectos relevantes mencionados, o conceito de proficiência que subjaz ao exame "consiste no uso adequado da língua para desempenhar ações no mundo" (BRASIL, 2013b, p. 8).

Para que se possa avaliar a proficiência oral do examinando, é necessário que o AI conduza a interação de forma a permitir que haja a adequação da interlocução ao contexto avaliativo, aos temas tratados e ao propósito comunicativo.

Por fim, por cultura, indissociável à língua, entende-se as experiências de mundo e práticas compartilhadas pelos membros de uma comunidade. Os sujeitos agem em contexto e, como tal, são influenciados por sua própria biografia, que é marcada pelo contexto social e histórico no qual estão inseridos (BRASIL, 2013b, p. 7). Ao longo da interlocução, podem surgir ruídos interculturais, daí a importância da formação dos avaliadores, para que não penalizem visões de mundo diferentes da sua.

Pelo fato de o Celpe-Bras ser um exame de natureza comunicativa e, tomando como base a parte oral, é preciso que o AI estimule o diálogo, de tal forma que o examinando possa

---

[27] Disponível em http://www.ufrgs.br/acervocelpebras/arquivos/guias/guia-de-capacitacao-para-examinadores-da-parte-oral. Acesso em: 03 nov. 2014.

demonstrar o seu conhecimento a respeito dos assuntos tratados. De acordo com o Guia citado (Brasil, 2013b), compete a esse avaliador: sustentar a interação, sem julgar as opiniões do examinando; articular as respostas do examinando aos novos tópicos da conversa; levar o examinando a se expressar, explorando o máximo do conhecimento de língua e das práticas de comunicação; atribuir uma nota ao desempenho global do examinando (BRASIL, 2013b, p. 14).

Da mesma maneira, as atribuições do observador também constam do Guia (Brasil, 2013b), que apresenta os critérios da grade de avaliação e determina que as notas devem ser atribuídas ao fim da interação, devido à possível oscilação do desempenho do examinando, e sem que haja qualquer tipo de comunicação entre os avaliadores (AI e AO). Entretanto, os pesos relativos a cada critério não são apresentados no documento, pois, conforme mostrado no Quadro 4, essas informações passaram a ser publicadas para os candidatos a partir da edição 2015/2.

A partir dessas atribuições, é preciso que o avaliadores considerem a concepção teórica do exame, o conceito de tarefa e o de língua-cultura e apliquem, com propriedade, os critérios constantes da grade de avaliação, sob pena de prejudicar o desempenho do examinando. Esses conhecimentos necessários para uma boa atuação dos avaliadores lhe são passados nos eventos de capacitação. Esses sujeitos avaliadores são, portanto, peça fundamental para que a entrevista (parte oral) e a avaliação do examinando reflitam em resultados considerados confiáveis.

Na avaliação da proficiência do examinando, pode haver uma tensão entre as qualidades *validade* e *confiabilidade*, conforme aponta Sakamori (2006). A validade tende a aumentar, no momento em que a interação face a face aproxima-se de uma situação real de comunicação. Mas, por outro lado, a confiabilidade pode ser diminuída porque, nessa interação, podem aparecer outras variáveis, como, por exemplo, algumas citadas pela pesquisadora: os estilos dos entrevistadores, o não cumprimento do tempo destinado a cada etapa da prova (quebra-gelo e elementos provocadores), a forma de elaboração das perguntas (se bem elaboradas, claras, diretas, ou o contrário), a forma de atribuição de notas (se são avaliadores mais ou menos rigorosos).

Levando em conta a importância dos avaliadores no exame, alguns aspectos da sua conduta são objeto da Portaria nº 334[28], de 2 de julho de 2013, que dispõe sobre o credenciamento, recredenciamento e descredenciamento de postos aplicadores e define procedimentos para aplicação do exame Celpe-Bras. Em seu Anexo I, item II, é previsto que:

> Art.5 Os examinadores da Parte Oral devem possuir as habilidades necessárias para conduzir o processo de aplicação das provas, conhecer o construto teórico do Exame, saber planejar e conduzir as interações, manejar os equipamentos utilizados, conhecer a grade de avaliação, compreender bem as delimitações de níveis do Celpe-Bras e agir com cordialidade, lembrando-se de que estão em situação formal de interação.
> Art.6 É imprescindível que os examinadores tenham em mãos um roteiro de orientações durante a realização da avaliação da Parte Oral (BRASIL, 2013c, p. 17).

Ao lado desse documento legal, caminham os manuais, guias e roteiros de interação face a face disponibilizados pelo INEP, com vistas a padronizar os procedimentos de realização do exame, o que pode contribuir para a confiabilidade dos seus resultados.

Diante do que apresentamos até aqui, consideramos a avaliação de proficiência oral uma atividade complexa, em que todas as ações devem estar coerentes com o construto do exame. Nessa atividade, existem as características institucionais, do órgão que o administra, e as dos sujeitos que participam do exame. Além disso, cada uma dessas instâncias carrega consigo suas próprias crenças e valores que podem interferir no processo de avaliação.

Por meio das características institucionais e individuais e também das responsabilidades impetradas a cada um dos protagonistas, constatamos que discutir sobre a confiabilidade a partir do exame Celpe-Bras requer que consideremos que:

- **os administradores do exame** devem ser capazes de elaborar um instrumento que revele a concepção de língua, linguagem e proficiência que o país adota, de acordo com a finalidade do próprio instrumento. Para isso, é preciso que sejam criados (e acompanhados) procedimentos e regras claros a que todos os envolvidos tenham acesso, bem como criadas políticas transparentes de seleção dos especialistas que participarão do processo de elaboração e avaliação do exame. Esses sujeitos devem ser capacitados frequentemente, para que as normas e os procedimentos sejam seguidos de forma padronizada, evitando tratamento diferenciado aos examinandos nos dois contextos de aplicação: Brasil e exterior. Além disso, cabe aos

[28] A referida Portaria foi exarada pelo presidente do INEP à época, o Sr. Luiz Cláudio Costa, e publicada no Diário Oficial da União, em 4/7/13, seção 1, páginas 16 a 17.

administradores a realização de estudos estatísticos contínuos, para se avaliar as qualidades da escala de avaliação.

- **os avaliadores, entrevistador e observador,** carregam consigo suas crenças e valores que refletem sua própria experiência de vida no campo do trabalho e do meio acadêmico. Isso é inerente ao ser humano, mas as suas atitudes no processo de avaliação devem estar alinhadas aos pressupostos teóricos e procedimentais adotados pelo exame.
- **os examinandos** também carregam consigo suas crenças e valores marcados, muitas vezes de maneira explícita, por questões culturais. Além disso, têm de saber lidar com as especificidades do gênero híbrido *entrevista-conversa*, com as questões de simetria e assimetria enunciativas que emergem da interação face a face, demonstrando suas habilidades linguísticas e seu desempenho oral com base no que o exame se propõe a avaliar.

Essas considerações, então, refletem o que configura a avaliação oral no exame Celpe-Bras, em que todos os fatores são interdependentes e interligados entre si. Dada essa gama de fatores relacionados ao exame, bem como a sua importância nos cenários nacional e internacional, diversos pesquisadores têm se dedicado a estudar sobre esse instrumento, como tratado na próxima seção.

2.5 A parte oral do exame Celpe-Bras: diálogo de pesquisas

No cenário da avaliação de proficiência oral, as provas do exame Celpe-Bras já serviram de objeto de discussão em algumas pesquisas, como as apresentadas no Quadro 5, a seguir, o que tem contribuído para a área da Linguística Aplicada em geral e para fortalecer as discussões sobre os temas que emergem do próprio exame.

Quadro 5 - Trabalhos publicados sobre a parte oral do exame Celpe-Bras

(Continua)

Ano de publicação	Nome(s) do(s) autor(es)	Nível/tipo do trabalho	Título do trabalho
2003	SCHOFFEN, Juliana Roquele	Mestrado	Avaliação de proficiência oral em língua estrangeira: descrição dos níveis de candidatos falantes de espanhol no exame Celpe-Bras
2006	SAKAMORI, Lieko	Mestrado	A atuação do entrevistador na interação face a face do exame Celpe-Bras
2008	LIMA, Ronaldo Amorim	Doutorado	Representações do Brasil em textos do exame Celpe-Bras
2009	COURA-SOBRINHO, Jerônimo; DELL'ISOLA, Regina Lúcia Péret	Capítulo de livro	O contrato de comunicação na avaliação de proficiência em língua estrangeira
2009	FORTES, Melissa Santos	Doutorado	Uma compreensão etnometodológica do trabalho de fazer ser membro na fala-em-interação de entrevista de proficiência oral em português como língua adicional
2010	GAYA, Karina Figueiredo	Mestrado	Atividades de compreensão oral como insumo para a produção oral/escrita em Português língua estrangeira: preparação para o Exame Celpe-Bras
2011	FURTOSO, Viviane Aparecida Bagio	Capítulo de livro	Avaliação de proficiência em português para falantes de outras línguas: relação com ensino e aprendizagem
2012	FERREIRA, Laura Márcia Luiza	Mestrado	Habilidades de leitura na proposta de interação face a face do exame Celpe-Bras

(Continua)

2012	SCHOFFEN, Juliana Roquele	Capítulo de livro	Níveis de proficiência oral de examinandos falantes de espanhol no exame Celpe-Bras
2014	BOTTURA, Eleonora Bambozzi	Mestrado	Exame Celpe-Bras: uma investigação sobre o papel do entrevistador na interação face a face
2014	CAIRES, Martha da Rocha	TCC	Percepções quanto à proficiência de PFOL: uma análise comparativa com avaliadores iniciantes e experientes do exame oral do CELPE-BRAS
2014	COSTA, Augusto da Silva	Capítulo de livro	A composição das imagens nos elementos provocadores e a interação na parte oral do Celpe-Bras
2014	COURA-SOBRINHO, Jerônimo	Artigo	*The face to face interaction to evaluate the oral Portuguese language proficiency*
2014	DUARTE, Ana Paula Andrade; OLIVEIRA, Regina Purri Brant Hemetério de; MIRANDA, Yara Carolina Campos de.	Capítulo de livro	Os gêneros textuais na interação face a face do Celpe-Bras
2014	FERREIRA, Laura Márcia Luiza	Capítulo de livro	Avaliação da proficiência oral: atividades de pós-leitura de listas e gráficos no exame Celpe-Bras
2014	NIEDERAUER, Marcia	Artigo	Competência interacional: critério para avaliação da produção oral em língua adicional
2014	PONCIANO, Leila; LONGORDO, Monique	Capítulo de livro	Representações da cultura brasileira nos elementos provocadores do Celpe-Bras de 2013
2014	TROUCHE, Lygia Maria Gonçalves	Artigo	Análise da interlocução em elementos provocadores do exame oral Celpe-Bras
2015	CANDIDO, Marcela Dezotti	Mestrado	Avaliação da interação face a face no exame Celpe-Bras: as características dos elementos provocadores e a atuação dos examinadores-interlocutores

(Conclusão)

2015	COSTA, Augusto da Silva	Mestrado	Avaliação de proficiência oral no Exame Celpe-Bras: análise da condução das interações face a face
2015	SANTOS JUNIOR, Elysio Soares	Artigo	Descendo do salto: uma análise sobre mal-entendidos na interação face a face do Celpe-Bras
2016	VIEIRA, Ana Luíza Gabatteli	Mestrado	Curso online para a parte oral do Celpe-Bras: contribuições da avaliação de proficiência para o ensino-aprendizagem de PLE
2017	CAMPOLINA, Isabela Bertho	TCC	Competência intercultural na prova oral do exame Celpe-Bras: um estudo comparativo
2017	NEVES, Liliane de Oliveira; AGOSSA, Mahulikplimi Obed Brice; COURA-SOBRINHO, Jerônimo	Capítulo de livro	Enquadramento temático na Parte Oral do Exame que Confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras)
2018	FERREIRA, Laura Márcia Luiza	Doutorado	Avaliação da proficiência oral: uma análise fatorial e de discriminação de itens do exame Celpe-Bras

Fonte: elaborado pela autora, 2018.
Nota: as buscas foram realizadas em março de 2017, nas seguintes bases: Acervo Celpe-Bras (http://www.ufrgs.br/acervocelpebras/acervo), Biblioteca Digital Brasileira de Teses e Dissertações, Google Acadêmico e em livros.

Os trabalhos somam 25 e esta pesquisa estabelece diálogo com vários deles, sendo que seis são destacados a seguir, por fazerem considerações a respeito do papel dos avaliadores.

Sakamori (2006) objetivou analisar a atuação dos entrevistadores nas interações face a face, com enfoque em dois elementos distintos: um relacionado com os procedimentos do exame (se os entrevistadores cumpriram o tempo determinado para a prova, ou seja, cinco minutos para o "quebra-gelo" e para cada elemento provocador, totalizando vinte minutos para o total da prova; se os entrevistadores respeitaram a ordem das etapas do Manual do Aplicador) e outro relacionado com os estilos (colaborativos e não colaborativos) que os entrevistadores apresentaram nas interações. Para isso, foram analisadas 58 entrevistas, gravadas em vídeo, de exames realizados em 2004, em três diferentes universidades, localizadas em um mesmo estado brasileiro.

Algumas constatações merecem destaque e estão descritas a seguir. Quanto aos procedimentos do exame: alguns entrevistadores não cumpriram o tempo e as etapas; alguns entrevistadores selecionaram os elementos provocadores no momento da interação e não antes da entrada do examinando na sala, desviando a atenção e o olhar ao procurar o material, o que fez com que as perguntas ficassem desfocadas; alguns observadores manifestaram-se durante a interação face a face, o que contraria os procedimentos do exame. Quanto ao estilo dos entrevistadores, alguns foram caracterizados como colaborativos (quando, por exemplo, faz comentários, deixando a interação menos assimétrica, faz perguntas claras e diretas) e não colaborativos (quando, por exemplo, age apenas como perguntador, sem envolvimento com o candidato e faz perguntas longas e confusas). Segundo a pesquisadora, essas constatações dizem respeito a fatos que podem interferir no desempenho do examinando e, consequentemente, trazer implicações para a confiabilidade dos resultados do exame.

Furtoso (2011b), por sua vez, identificou, a partir de algumas interações face a face, estratégias das quais os entrevistadores lançam mão na parte oral do exame. Para tratar de coerência entre ensino e avaliação, a pesquisadora examinou algumas interações, procurando analisar como a relação entre o conteúdo dos instrumentos de avaliação, o avaliador e o examinando pode ir ao encontro do construto que fundamenta o Exame (FURTOSO, 2011b, p. 233). Foram analisadas seis interações, a partir de dois elementos provocadores, da aplicação de abril de 2010.

Segundo a pesquisadora, do ponto de vista dos instrumentos de avaliação da parte oral do exame, seus conteúdos contemplam língua e cultura sendo indissociáveis. Quanto à figura dos avaliadores, ressalta:

> a intervenção do avaliador, conhecedor das concepções que fundamentam o Exame, é de suma importância para assegurar que todo o processo, desde o (re)conhecimento do perfil do examinando até a avaliação de sua proficiência, seja realmente um espaço de negociação de significados e de busca pela cooperação entre línguas-culturas diferentes, contemplando, assim, quem fala, de onde, com quem e para quem. **O entrevistador, nesse sentido, é o que assume a voz das concepções que fundamentam o Exame, enxergando e ouvindo o outro para que a interação faça sentido e os conflitos culturais sejam minimizados**. (FURTOSO, 2011b, p. 234 – grifo nosso).

Já Bottura (2014), levando em consideração as pesquisas de Sakamori (2006) e Furtoso (2011b), objetivou investigar o papel dos entrevistadores nas interações face a face de uma aplicação do exame (segundo semestre de 2012) em um posto aplicador do Estado de São Paulo, dando enfoque na atuação, nas reações e nas atitudes dos entrevistadores durante a interação. Das várias considerações que a pesquisadora aponta, é pertinente ressaltar: a necessidade de pesquisas com enfoque na formação de examinadores, dada a complexidade e influência configuradas na interação entre entrevistador e examinando; é preciso que os entrevistadores tomem pra si concepções elementares que fundamentam o exame, tais como língua, cultura e proficiência; é preciso que os entrevistadores tenham conhecimento não só das etapas do exame, mas o objetivo de cada uma delas, para que se tornem sujeitos críticos e reflexivos que consigam mobilizar estratégias e elementos em seu discurso indispensavelmente coerentes com o construto do exame.

Coura-Sobrinho (2014), por sua vez, apresenta alguns resultados da sua pesquisa de pós-doutorado, discutindo a maneira como as habilidades orais dos examinandos são elicitadas e avaliadas, considerando-se os dois contextos de aplicação do Celpe-Bras: Brasil e exterior. Para o desenvolvimento do trabalho, o pesquisador analisou uma amostra de 15 interações face a face de uma mesma edição, levando-se em conta os movimentos de simetria e assimetria enunciativa, assumindo o fato de que há variabilidade de comportamento dos entrevistadores.

Dos resultados da pesquisa, destacam-se: (i) há predominância dos momentos de assimetria competitiva provocados pelo entrevistador, em relação à assimetria competitiva provocada pelo candidato. Esse fato ocorre com mais frequência nas entrevistas realizadas no exterior e isso mostra que o candidato ao Celpe-Bras, ao se submeter ao exame estando fora do Brasil, coloca-se numa posição enunciativa submissa ao entrevistador (COURA-SOBRINHO, 2014, p. 8); (ii) são raros os momentos de simetria nos dois contextos de aplicação, sobretudo por parte dos examinandos. Esses movimentos de simetria e assimetria

enunciativa, observados pelo pesquisador, foram capazes de revelar nuances dos contextos de aplicação: Brasil e exterior.

Também preocupado com os contextos de aplicação do exame Celpe-Bras, Costa (2015) objetivou analisar as formas como os entrevistadores conduzem as interações face a face e como as estratégias por eles utilizadas podem refletir no resultado dos examinandos. Para isso, foi analisada uma amostra de 27 interações de três diferentes postos no exterior, sendo que todas elas apresentaram notas discrepantes entre as partes escrita e oral. Nas análises, o pesquisador considerou: a realização ou não da pergunta obrigatória do roteiro do entrevistador, a utilização das perguntas do roteiro, a quantidade de perguntas extras formuladas pelo entrevistador, o tempo dedicado a cada etapa da interação, entre outros. Um dos resultado a que chegou o pesquisador é que

> [...] **a atuação dos entrevistadores parece ser, então, uma das razões para a geração das discrepâncias encontradas em nossos dados**. As perguntas com foco no candidato, com estrutura simplificada e as que exigem do candidato a produção de respostas objetivas, além do pouco tempo de fala dada ao candidato, podem supervalorizar a produção oral dos examinandos que, embora não demonstrem altos níveis de proficiência ao longo das interações, obtêm notas altas em sua avaliação. Corrobora para essa constatação as notas mais baixas que esses mesmos candidatos obtêm na Parte Escrita do exame que exige maior esforço linguístico do candidato para cumprir as tarefas (COSTA, 2015, p. 185 – grifo nosso).

Além disso, Costa (2015) afirma que o não cumprimento de algumas das etapas da interação, conforme orientações disponibilizadas aos entrevistadores no roteiro da interação face a face, pode comprometer a confiabilidade dos resultados do exame, uma vez que conduções muito diferentes revelam falta de isonomia na avaliação dos examinandos.

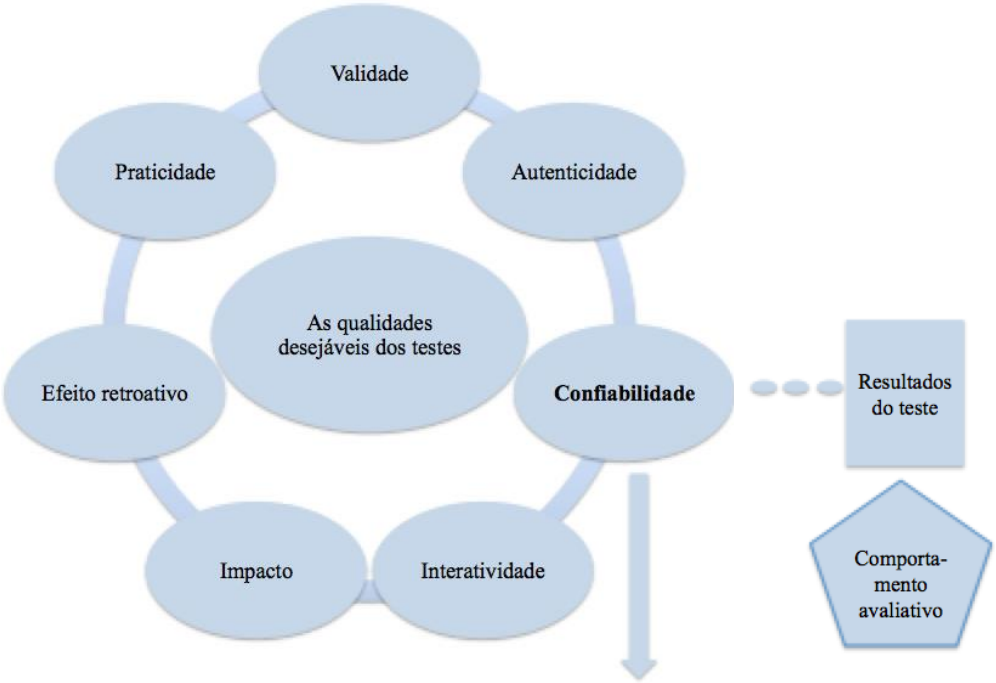
Por fim, Ferreira (2018) desenvolve uma pesquisa quantitativa sobre a estrutura fatorial da escala de avaliação da prova oral e do quanto os critérios avaliados fornecem informações relevantes sobre a proficiência dos examinandos. Entre os resultados da pesquisa, destacam-se: (i) a escala de avaliação apresenta-se unidimensional, ou seja, é avaliado um único construto: a proficiência oral; (i) do critérios de avaliação, *compreensão* é o que possui menor poder de explicação da nota final do desempenho do examinando; (iii) baseando-se nos valores de carga fatorial, a autora discute a possibilidade de revisão dos atuais pesos de cada um dos critérios da grade. Ressaltamos que, do levantamento apresentado no Quadro 5, essa é a única pesquisa sobre o Celpe-Bras que adota uma metodologia quantitativa de análise de dados, o que mostra que são escassos trabalhos dessa natureza no âmbito do exame.

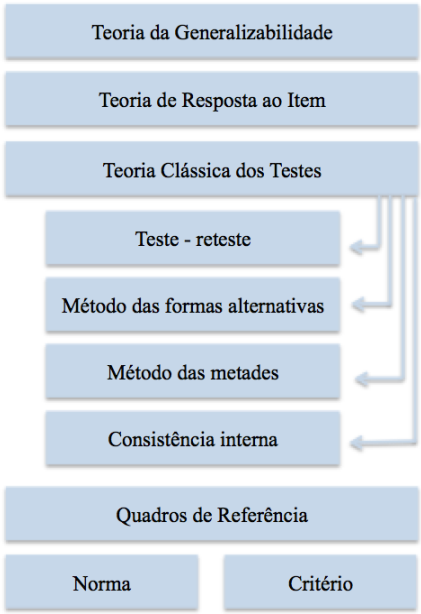
Nota-se que as seis pesquisas levam em conta o papel do avaliador na parte oral do Celpe-Bras e, consequentemente, o construto do exame. A proposta desta tese vai ao encontro dessas perspectivas, ao estabelecer como norte a seguinte pergunta: *há variabilidade no comportamento de avaliação das interações face a face?* Avaliação, aqui, é entendida pelas três instâncias que compõem o processo, conforme mostrado no início deste capítulo.

Essa proposta dialoga com as pesquisas ora citadas e a sua diferença está configurada em dois eixos: (i) a aplicação de uma metodologia quantitativa, de forma a identificar o comportamento avaliativo dos atribuidores de nota e mostrar qual a relação existente entre esse comportamento e a confiabilidade dos resultados do teste e (ii) um estudo de sete edições consecutivas do exame. Do levantamento apresentado no Quadro 5, nota-se que não há pesquisas sobre estimativas da confiabilidade dos resultados do Celpe-Bras, tanto em relação à prova escrita quanto à oral. Esta tese, portanto, cumpre o papel de preencher lacunas nessa área.

Neste capítulo, apresentamos as particularidades da parte oral do Celpe-Bras, enfatizando o seu processo de aplicação, de avaliação e os sujeitos que dele participam, além de elencarmos algumas pesquisas que dialogam entre si pelo próprio objeto de análise: as interações face a face do exame. O próximo capítulo é dedicado ao referencial teórico da pesquisa.

CAPÍTULO 3
REFERENCIAL TEÓRICO





3 REFERENCIAL TEÓRICO

Este capítulo é dedicado ao referencial teórico da pesquisa. Nele discutimos as qualidades desejáveis dos testes de língua, apresentamos a definição de *comportamento avaliativo dos atribuidores de notas* e damos ênfase à confiabilidade, apontando como esta pode ser estimada.

3.1 As qualidades dos testes de língua

Qualquer teste carrega consigo o seu próprio construto e é a partir disso que deve ser estruturado, administrado, aplicado, avaliado e utilizado. Bachman e Palmer (1996, p. 19-36) tratam das seis qualidades desejáveis quando se trata de testes de língua: confiabilidade, validade de construto, autenticidade, interatividade, impacto e praticidade, comentadas a seguir.

A *confiabilidade* normalmente é definida pela consistência da avaliação e está relacionada às variáveis que possam interferir no que se quer avaliar, sendo uma qualidade essencial dos resultados dos testes.

A *validade de construto* diz respeito ao significado e adequação das interpretações que fazemos com base nos resultados de testes, sempre levando em consideração o que se quer avaliar, ou seja, considera-se sempre o construto do teste.

A *autenticidade*, por sua vez, refere-se ao grau de correspondência das características de determinada tarefa do teste para as características de tarefas da língua em uso. Nesse sentido, as tarefas devem refletir a língua em uso e não serem criadas/inventadas para uso na avaliação.

Já a *interatividade* é entendida como a extensão e o tipo de envolvimento do examinando ao realizar uma tarefa, ou seja, como as suas estratégias metacognitivas (seleção de informações, estabelecimento de relações, por exemplo), conhecimento de língua, esquemas afetivos estão envolvidos com a tarefa.

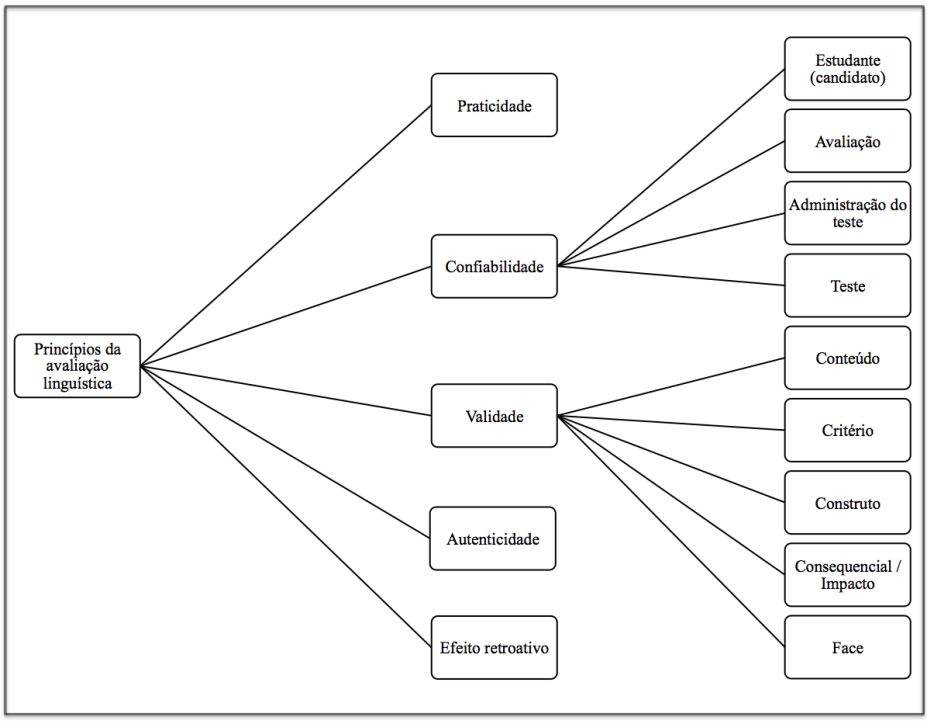
Outra qualidade é o *impacto*, que pode ser de dois níveis: o micro, em termos do efeito que o teste provoca no examinando, e o macro, em termos do efeito provocado na sociedade e nos sistemas educacionais. O *impacto* também contempla outra característica, o *efeito retroativo*, que tem sido de grande interesse de pesquisadores da área da avaliação, a exemplo de Scaramucci (2000/2001; 2004; 2008), Silva (2006) e Agossa (2017). Bachman e Palmer (1996) apresentam uma distinção conceitual dos dois termos:

> [...] em resumo, **o impacto** do uso de teste precisa ser considerado dentro dos valores e objetivos da sociedade e do programa educacional em que ele ocorre e de acordo com as potenciais consequências de tal uso. Ao avaliar o impacto do uso de teste, devemos considerar as características específicas do teste (finalidade, domínio da língua estrangeira, os examinandos, definição de construto) em termos de valor e objetivos dos indivíduos e do sistema educacional e da sociedade, e das potenciais consequências para os indivíduos, o sistema educacional e sociedade. **A noção de efeito retroativo** em testes de língua pode ser caracterizada em termos de impacto e inclui o impacto potencial sobre os examinandos e suas características, em atividades de ensino e aprendizagem, e sobre os sistemas educacionais e da sociedade (BACHMAN; PALMER, 1996, p. 35 – grifo nosso).

Por fim, enquanto as qualidades citadas referem-se aos usos que se faz dos resultados do teste, a *praticidade* pertence à forma como o teste é implementado em dadas situações e se será desenvolvido e utilizado em todas elas. Entendemos que fazem parte dessa qualidade: a proporção entre quantidade de examinandos e examinadores, os recursos financeiros, o tempo disponível para a aplicação do teste, facilidade de administração e a sua estrutura tecnológica (equipamentos, por exemplo).

De um lado, Bachman e Palmer (1996) abordam essas características como as *qualidades* de testes e, de outro, Brown e Abeywickrama (2010) tratam-nas como *princípios* que norteiam a avaliação linguística, conforme Figura 3, a seguir.

Figura 3 - Princípios que norteiam a avaliação linguística



Fonte: elaborado pela autora, 2018.
Nota: esquema elaborado com base em Brown e Abeywickrama, 2010, p. 25-39.

Como se observa nessa figura, validade e confiabilidade são princípios que se desdobram em outros, por serem, acreditamos, os mais complexos e, por isso, exercem influência nas tomadas de decisão por parte das instâncias que administram os testes e dos demais *stakeholders*. Não é à toa que têm ocupado espaços significativos em livros da área do ensino e avaliação de línguas (Bachman, 1990; Bachman e Palmer, 1996; McNamara, 2000; Hughes, 2003; Bachman, 2004; Brown, 2005; Fulcher, 2010; Brown e Abeywickrama, 2010) e despertado o interesse de pesquisadores (a exemplo dos brasileiros: Scaramucci[29], 2004, 2009, 2011; Sakamori, 2006; Fortes, 2009; Schoffen, 2009; Bottura, 2014; Caires, 2014; Coura-Sobrinho, 2014; Costa, 2015; Cândido, 2015, entre outros). Apresentamos, então, algumas considerações sobre validade e discorremos, com mais detalhes, sobre a confiabilidade.

Ao tratar de testes de língua, Scaramucci (2009; 2011) clarifica o conceito tradicional e moderno de validade. Segundo a pesquisadora, a validade, numa visão tradicional, tem sido definida como uma característica ou qualidade de um teste, um critério para sua aceitabilidade.

> Validade em geral refere-se à adequação de um teste ou de algum de seus componentes como uma medida do que esse teste deve medir. Um teste é válido na medida em que mede o que deve medir. Assim, o termo válido, quando usado para descrever um teste, deve geralmente vir acompanhado pela preposição "para". Qualquer teste, dessa forma, pode ser válido para alguns propósitos, mas não para outros (HENNING, 1987 apud SCARAMUCCI, 2011, p. 105).

Ainda segundo Scaramucci (2009; 2011), a validade normalmente tem sido abordada em relação à confiabilidade e um teste não pode ser considerado válido sem primeiro ter resultados confiáveis, ou seja, é necessário que avalie o que se pretende com precisão e de maneira consistente. O exemplo dado pela autora para ilustrar esses dois conceitos é relativo à correção de uma prova por dois avaliadores distintos: se as notas apresentam discrepância significativa (10 e 0, numa escala de 0 a 10), por exemplo, qual resultado reflete o real desempenho do candidato? Essa questão recai sobre o conceito de confiabilidade. Por outro lado, um teste com resultados confiáveis pode não ser considerado válido, quando, por

[29] Ressaltamos que, no Brasil, Scaramucci é uma das pesquisadoras com trajetória considerável na área da avaliação de línguas e de validação. Além dos textos publicados na área, um trabalho de relevância foi a sua participação no processo de validação do Exame de Proficiência em Língua Inglesa do Sistema de Controle do Espaço Aéreo Brasileiro (EPLIS), exame sobre o qual tecemos alguns comentários no Capítulo 1.

exemplo, um teste de produção escrita em língua estrangeira solicita aos candidatos apenas a tradução de termo a termo, sendo que aprender a escrever em língua estrangeira ultrapassa os limites da tradução.

> Um teste de desempenho de produção escrita que, por outro lado, tem por objetivo a redação de um texto pode ser válido, embora não necessariamente confiável, se não forem estabelecidos critérios claros para a correção, se os corretores não forem treinados para a tarefa, e assim por diante. Dessa forma, um aumento de validade geralmente leva a uma diminuição de confiabilidade e vice-versa, revelando a tensão existente entre os dois parâmetros (SCARAMUCCI, 2011, p. 105).

Nesse sentido, nota-se que o parâmetro da confiabilidade está relacionado também à elaboração de critérios claros de correção e à capacitação dos avaliadores. Ao parâmetro da validade, outros desdobramentos também são apontados pela pesquisadora, a partir do que trata Hughes (1989/1994 *apud* Scaramucci, 2011), são eles: validade de construto, de conteúdo, relacionada a critério e de face. Estabelecendo um diálogo com esses termos, Damazo (2012) tratou de algumas definições dadas ao termo validade.

> Um instrumento de avaliação precisa ser confiável e válido. Para ser válido, o instrumento precisa ter qualidades relativas a quatro dimensões: validade de critério, validade de conteúdo, validade de face e validade de construto. A validade de critério diz respeito à questão *avaliar para quê?* A validade de conteúdo refere-se à questão *avaliar o quê?* A validade de face diz respeito ao significado que os avaliados atribuem ao resultado da avaliação. A validade de construto tem estreita relação com os pressupostos teóricos subjacentes à elaboração do instrumento e diz respeito às concepções que os elaboradores de itens possuem sobre o que seja avaliar (DAMAZO 2012, p. 30).

Scaramucci ainda ressalta que o conceito tradicional de validade era considerado fragmentado e incompleto, por restringir-se a uma visão psicométrica e, então, não considerava as implicações de valor do significado dos resultados ou escores, nem as consequências sociais do uso desses resultados, ou seja, a dimensão social e política que devia estar presente na avaliação de línguas pelo fato de ser uma prática social (SCARAMUCCI, 2011, p.110).

A partir de uma visão moderna e atravessada pelos pressupostos que subjazem à *avaliação de desempenho e efeito retroativo e impactos/consequências sociais*, a pesquisadora afirma que o novo conceito, embora contemplando múltiplas facetas, unifica-se em torno da validade de construto, passando a incluir as consequências intencionais e não intencionais, assim como as consequências reais das avaliações no ensino, na aprendizagem, na vida das pessoas em geral como parte das evidências de validade no processo de validação (SCARAMUCCI, 2011, p. 116).

Essa nova visão de validade, portanto, leva-nos a afirmar que o construto é o elemento definidor de quaisquer decisões e análises que se possam fazer de um teste. Como bem ressalta Scaramucci (2009) ao questionar os resultados de uma pesquisa que investigou instrumentos de avaliação de leitura em inglês como língua estrangeira, cada teste tem seu próprio construto, definido na fase de elaboração do instrumento. Consequentemente, ao analisar a validade [ou qualquer outra qualidade] de um instrumento, é preciso ter o seu construto como ponto de partida.

Validade e confiabilidade são, portanto, características e qualidades essenciais para o uso de testes. Enquanto a confiabilidade está relacionada à qualidade de escores de um teste, a validade relaciona-se com a qualidade de interpretações ou usos que são feitos desses escores. Determinar, então, o grau necessário dessas qualidades para um contexto específico envolve julgamento de valor por parte de que quem utilizará o teste (BACHMAN, 1990, p. 26).

3.2 Confiabilidade: característica do teste ou dos seus resultados?

Quando se fala em *confiabilidade,* uma das preocupações de alguns autores é esclarecer certa "confusão conceitual" em torno do termo. Para Thompson (2003a, p. 3), é comum encontrar na literatura autores fazendo afirmações do tipo "a confiabilidade do teste" ou "o teste é confiável". Para o autor (2003a), isso se trata de uma inverdade, pois a confiabilidade é uma propriedade que se aplica aos escores do teste e não ao teste em si. Na prática, a confiabilidade de escores é uma questão de grau, porque todos os escores contêm algumas flutuações aleatórias e não são perfeitamente confiáveis. Não há escores perfeitos (THOMPSON, 2003a, p. 5). Essa ideia de que a confiabilidade está relacionada aos escores e não ao teste é reforçada por Thompson (2003b), Sawilowsky (2003) e Bachman (2004).

Urbina (2007) também chama a atenção para esse fato e afirma que, embora a fidedignidade[30] na testagem dependa, em um grau significativo, das características do teste, a fidedignidade dos escores – que é o que resulta do uso do instrumento e é o que realmente importa – também pode ser afetada por muitas outras variáveis (URBINA, 2007, p. 124).

---

[30] Alguns autores, como Marôco e Garcia-Marques (2006) e Urbina (2007), utilizam o termo *fidedignidade*, inclusive para a tradução de *reliability*. Nesta tese, optamos por *confiabilidade*, por ser mais utilizado no Brasil, na área da Linguística Aplicada.

Portanto, analisar a confiabilidade significa analisar os escores de determinado teste e isso requer o uso de técnicas estatísticas apropriadas a cada tipo de teste.

3.3 Algumas variáveis intervenientes na confiabilidade

Para Bachman (1990), tratar sobre o parâmetro da confiabilidade significa observar a pergunta: o quanto do desempenho individual em um teste está relacionado ao erro de mensuração ou a outros fatores, além da habilidade linguística que se quer medir? (BACHMAN, 1990, p. 160). A partir desse questionamento, podemos afirmar que o resultado final do desempenho de um candidato ao Celpe-Bras, demonstrado por sua habilidade linguística, pode conter erros de mensuração e sofrer influência de outros fatores que, ao nosso ver, estão relacionados a variáveis intervenientes na totalidade do processo de avaliação. Para esse autor, quanto menos esses fatores afetarem as pontuações de um teste, maior será o efeito relativo das habilidades linguísticas que se quer medir, e consequentemente, a confiabilidade dos seus resultados (BACHMAN, 1990, p. 160).

Para Urbina (2007), essas variáveis são relativas a fatores atinentes ao sujeito que se submete ao teste (fadiga, falta de motivação etc.) ou a condições da própria testagem (ruídos no local de aplicação do teste, personalidade do examinador etc.), ao que acrescentamos o comportamento avaliativo dos atribuidores de nota.

Em se tratando de testes que utilizam a entrevista como o instrumento para medir o desempenho oral de candidatos, como é o caso do Celpe-Bras, uma dessas variáveis pode ser o comportamento do entrevistador, isto é, a maneira que é feita a condução da interação. Ao avaliar o impacto da variabilidade do entrevistador na avaliação de candidatos ao *International English Language Testing System* (*IELTS*)[31], Brown (2005) aponta três aspectos do comportamento do entrevistador que podem ser uma ameaça à confiabilidade, quais sejam: falta de equivalência dos diferentes conjuntos das perguntas elaboradas, inconsistência de perguntas e variação na seleção de conteúdo. A partir dos resultados de sua pesquisa, a pesquisadora ressalta que é preciso realizar estudos que não somente examinem o impacto da

[31] O *International English Language Testing System* possui 10 escalas de avaliação da proficiência linguística, quais sejam: 0- *Did not attempt the test*; 1- *Non user*; 2- *Intermittent user*; 3- *Extremely limited user*; 4- *Limited user*; 5- *Modest user*; 6- *Competent user*; 7- *Good user*; 8- *Very good user* e 9- *Expert user*.

variabilidade do entrevistador em avaliações atribuídas aos candidatos, mas que também procurem estabelecer as causas de cada variação de escore. Isso reforça o que Lumley e McNamara (1995) já haviam apontado, de que é extensa a variabilidade nos resultados dos exames associada ao avaliador.

Meiron e Schick (2000) realizaram um estudo exploratório no *Institute for Egyptian Teachers of English as a Foreign Language*, cujos sujeitos da pesquisa foram 25 professores egípcios de Inglês como Língua Estrangeira (ILE) que participaram de um programa de treinamento de professores de 11 semanas. Os pesquisadores trabalharam com dados qualitativos e quantitativos de um teste de proficiência oral, no intuito de investigar se os escores quantitativos representavam desempenhos qualitativamente diferentes no teste e como uma análise qualitativa do desempenho pode explicar qualquer melhoria quantitativa nos escores das habilidades orais por parte dos examinandos.

Van Lier (1989 apud Meiron e Schick, 2000) aponta que um dos problemas mais comuns em testes de proficiência oral é a dificuldade que a avaliação de características conversacionais apresenta aos avaliadores e é justamente sobre esse aspecto que Meiron e Schick voltam sua atenção na pesquisa. Os autores focam, então, na avaliação e nos avaliadores ao invés de ser na interação entre entrevistador e candidato, e se interessam em analisar algumas variáveis que podem influenciar a maneira como os sujeitos avaliam características conversacionais como parte integrante da proficiência oral.

Segundo os pesquisadores, os examinandos podem apresentar desempenhos qualitativamente diferentes, ainda que obtenham escores quantitativos similares.

> Esta é uma preocupação importante, porque 'o avaliador é um componente integral dos escores de avaliação de proficiência' (Chaloub-Deville, 1995, p. 255) **e a variação na avaliação pode colocar em xeque a confiabilidade e validade do resultado**. No entanto, se a entrevista for, como diz van Lier (1989), um 'veículo apropriado para a demonstração da habilidade de fala em contexto' (pág. 489), então precisamos examinar a maneira como os avaliadores são treinados para reconhecer características conversacionais e a forma como eles interpretam e avaliam a proficiência oral em testes de desempenho (MEIRON; SCHICK, 2000, p. 155 – grifo nosso).

Meiron e Schick (2000) ainda afirmam que um dos fatores que podem influenciar a avaliação é o perfil do avaliador: se professor ou não, o que é, portanto, nos termos de Bachman (1990) e Urbina (2007), uma das variáveis que podem interferir na condução do teste e na confiabilidade dos seus resultados.

Essa variação na avaliação, de que tratam Meiron e Schick (2000), pode corresponder a uma preocupação de Bachman (1990) ao apresentar o seguinte questionamento: *o que é*

mais problemático: certificar um candidato que não é proficiente, ou o contrário? Ou as duas situações? Portanto, entendemos que a variável *avaliador* pode ser uma peça-chave para a garantia da confiabilidade dos resultados de um teste.

He e Young (1998), ao citarem as qualidades de testes tratadas por Bachman e Palmer (1996), afirmam que a confiabilidade é a consistência com que um teste mede a habilidade do candidato, e uma maneira em que as entrevistas podem não ter resultados confiáveis é quando dois avaliadores diferentes julgam, de forma diferente, a habilidade oral de um mesmo sujeito. Para os autores, esse fato é considerado uma ameaça à confiabilidade, sendo que uma forma de ser evitada é por meio da capacitação de avaliadores e do desenvolvimento de escalas de avaliação que minimizem a discordância entre eles, como também apontado por Scaramucci (2011).

Outra variável que pode interferir nos resultados do teste e, consequentemente, na confiabilidade, é a maneira como grades de avaliação são interpretadas e utilizadas para a mensuração do desempenho oral dos candidatos. Amaral (2015) desenvolveu uma pesquisa sobre o exame IELTS, com o objetivo de analisar como examinadores que atuam no Brasil utilizam a grade para fazerem a avaliação. A grade desse exame é composta por quatro critérios, a saber: *Fluência e Coerência*, *Recursos Lexicais*, *Abrangência e Correção Gramaticais* e *Pronúncia*. Para a realização da pesquisa, foram entrevistados cinco avaliadores que atribuíam e justificavam suas notas ao desempenho de um candidato[32]. Entre os resultados que a pesquisa aponta, destacam-se: (i) os avaliadores tiveram mais dificuldades em avaliar os critérios *Fluência e Coerência* e *Pronúncia*, mas o contrário ocorreu em relação a *Abrangência e Correção Gramaticais* e *Recursos Lexicais*; (ii) os examinadores afirmaram ter dificuldades em manter, ao mesmo tempo, os papéis de interlocutor e avaliador, pois devem seguir o roteiro, interagir com o candidato de forma natural, avaliar os critérios constantes da grade e controlar os procedimentos da entrevista. Ou seja, tanto os critérios de avaliação quanto a atuação do entrevistador são vieses importantes de análise.

Também interessados em particularidades atinentes às variáveis que podem ameaçar a confiabilidade dos resultados de instrumentos de avaliação, os três pesquisadores brasileiros citados a seguir dedicaram-se a analisar o papel do entrevistador na interação face a face do

[32] Amaral (2015) utilizou uma simulação de prova oral, nos moldes do IELTS, para que pudesse realizar a pesquisa. A prova foi gravada em vídeo, sendo que os participantes foram: uma entrevistadora com experiência na aplicação do exame e um aluno que se preparava para prestar o exame na mesma semana em que ocorreu a simulação, ambos brasileiros. No IELTS, assim como no Celpe-Bras, o entrevistador também é um avaliador. A diferença é que, no Celpe-Bras, há também a figura do observador.

exame Celpe-Bras, conforme detalhes apresentados no Capítulo 2. Com o objetivo de investigar a atuação dos entrevistadores, Sakamori (2006) tomou como base dois aspectos: os procedimentos de aplicação do exame e os estilos que os entrevistadores apresentaram nas interações. Bottura (2014), por sua vez, discorreu sobre a atuação, as reações e as atitudes dos entrevistadores durante a interação face a face. Costa (2015), por fim, analisou as formas como os entrevistadores conduzem as interações face a face e como as estratégias por eles utilizadas podem refletir no resultado dos examinandos.

Como se pode notar, o papel do entrevistador exerce importância fundamental no processo de aplicação do exame Celpe-Bras e também na atribuição de notas, já que o entrevistador, além de conduzir a interação, avalia o candidato. A capacitação de AI e AO merece atenção, tendo em vista que a conduta desses avaliadores pode estar, nos termos de Brown (2005), vinculada a aspectos considerados uma ameaça à confiabilidade.

Portanto, devido a uma gama de variáveis que interferem no processo de mensuração de um teste, desde a elaboração do instrumento até a avaliação do desempenho dos candidatos, a garantia da confiabilidade dos resultados torna-se um desafio para os seus desenvolvedores. Nos testes em que é adotada a interação face a face para avaliação de desempenho linguístico, esse desafio é marcado pela sua própria natureza subjetiva.

Conforme aponta McNamara (1996 apud Schoffen, 2009, p. 29), os julgamentos que envolvem subjetividade são complexos, estão condicionados a interpretações por parte dos sujeitos avaliadores e, consequentemente, sujeitos à discordância. O autor propõe três procedimentos para tentar reduzir essa subjetividade a níveis aceitáveis. São eles: (i) o uso de descritores de desempenho cuidadosamente formulados para cada nível de avaliação, incluindo exemplos ilustrativos das características do desempenho; (ii) treinamento cuidadoso dos avaliadores no uso dos procedimentos de avaliação e (iii) avaliação de cada candidato mais de uma vez e adoção de procedimentos para lidar com as possíveis discrepâncias de notas. Esses procedimentos, portanto, podem auxiliar na diminuição da subjetividade, para que esta não seja uma forte fonte de erro de mensuração.

3.4 Uma definição de *comportamento avaliativo dos atribuidores de notas*

Como mostrado na Figura 3, Brown e Abeywickrama (2010) indicam quatro fatores que podem interferir na confiabilidade dos resultados de um teste e um deles está vinculado diretamente aos avaliadores (*rater reliability*). Segundo os autores (2010), Bachman (1990) e Moskal e Leydens (2000), podem fazer parte do processo de avaliação o erro humano, a

subjetividade e entendimentos enviesados. Daí a preocupação com a confiabilidade entre os avaliadores (*inter-rater reliability*) e com o próprio avaliador em si (*intra-rater reliability*), ou seja, a preocupação com a consistência da avaliação (*rater consistency*).

O primeiro fator, *inter-rater reliability*, refere-se à preocupação de que os escores atribuídos a um determinado sujeito podem variar de avaliador para avaliador. Segundo Moskal e Leydens (2000), para minimizar esse tipo de ocorrência e, consequentemente, diminuir a quantidade de discrepâncias, devem ser criadas grades de avaliação com critérios bem definidos, embora elas não sejam capazes de eliminar completamente as variações de escores atribuídos pelos avaliadores.

O segundo fator, por sua vez, *intra-rater reliability*, está relacionado à maneira pela qual um avaliador faz a avaliação de determinado sujeito. Por exemplo, o avaliador pode ter interferência do seu próprio cansaço, do humor, da comparação que estabelece entre os candidatos, e atribuir notas diferentes para sujeitos com igual desempenho. Ou seja, trata-se da existência de inconsistências de influências internas ao avaliador e, segundo os autores (2000), uma maneira de minimizar esse tipo de ocorrência pode ser a consulta à grade de avaliação, ao longo do processo, para que a consistência na avaliação seja mantida.

Tratar da consistência da avaliação leva-nos a abordar algo inerente a qualquer teste subjetivo: o *comportamento avaliativo*. Embora na Psicologia e na Sociologia, por exemplo, a palavra *comportamento* diga respeito a particularidades dos sujeitos avaliadores, como sexo, idade, formação acadêmica, tempo de experiência na avaliação, contexto de atuação (se no Brasil ou no exterior) etc., utilizamos esse termo de forma genérica, tendo em vista que, no banco de dados analisado, não temos informações de quem são os avaliadores da prova oral do exame Celpe-Bras, mas apenas qual posição ocupam: se entrevistadores ou observadores.

Portanto, *comportamento avaliativo*, nesta pesquisa, refere-se à maneira como os avaliadores atribuem notas ao desempenho oral dos examinandos, nas diferentes instâncias de avaliação, ou seja, se os avaliadores concordam ou discordam entre si (observador x entrevistador, em determinada instância) ou entre suas funções (observador x observador; entrevistador x entrevistador, nas diferentes instâncias de avaliação). Trata-se, então, da maneira como as notas (ou os conceitos que elas carregam) são atribuídas aos examinandos, pelos diferentes avaliadores e nas diferentes instâncias de avaliação, o que se materializa na observação da consistência *inter-rater*.

Analisar o comportamento avaliativo permite que façamos inferências sobre o entendimento que os avaliadores têm sobre a avaliação e os critérios que servem de base para

a mensuração do desempenho dos examinandos. E tudo isso poderia ser enriquecido se o *corpus* de análise permitisse a identificação das particularidades dos sujeitos avaliadores citadas anteriormente, bem como do contexto de aplicação.

Entendemos que uma significativa variabilidade desse comportamento possa provocar, dentre outros fatores, erro de mensuração e, consequentemente, interferir nos resultados das habilidades que o teste se propõe a avaliar. Portanto, o comportamento avaliativo dos atribuidores de nota deve ser pautado pelo construto do exame, por meio de uma grade de avaliação com critérios claros, e ser *refinado* por meio de guias, manuais e capacitações práticas promovidas pelos responsáveis pelo exame, para que as características individuais dos avaliadores (traços psicossociais e posicionamentos teóricos) não sejam responsáveis pelo incremento da subjetividade do processo avaliativo.

Ressaltamos que a conceituação dada para o termo *comportamento avaliativo* foi elicitada pelo próprio objeto em análise, ou seja, a análise exploratória dos dados desta pesquisa produziu insumos para descrever como os avaliadores da prova oral do exame Celpe-Bras comportam-se nas diferentes instâncias avaliativas. McNamara (1995), ao abordar as características de avaliadores, utilizou termo semelhante (*judges' behaviour*), embora não tenha apresentado uma conceituação para ele. De acordo com o pesquisador (1995, p. 3), diferenças entre avaliadores podem ser entendidas em termo de severidade (ou leniência), por um lado, e aleatoriedade (erro), por outro. Essas características dos avaliadores podem ser evidenciadas para um grupo de candidatos, e não para outros, em relação a determinadas tarefas, e não a outras, em algumas ocasiões, e não em outras. Ou seja, pode haver uma interação entre o avaliador e outra faceta do processo de avaliação.

Ainda segundo o autor (1995), em avaliações de desempenho, uma das formas de reduzir a variabilidade do comportamento dos avaliadores é, normalmente, por meio de treinamento, em que eles são submetidos à realização de avaliações práticas e, posteriormente, é feita uma estimativa da confiabilidade dos resultados dessas avaliações, para determinar se os avaliadores poderão ou não participar do processo.

A definição que propomos de *comportamento avaliativo* também dialoga com a noção de *variabilidade do avaliador*, proposta por Eckes (2015). Segundo o pesquisador, é complexo o processo de mensuração do desempenho de um candidato: o avaliador deve entender esse desempenho, interpretá-lo e atribuir notas com base em um construto e uma grade de avaliação. Esse processo complexo revela a necessidade de se investigar cuidadosamente a qualidade psicométrica das avaliações que são mediadas por avaliadores. O termo *variabilidade do avaliador* geralmente refere-se à variabilidade das pontuações

atribuídas aos candidatos, variabilidade esta que está associada às características dos próprios avaliadores e não ao desempenho dos candidatos (ECKES, 2015, p. 39).

Portanto, o comportamento avaliativo é uma variável importante do processo de mensuração, devendo, com isso, ser analisada, tendo em vista ter suas origens na subjetividade de interpretações e, consequentemente, poder provocar erros, interferindo, assim, na confiabilidade dos resultados do teste.

3.5 Como estimar a confiabilidade dos resultados de um instrumento de avaliação

Vianna (2003), ao discutir avaliações para acesso ao ensino superior, demonstrou preocupação com os resultados dos instrumentos, pela quase ausência de iniciativas para estimativa da confiabilidade, ficando esta no campo das inferências *por intermédio de uma análise qualitativa crítica*. Então, como sair desse campo das inferências e ir para a análise prática? Para responder a essa pergunta, pautamo-nos nos estudos da linguística aplicada e da psicometria, estando a estatística, portanto, presente de forma transversal.

A partir do que tratam alguns pesquisadores (Bachman, 1990; 2004; Moskal e Leydens, 2000; Thompson, 2003a; 2003b; Sawilowsky, 2003; Murphy; Davidshofer, 2005; Urbina, 2007; Scholtes; Terwee; Poolman, 2011; Hauck Filho; Zanon, 2015; Zanon; Hauck Filho, 2015), a maneira de se estimar a confiabilidade é partir dos resultados do teste e da verificação do quanto eles sofrem interferência das variáveis intervenientes no processo avaliativo. Quanto menor for o erro de mensuração, maior será a confiabilidade dos resultados e, consequentemente, a qualidade do instrumento.

Para Lyman (1978 apud Walsh; Betz, 1995), erros em escores de testes estão relacionados a cinco fatores principais: influência do tempo, conteúdo do teste, o examinador do teste ou o responsável por atribuir os escores, a situação em que o teste ocorre e o próprio examinando.

Urbina (2007), por sua vez, categoriza os erros em três fontes. São elas:

> (a) o contexto da testagem (incluindo fatores relacionados ao administrador do teste, ao avaliador e ao ambiente, bem como aos motivos da aplicação do teste); (b) o testando e (c) o teste em si. Alguns erros oriundos destas fontes podem ser minimizados ou eliminados desde que práticas apropriadas de testagem sejam observadas pelas partes envolvidas no processo de desenvolvimento, seleção, administração e pontuação dos instrumentos (URBINA, 2007, p. 125).

Marôco e Garcia-Marques (2006) também chamam a atenção para a existência do erro no processo de mensuração. Segundo os pesquisadores, o erro aleatório, associado à variabilidade observada, é uma característica desejada, mas que deve ser reduzida. Porém, o erro pode ser sistemático e, dessa forma, traduzir uma questão não de confiabilidade, mas de validade.

> O instrumento com erro sistemático é um instrumento com validade reduzida, é um instrumento que está a medir algo que não era suposto medir (mesmo que o faça de forma fiável). Qualquer medida para ser válida, enquanto medida de um dado construto, tem necessariamente de ser fiável. Pelo que, a fiabilidade surge como condição necessária, mas não suficiente, para a validade (MARÔCO; GARCIA-MARQUES, 2006, p. 67).

O ponto inicial de quase todas as teorias sobre confiabilidade, para Murphy e Davidshofer (2005, p. 119), é a ideia de que os escores refletem a influência de dois fatores:
1 - os que contribuem para a consistência: características estáveis do indivíduo ou o atributo que está sendo medido;
2 - os que contribuem para a inconsistência: características que podem afetar os resultados dos testes, mas que não têm relação com o atributo que está sendo medido.

Segundo os autores (2005), essa conceituação é tipicamente representada por: *escore observado = escore verdadeiro + erro de mensuração*.

Uma das primeiras tentativas formais de mensuração em psicologia é a Teoria Clássica dos Testes (TCT), cujo foco está nos escores observados produzidos pelos instrumentos psicométricos e no quanto de erro de medida eles apresentam (HAUCK FILHO; ZANON, 2015, p. 25). Segundo os pesquisadores, o experimento mental da TCT pode ser representado pela equação: $t = X - E$, em que:

t = escore verdadeiro X = escore observado E = erro aleatório.

Hauck Filho e Zanon (2015, p. 26) ainda explicam que o objetivo da TCT é estimar o erro contido nos escores observados, a fim de conhecer melhor o escore verdadeiro *t*. A medida usada para essa finalidade é chamada de *fidedignidade*, ou *confiabilidade*. Em outras palavras, o objetivo de estimar a confiabilidade é determinar o quanto da variabilidade nos resultados dos exames é devido a erros na mensuração e o quanto é devido à variabilidade nos escores verdadeiros (MURPHY; DAVIDSHOFER, 2005, p. 122).

Hauck Filho e Zanon (2015) fazem uma importante consideração acerca da equação ora apresentada, conforme a seguir.

> A fidedignidade determina o quanto da variância ou variabilidade nos escores observados X (ao longo das inúmeras replicações) é devida ao escore verdadeiro *t*, e não ao erro aleatório, ou seja, Var(X)/Var(*t*) (Graham, 2006). Assim, se Var(X) = Var(*t*), então a fidedignidade é igual a 1,00, ocasião em que os escores produzidos são maximamente fidedignos. De fato, na TCT, a fidedignidade é medida por coeficientes cujos valores situam-se entre 0 e 1, sendo aceitos como desejáveis valores acima de 0,70. O único detalhe é que, na vida real, a fidedignidade é calculada para uma amostra de indivíduos que responderam ao instrumento apenas uma vez (...). A alteração requerida na equação, em virtude disso, é tornar o *t* minúsculo (que indica uma constante com valor fixo para o indivíduo) em um *T* maiúsculo, definindo uma variável aleatória, cuja variância agora se dá entre indivíduos, e não apenas intraindivíduos (Borsboom, 2005) (HAUCK FILHO; ZANON, 2015, p. 26-27).

Esses termos-chave são assim explicados por Urbina (2007):

> [...] na teoria clássica dos testes, o *escore verdadeiro* de um indivíduo é conceitualizado como o escore médio em uma distribuição hipotética que seria obtida se o indivíduo se submetesse ao mesmo teste um número infinito de vezes. Na prática, obviamente, é impossível obter tal escore até mesmo para um único indivíduo, o que dirá para muitos. Ao invés de escores verdadeiros, o que derivamos dos testes são os *escores observados* (isto é, os escores que os indivíduos efetivamente obtêm).
>
> Em relação a um único escore, as ideias apresentadas até este ponto podem ser representadas sucintamente por meio da seguinte equação:
>
> $$X_o = X_{verdadeiro} + X_{erro}$$
>
> que expressa o conceito de que qualquer escore observado (X_o) tem dois componentes: um componente de escore verdadeiro ($X_{verdadeiro}$) e um componente de erro (X_{erro}). De um ponto de vista realista, as magnitudes destes dois componentes sempre são desconhecidas. Não obstante, em teoria, o componente do escore verdadeiro é entendido como aquela parte do escore observado que reflete a habilidade, traço ou característica avaliada pelo teste. Inversamente, o componente de erro, que é definido como a diferença entre o escore observado e o escore verdadeiro, representa quaisquer outros fatores que possam influenciar o escore observado como consequência do processo de mensuração (URBINA, 2007, p.122).

Portanto, levando-se em conta o escore verdadeiro, o escore observado e o erro, a partir da análise dos escores, alguns métodos para estimar a confiabilidade são, de acordo com Murphy e Davidshofer (2005, p. 122-128):

(i) **teste-reteste** (*test-retest methods*): consiste em verificar o grau em que os escores de um teste são consistentes de uma aplicação para a outra. Esse método envolve: aplicar o teste para um grupo de indivíduos; reaplicar o mesmo teste para o mesmo grupo, em outro momento; correlacionar o primeiro conjunto de escores com o segundo.

(ii) **Método de formas alternativas** (*alternate forms methods*): consiste em aplicar duas formas de um teste, equivalentes em termos de conteúdo, processos de resposta e

características estatísticas. Esse método envolve: aplicar uma forma do teste (forma A) para um grupo de indivíduos; aplicar, algum tempo depois, a outra forma do teste (forma B) para o mesmo grupo de indivíduos; correlacionar os escores das formas A com os da B.

(iii) **Método das metades** (*Split-half methods*): esse método soluciona, de certa forma, dois problemas práticos do de formas alternativas: a dificuldade em se desenvolver formas alternadas de um teste e a necessidade de aplicações separadas. Envolve: aplicar o teste para um grupo de indivíduos; dividir o teste ao meio; correlacionar os escores de uma metade com os da outra metade do teste.

(iv) **Consistência interna** (*internal consistency methods*): estima a confiabilidade baseada somente no número de itens do teste e a média da intercorrelação entre os itens. Esse método envolve: a aplicação do teste para um grupo de indivíduos; contabilização da correlação entre todos os itens e contabilização da média dessas intercorrelações; estimativa da confiabilidade, utilizando alguma fórmula matemática, como, por exemplo, o coeficiente alfa, descrito mais adiante. A principal vantagem desse método é a sua praticidade, pois requer uma única aplicação do teste e é possível de ser utilizado todas as vezes que o teste for aplicado.

Esses métodos também são tratados por Urbina (2007), que explica as fontes de erro que podem tornar os testes inconsistentes, bem como os coeficientes tipicamente utilizados para estimar a confiabilidade dos escores. É o que consta do quadro a seguir.

Quadro 6 - Fontes de erro de mensuração e coeficientes de confiabilidade

Fonte de erro	Tipo de teste propenso a cada fonte de erro	Medidas apropriadas para estimar erros
Diferenças entre avaliadores	Testes avaliados com algum grau de subjetividade	Fidedignidade do avaliador
Erro de amostragem de tempo	Testes de traços ou comportamentos relativamente estáveis	Fidedignidade de teste-reteste ou coeficiente de estabilidade
Erro de amostragem de conteúdo	Testes para os quais a consistência de resultados é desejada como um todo	Fidedignidade de forma alternativa ou fidedignidade pelo método das metades (*split-half*)
Inconsistência entre itens	Testes que requerem consistência entre os itens	Fidedignidade pelo método das metades ou medidas mais rígidas de consistência interna, como a fidedignidade de Kuder-Richardson 20 (K-R 20) ou o coeficiente alfa (α)
Inconsistência entre itens e heterogeneidade de conteúdo combinadas	Testes que requerem consistência e homogeneidade entre os itens	Medidas de consistência interna e evidências adicionais de homogeneidade
Erros de amostragem de tempo e conteúdo combinados	Testes que requerem estabilidade e consistência dos resultados como um todo	Fidedignidade de forma alternativa com intervalo

Fonte: URBINA (2007, p. 127).

Como se pode notar, há várias fontes de erro que podem interferir nos resultados de um teste e, a depender das suas características e do tipo de erro, há medidas apropriadas para estimar a confiabilidade.

Todos esses métodos apresentados até aqui são utilizados pela Teoria Clássica dos Testes. De acordo com Urbina (2007), uma extensão da TCT é a *teoria da generalizabilidade (teoria G)*, sendo uma abordagem mais abrangente, e que usa métodos de análise de variância (ANOVA) para avaliar os efeitos combinados de múltiplas fontes de variância de erro em escores de teste simultaneamente (URBINA, 2007, p. 142). Apesar de ser considerada um procedimento mais completo para a identificação do componente de variância de erro, é preciso, para a sua utilização, obter múltiplas observações do mesmo grupo de indivíduos em todas as variáveis independentes que podem contribuir para a variância de erro em um dado teste (p. ex., escores em todas as ocasiões, por todos os avaliadores, em formas alternativas etc.) (URBINA, 2007, p. 142).

Bachman, Lynch e Mason (1995) consideram que a *teoria G* é adequada para estimar os efeitos relativos de diferentes fatores nos resultados do teste, sendo esses fatores relacionados, por exemplo, a vários itens no teste, tarefas diferentes e avaliadores diferentes, níveis diferentes de dificuldade para diferentes grupos de candidatos, avaliação diferenciada de grupos distintos etc.

A partir do que apontam Urbina (2007), Bachman, Lynch e Mason (1995), e, tendo em vista a natureza dos dados obtidos para esta tese e o objetivo da pesquisa, é levada em consideração a Teoria Clássica dos Testes, cujos procedimentos são detalhados no próximo capítulo.

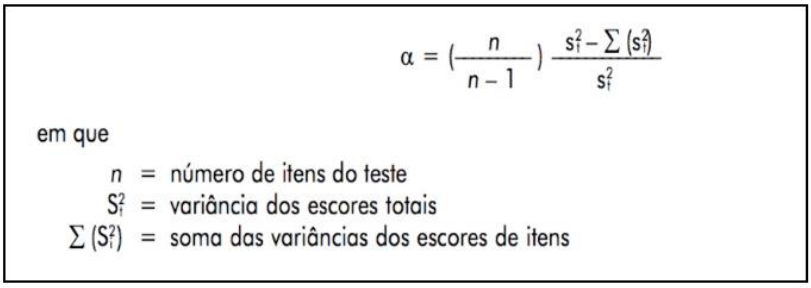
Scholtes, Terwee e Poolman (2011), considerando a TCT, apresentam uma síntese das propriedades mais comuns para estimar a confiabilidade dos resultados de testes em: (a) consistência interna, (b) consistência da avaliação e (c) erro de mensuração, conforme descritas a seguir.

a - Consistência interna: estima o grau de interrelação entre os itens da escala de avaliação. Esse método assume que todos os itens são parte de um construto subjacente, ou seja, unidimensional. Devido a isso, antes mesmo de se verificar a consistência interna, os autores (2011) sugerem que a dimensionalidade da escala seja verificada e isso pode ser feito via Análise dos Componentes Principais.

Posteriormente à verificação da dimensionalidade da escala, a consistência interna pode ser estimada a partir do cálculo de um coeficiente chamado de *Alfa de Cronbach*[33], ou simplesmente *Alfa* (α), sendo o que Pasquali (2010) considera como uma das técnicas mais utilizadas para este fim.

O coeficiente *Alfa de Cronbach* é uma medida que leva em consideração a correlação entre itens, isto é, a correlação entre o desempenho em *todos os itens* de um teste (URBINA, 2007, p. 134). Segundo a pesquisadora (2007, p. 136), o cálculo dessa medida leva em conta o número de itens do teste e a razão entre a variabilidade no desempenho dos testandos em todos os itens e a variância total nos escores do teste. É expresso pela seguinte fórmula:

Figura 4 - Fórmula para cálculo do coeficiente *Alfa de Cronbach*



Fonte: Urbina (2007, p. 136).

Urbina (2007) e Scholtes, Terwee e Poolman (2011) sugerem um α acima de 0,70 para demonstrar que os itens da escala estão suficientemente correlacionados. Já Marôco e Garcia-Marques (2006) detalham os seguintes valores de referência.

Quadro 7 - Valores de referência para verificação da consistência interna

Valor do coeficiente *Alfa de Cronbach*	Interpretação
<0.6	Confiabilidade inaceitável
0.7	Confiabilidade baixa
0.8 a 0.9	Confiabilidade moderada a elevada
>0.9	Confiabilidade elevada

Fonte: MARÔCO; GARCIA-MARQUES, 2006, p. 73 (adaptado).

Os valores do α, portanto, variam entre 0 e 1. Quanto mais próximos de 1, mais consistência interna possuem os itens da escala.

[33] Hogan, Benjamin e Brezinski (2003) apresentaram um estudo realizado no ano 2000, em que foi feito um levantamento da frequência de uso e dos tipos de coeficiente de confiabilidade citados em artigos científicos. Os resultados apontaram que, em 66,5% dos casos (533 em 801), há referência ao Alfa.

b - Consistência da avaliação:

Segundo Scholtes, Terwee e Poolman (2011), a consistência da avaliação pode ser verificada a partir de algumas vertentes, como:

(i) consistência **teste-reteste**: estima a consistência da avaliação a partir da aplicação do teste em dois momentos distintos;

(ii) consistência **inter-avaliadores**: verifica se dois (ou mais) avaliadores têm consenso na atribuição de notas quando utilizam o mesmo instrumento de avaliação para avaliarem o(s) mesmo(s) sujeito(s), num mesmo momento, e

(iii) consistência **intra-avaliadores**: verifica a consistência com que um avaliador faz a avaliação em momentos diferentes.

Os pesquisadores (2011) afirmam que, tanto a consistência inter quanto intra-avaliadores depende primariamente de uma boa capacitação dos avaliadores e uma boa padronização de procedimentos. Segundo eles, há duas maneiras de se estimar a confiabilidade nessas três vertentes. A primeira delas é, para dados categoriais, por meio do ***Coeficiente Kappa***, que verifica o grau de concordância em que os escores foram atribuídos. Os resultados desse coeficiente devem, preferencialmente, atingir valores >0,70. Esse método foi desenvolvido para estimar o grau de concordância entre dois avaliadores depois de corrigir a porcentagem (TOFFOLI, 2015, p. 131). De acordo com a pesquisadora (2015, p. 131-132), a interpretação desse coeficiente é a seguinte:

Kappa = 0: não indica que os dois avaliadores discordam completamente um do outro, mas que eles concordam entre si com a mesma frequência que seria esperada ao acaso;

Kappa >0: indica que os avaliadores concordam entre si com maior frequência do que o esperado ao acaso;

Kappa <0: indica que os avaliadores concordam entre si com menor frequência do que o esperado ao acaso.

Numa escala de avaliação, as notas atribuídas pelos avaliadores ao desempenho de um candidato, por exemplo, podem ser transformadas em categorias, como os níveis de proficiência adotados pelo exame Celpe-Bras. No momento posterior à classificação das unidades em análise nas diversas categorias, é frequente optar-se por uma estratégia que avalie a objetividade dessa classificação a partir de um grau específico de concordância entre dois ou mais elementos avaliadores (*juízes*) – o **acordo inter-juízes** (FONSECA; SILVA; SILVA, 2007, p. 82 – grifos no original). Segundo os autores, o coeficiente *Kappa* é o mais

utilizado para esse fim e pode ser definido como a proporção de acordo entre os juízes após ser retirada a proporção de acordo devido ao acaso (FONSECA; SILVA; SILVA, 2007, p. 83). Ainda de acordo com os autores (2007), não há um valor objetivo específico a partir do qual os valores do *Kappa* são considerados como adequados. Entretanto, Fleiss (1981 apud Fonseca; Silva; Silva, 2007, p. 85) sugere os seguintes valores e classificações.

Quadro 8 - Valores de referência do Coeficiente *Kappa*

Valor de referência do *Kappa*	Interpretação
<.40	Pobre
.40-.75	Satisfatório a bom
>.75	Excelente

Fonte: FLEISS (1981apud FONSECA; SILVA; SILVA, 2007, p. 85).

Retomando as maneiras de se estimar a confiabilidade nas vertentes teste-reteste, inter-avaliadores e intra-avaliadores, segundo Scholtes, Terwee e Poolman (2011), a segunda maneira é, para dados contínuos, por meio da correlação intra-classe (ICC), embora também sejam utilizados os coeficientes de correlação de Pearson ou Spearman.

Das três vertentes apresentadas por Scholtes, Terwee e Poolman (2011), os dados desta tese permitem estimar a consistência da avaliação apenas pela consistência inter-avaliadores, por não ser possível reaplicar o teste e nem identificar os sujeitos avaliadores. Se a avaliação de um teste envolve julgamentos subjetivos, a fidedignidade do avaliador deve ser considerada (URBINA, 2007, p. 139) e, dessa maneira, a consistência inter-avaliadores é tratada nesta pesquisa de forma a promover discussões atinentes à confiabilidade dos resultados do teste e ao comportamento avaliativo.

c – Erro de mensuração:

Para Scholtes, Terwee e Poolman (2011), o erro de mensuração aborda a quantidade absoluta de medida de erros, sendo que a estatística preferida para expressar esse erro é por meio do *erro padrão da média* (EPM / *standard error of measurement – SEM*). Por exemplo, num banco de dados com 1000 pessoas, pretende-se verificar a média de idade de determinado grupo. Para isso, pode-se sortear 300 delas, calcular a média e repetir esse procedimento. Ao final, verifica-se qual foi a variabilidade da média em todos os grupos sorteados. Tem-se, então, o EPM.

3.6 Estimativa da confiabilidade a partir de quadros de referência: norma e critério

De acordo com Bachman (2004), os resultados de testes podem ser interpretados a partir de dois quadros de referência. Se a performance de um grupo de indivíduos for utilizada como base para interpretações de escores, diz-se que elas são referenciadas em normas. Por outro lado, se a base for um determinado domínio ou um nível de habilidade, diz-se que as interpretações dos escores são referenciadas em critério.

Hughes (2003) explica que os testes referenciados em norma, são aqueles em que o desempenho de um candidato é relacionado com o desempenho dos outros, a exemplo dos exames de entrada em universidades, em que há limite de vagas. Segundo Bachman (2004), os resultados desses testes são interpretados em termos de: o quanto os escores dos candidatos estão distantes, acima ou abaixo da média de escores de um determinado grupo.

Por outro lado, os testes referenciados em critério são aqueles em que o desempenho do candidato não é comparado ao de outros, ou seja, não há uma classificação, um *ranking* de notas para uma determinada cota de vagas. O objetivo desse tipo de teste é estabelecer um critério pelo qual um determinado indivíduo possa ser classificado, de acordo com o que se pretende medir, a exemplo de testes de proficiência. Segundo o autor (2003), esses testes, referenciados em critério, apresentam duas vantagens: são capazes de estabelecer padrões significativos em termos do que as pessoas podem fazer, que não mudam com diferentes grupos de candidatos, e de motivar os alunos a atingirem esses padrões.

O quadro a seguir apresenta a distinção entre os dois quadros de referência.

Quadro 9 - Interpretação de testes referenciados em normas *versus* testes referenciados em critérios

Testes referenciados em normas	Testes referenciados em critério
Buscam localizar o desempenho de um ou mais indivíduos em relação ao construto que o teste avalia, em um contínuo criado pelo desempenho de um grupo de referência.	Buscam avaliar o desempenho dos indivíduos em relação a padrões relacionados ao construto em si.
Na sua interpretação, o referencial são sempre as pessoas.	Na sua interpretação, o referencial pode ser: - o conhecimento sobre um domínio de conteúdo, demonstrado em testes padronizados objetivos; - o nível de competência exibido na qualidade do desempenho ou de um produto.
O objetivo primário é fazer distinção entre os indivíduos em termos da capacidade ou traço avaliado por um teste.	O objetivo primário é avaliar o grau de competência de uma habilidade ou conhecimento em termos de um padrão estabelecido de desempenho.

Fonte: Urbina (2007, p. 110, 116-117 – modificado).

Pelas suas características, o Celpe-Bras enquadra-se na classificação de teste *referenciado em critério*, pois, como aponta Bailey (1998, p. 36), determinada pontuação atribuída ao desempenho de um candidato é interpretada em relação a uma meta ou um objetivo preestabelecido (o critério) e não ao desempenho dos outros candidatos. Dito de outra maneira, os dois parâmetros que melhor definem esse tipo de teste e que podem ser visualizados no Celpe-Bras, são o domínio delimitado de competências em que incide a avaliação e a existência de um nível prévio definindo o desempenho satisfatório e não satisfatório (ALMEIDA; VIANA, 2010, p. 243).

Bachman (2004) também faz uma distinção entre esses dois quadros de referência, apresentando, para cada um deles, os procedimentos estatísticos utilizados para estimativa de confiabilidade, conforme destacamos a seguir.

3.6.1 Testes referenciados em norma

Bachman (2004) aponta três modelos que apresentam possibilidades de estimar os efeitos de erros nos escores de um teste e estimar a confiabilidade desses escores. O primeiro modelo é a Teoria Clássica dos Testes, que assume que todo erro de mensuração é aleatório. Os procedimentos para a estimativa da confiabilidade podem ser a partir de:

- consistência interna (método das metades, coeficiente alfa, KR20, KR21);
- estabilidade (teste-reteste);
- equivalência (formas paralelas) e
- consistência do (intra e inter) avaliador.

O autor (2004) apresenta um exemplo em que os procedimentos da TCT podem ser utilizados, o que dialoga com os dados desta pesquisa.

> Considere, por exemplo, um teste de entrevista oral, em que os examinandos são submetidos a várias questões e comandos que eles devem responder. Suponha, também, que diferentes examinandos são entrevistados por diferentes examinadores, e que suas respostas são avaliadas por diferentes avaliadores. Nessa situação, há potenciais fontes de inconsistência, incluindo: (1) diferentes comandos; (2) diferentes examinadores e (3) diferentes avaliadores. Essas potenciais fontes de erro de mensuração podem ser estimadas individualmente pelo modelo da TCT, como consistência interna, consistência inter examinador e consistência intra e inter-avaliadores (BACHMAN, 2004, p. 174-175).

O segundo modelo, de acordo com o autor (2004), é a Teoria G, a partir da qual se pode investigar os efeitos de múltiplas fontes de variância nos escores de um teste, como

tratado anteriormente. O terceiro modelo, por fim, é a Teoria de Resposta ao Item (TRI), que parte da premissa de que a performance de um examinando em determinado item é determinada por dois fatores: o nível de habilidade do examinando no que o item mede (traço latente) e a característica do item.

3.6.2 Testes referenciados em critério

De acordo com Bachman (2004), para estimar a confiabilidade a partir de testes referenciados em critério, é necessário considerar as classificações feitas a partir dos resultados de um teste, tomando como base o fato de que esses testes são utilizados para classificar candidatos em grupos. Em algumas situações, essas classificações são do tipo "aprovado" / "não-aprovado", ou mais detalhadas a partir de determinados níveis, como é o caso do Celpe-Bras, que adota níveis de proficiência linguística (básico, intermediário, intermediário superior, avançado e avançado superior). Para estimar a confiabilidade a partir dessas classificações, o autor (2004) apresenta dois *índices de concordância*, sendo que a escolha entre eles depende de como são feitas as interpretações relativas à seriedade dos erros de mensuração. Essa seriedade (problema) do erro está diretamente relacionada às decisões tomadas a partir dos resultados do teste. Por exemplo: classificar um candidato em um nível superior, enquanto a sua habilidade em determinado domínio reflete um nível inferior, ou o contrário. São as classificações *falso-positiva* e *falso-negativa*, respectivamente. Outro aspecto a ser considerado nessas classificações falsas é o quanto elas se aproximam ou se distanciam de uma nota de corte. Um exemplo pode ser dado a partir dos seguintes questionamentos: tomando como base uma nota de corte de 50, numa escala de 0 a 100, uma classificação *falso-negativa* de 49 pontos será considerada mais ou menos séria do que a uma classificação de 40 pontos? Ou ambos os casos são problemáticos? A decisão quanto a essa seriedade conduz a escolha dos *índices de concordância*.

Se se considera que ambas as classificações erradas são sérias, Bachman (2004) propõe o cálculo do *índice de concordância de perda de limiar*. Para isso, é preciso obter dois conjuntos de escores dos mesmos indivíduos, com diferentes avaliadores, ou escores de diferentes edições de um teste. Exemplo:

		Avaliador 2	
		Aprovado	Não aprovado
Avaliador 1	Aprovado	15	2
	Não aprovado	1	2

Fonte: Bachman (2004, p. 200)

Uma das maneiras de se calcular esse coeficiente de concordância é somar a quantidade da classificação “aprovado”, pelos dois avaliadores, com o número de “não aprovado”, e dividir pelo número total de candidatos, ou seja, (15 + 2) / 20 = 0,850.

Segundo o autor (2004), uma das limitações desse índice é que não se leva em consideração a quantidade de concordância devida ao acaso. Admitindo-se a plausibilidade da concordância devida ao acaso, outro índice que pode ser utilizado é o coeficiente *Kappa*, como apresentado anteriormente. Segundo Bachman (2004, p. 202), entre os dois coeficientes, o *kappa* é mais apropriado para estimar a concordância em testes em que se exige uma competência mínima ou em testes para certificação de candidatos.

Por outro lado, quando se considera que uma pontuação de 40 é mais problemática do que a de 49, em relação à nota de corte de 50, utiliza-se o *índice de concordância de perda do erro quadrático* para cálculo da concordância. Segundo o autor (2004), os dois coeficientes utilizados para essa estimativa são o *phi lambda*, que pode ser obtido a partir de estudos da Teoria G, e o *kappa ao quadrado*.

A partir das características dos dois índices tratados, *de perda de limiar* e *de perda do erro quadrático*, e, tomando como base o exame Celpe-Bras, entendemos que a análise de dados deva considerar o primeiro deles, devido ao fato de que qualquer classificação errada é indesejada. Isso porque trata-se de um exame de alto impacto, a partir do qual decisões importantes são tomadas e, portanto, qualquer erro de classificação reflete diretamente na vida de sujeitos que se submetem ao teste.

Para Bachman (2004, p. 205), as abordagens de estimativa de confiabilidade desenvolvidas a partir de testes referenciados em normas geralmente não são apropriadas para uso em testes referenciados em critério, devido ao fato de que aqueles não fornecem informações sobre o quanto os escores são confiáveis como indicadores de níveis de habilidades dos candidatos.

No entanto, tendo em vista a especificidade dos dados desta pesquisa, é possível utilizar métodos das duas abordagens. Por um lado, considerando que a avaliação da prova

oral do Celpe-Bras envolve julgamentos de natureza subjetiva, é preciso considerar a fidedignidade do avaliador, conforme sugere Urbina (2007), e, para isso, podem ser adotadas técnicas utilizadas para análise de testes referenciados em normas. Por outro lado, considerando que os escores atribuídos ao desempenho dos examinandos são classificados em níveis de proficiência, é preciso verificar o nível de concordância dessas classificações e, para isso, podem ser adotadas as técnicas utilizadas para análise de testes referenciados em critério. Assim, as duas abordagens são adotadas nesta pesquisa.

3.7 Algumas pesquisas sobre confiabilidade

Esta seção apresenta resultados de três pesquisas científicas que utilizaram técnicas para estimativa de confiabilidade de resultados de testes.

Barnwell (1986) realizou um estudo experimental para verificar a confiabilidade inter-rater dos resultados de prova de proficiência oral, utilizando a grade de avaliação do American Council on the Teaching of Foreign Languages/Educational Testing Service (ACTFL/ETS). Participaram da pesquisa, como avaliadores, sete professores de espanhol que receberam um breve treinamento formal de uso da escala para avaliarem 21 entrevistas gravadas (com duração média de 14 minutos) de estudantes de espanhol de níveis variados de proficiência, que participaram como candidatos, de forma voluntária. O foco principal do estudo foi o comportamento de avaliação (*"rating behaviour"*), ou seja, o quanto as avaliações apresentavam concordância entre elas. O autor ressalta que a pesquisa não teve o objetivo de analisar se as entrevistas eram ou não bem conduzidas pelos entrevistadores, mas como eram avaliadas por outros sujeitos que não exerceram a função de entrevistar e que eram inexperientes como avaliadores. As avaliações foram realizadas em três seções, sendo utilizadas sete entrevistas em cada uma, e os avaliadores atribuíram notas individualmente, sem que houvesse interação entre eles.

Para verificar o nível de concordância das avaliações, o pesquisador (1986) adotou dois tipos de correlações: (i) os avaliadores foram divididos em todos os pares possíveis no grupo, somando-se 21 pares, sendo possível avaliar a concordância entre eles; (ii) foi avaliada a concordância entre cada avaliador individual e o grupo como um todo. Os resultados mostram que 41,5% dos pares apresentaram *concordância perfeita*, 44.9%, *concordância aceitável*, e 13,6%, *total discordância*, ou seja, em 86%, houve concordância entre os

avaliadores. Os valores de correlação[34] entre as avaliações foram, em 2 pares de avaliadores: superiores a 0.90; em 14 pares: entre 0.80 e 0.90; em 4 pares: entre 0.70 e 0.80; em 1 par: 0.58. No que se refere à correlação da avaliação individual com o grupo, tem-se que: seis dos sete avaliadores apresentaram valores 0.90 ou mais; um avaliador apresentou valor 0.81. A partir dos resultados, uma das conclusões a que o pesquisador chega é a de que um treinamento mínimo dos avaliadores é capaz de promover níveis altos de concordância entre eles, na maioria dos casos.

Cardoso (2013) propôs a tradução e a adaptação para a língua portuguesa da versão original norte-americana do *Late Life Function and Disability Instrument* (LLFDI)[35] e, por conseguinte, a estimativa da confiabilidade intra-avaliador e inter-avaliadores dos resultados produzidos a partir da versão brasileira. Para a tradução e adaptação do instrumento, participaram cinco tradutores que realizaram duas traduções, uma síntese delas, duas respectivas retro traduções e, por fim, uma revisão e avaliação da equivalência semântica e cultural por um comitê multidisciplinar. Os sujeitos analisados foram 45 idosos a partir de 60 anos de idade sem alterações cognitivas, de ambos os sexos, residentes em Belo Horizonte. O teste foi aplicado sob a forma de entrevista, por dois avaliadores previamente treinados e que seguiram orientações padronizadas propostas nas instruções iniciais do instrumento.

Para a análise da confiabilidade, foram consideradas três etapas: as duas primeiras com a administração do instrumento por dois examinadores independentes (1 e 2), num mesmo momento, e, a terceira, com a reaplicação do instrumento pelo examinador 1, num intervalo de 8 a 10 dias após a primeira aplicação. Para estimar a confiabilidade intra-examinador, foi calculado o Coeficiente de Correlação Intraclasse (ICC) e, para estimar a confiabilidade inter-examinador, o Coeficiente de Correlação de Concordância (CCC), ambos os coeficientes analisados via software SPSS. Dos resultados apresentados, destaca-se: os resultados da versão brasileira do LLFDI mostraram boa confiabilidade intra e inter-examinadores, com elevados índices nos dois componentes da escala (incapacidade e função). O fato de a confiabilidade dos resultados terem apresentado elevados índices contribui para qualificar o instrumento.

---

[34] O autor não detalha os procedimentos utilizados, mas informa que houve tratamento estatístico dos escores não paramétricos.

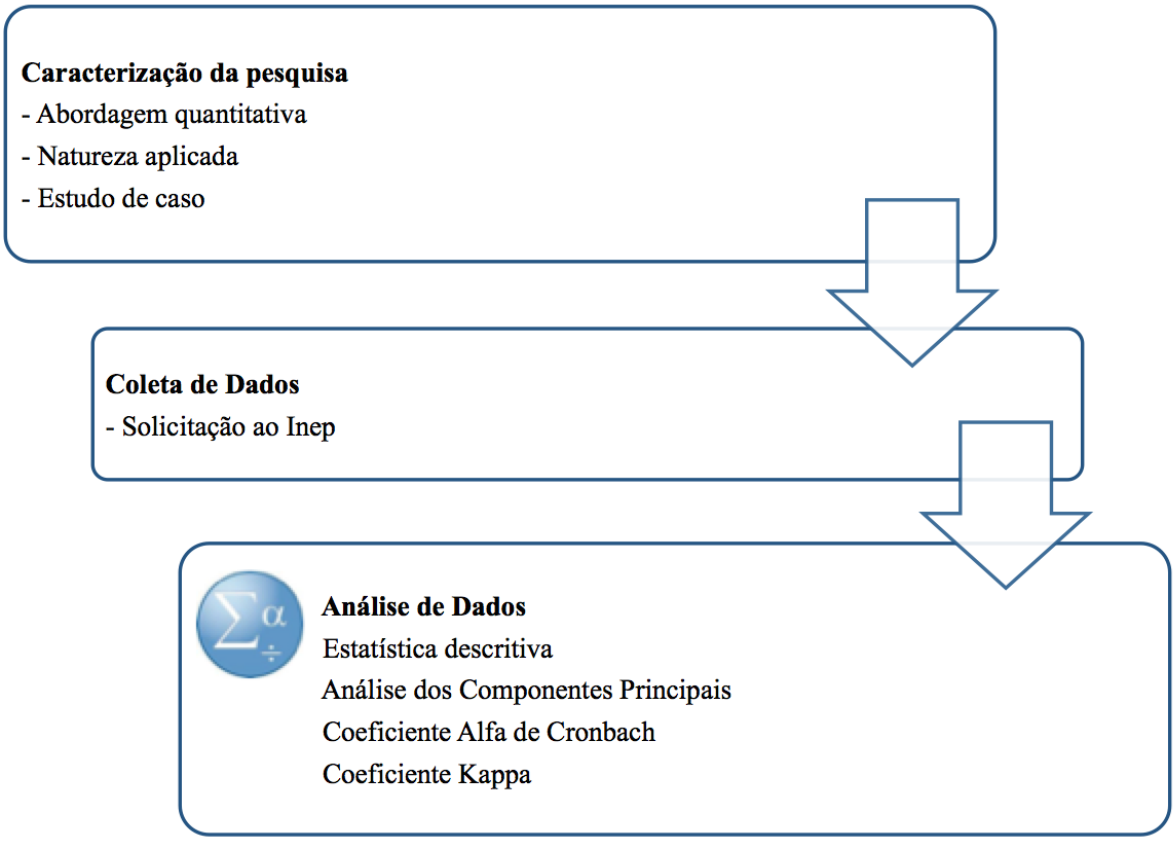
[35] De acordo com Cardoso (2013, p. 17), o LLFDI é um instrumento utilizado para avaliar a função e a incapacidade de pessoas idosas residentes na comunidade. A versão brasileira, a partir da qual a pesquisa foi realizada, foi desenvolvida por um grupo de pesquisa da UFMG.

Davis (2016) investigou, com base em materiais do teste TOEFL iBT, o efeito da *capacitação* e da *experiência avaliativa* nos padrões de avaliação em contexto de teste oral, com o objetivo de identificar de que maneira esses aspectos interferem: (i) na severidade do avaliador e na consistência interna; (ii) na consistência avaliativa, definida pela concordância com escores previamente estabelecidos e (iii) na consulta de respostas-padrão por parte dos avaliadores. Participaram da pesquisa 20 professores de inglês que não haviam atuado anteriormente como aplicadores do TOEFL iBT, mas já tinham experiência no ensino de inglês para alunos de nível intermediário ou superior. Para a realização do estudo, foram utilizadas gravações de provas orais que estão disponíveis no banco de dados do TOEFL (*TOEFL iBT Public Use Dataset*). O desempenho dos candidatos foi avaliado a partir de critérios constantes da grade holística do exame, que prevê três domínios: transmissão (pronúncia, entonação e fluência), uso da língua (gramática e vocabulário) e desenvolvimento do tópico (detalhamento e coerência do conteúdo). Os procedimentos seguidos incluíram: uma seção de orientação aos avaliadores, para certificar de que eles estavam familiarizados com os procedimentos do estudo (tempo, tarefas, grade de avaliação, software adotado); uma seção de avaliação, para definição dos critérios de avaliação, disponibilização de exemplos de avaliação (respostas-padrão) e a avaliação propriamente dita de 100 provas; uma seção de treinamento, para revisão da grade e discussão sobre os seus escores; mais três seções de avaliação, cada uma com 100 provas, totalizando, portanto, a avaliação de 400 provas que correspondem a 8000 escores atribuídos pelos 20 avaliadores.

Dos resultados da pesquisa, destacam-se: (i) a confiabilidade inter-rater (observada pelos valores de correlação de Pearson) e os níveis de concordância entre os avaliadores (pelo cálculo do coeficiente Kappa) apresentaram uma modesta melhora após o treinamento, o que indica que a variabilidade avaliativa diminui um pouco após o treinamento; (ii) o treinamento pareceu ter pouco efeito sobre a variabilidade na severidade do avaliador (verificada pelo modelo Rasch); (iii) após o treinamento, foi observada pouca alteração na consistência do avaliador ou severidade de avaliação, sugerindo que a experiência adicional teve pouco impacto nesses aspectos, mas que isso pode ser devido ao curto período em que os dados foram coletados; (iv) o treinamento pareceu ter pouco efeito sobre a frequência de consulta a respostas-padrão, pois essa frequência pareceu ser mais uma questão de estilo pessoal; (v) pelo fato de todos os avaliadores terem sido submetidos a um treinamento, não foi possível verificar o efeito da experiência na ausência de treinamento, o que mostra ser um importante objeto de pesquisa.

Neste capítulo, apresentamos o referencial teórico da pesquisa, tratando das variáveis que podem interferir na confiabilidade dos resultados dos testes, bem como das formas de estimativa de confiabilidade. O capítulo seguinte é destinado aos aspectos metodológicos.

CAPÍTULO 4
DADOS E PROCEDIMENTOS



4 DADOS E PROCEDIMENTOS

Este capítulo é dedicado às questões metodológicas da pesquisa, em que são explicitados a natureza, características e procedimentos da pesquisa, além do detalhamento do percurso traçado para coleta e exploração dos dados.

4.1 Natureza, características e procedimentos da pesquisa

Segundo Bachman (2004, p. 3), os testes de língua tornaram-se parte do sistema educacional e da nossa sociedade, sendo seus escores utilizados para fazermos inferência sobre a proficiência de indivíduos e para tomarmos decisões sobre eles. Para que sejam instrumentos que forneçam informações importantes, seus escores precisam ser confiáveis. Além disso, é preciso que as formas como os interpretamos e utilizamos sejam válidas.

Ainda segundo esse autor, se quisermos assegurar que o uso dos testes é adequado, é necessário mostrar evidências que suportem isso e

> [...] um importante tipo de evidência que coletamos para apoiar o uso dos testes é o que deriva de dados quantitativos – os escores dos testes e os testes como um todo – e as análises estatísticas apropriadas desses dados. A compreensão da natureza dos dados quantitativos e como analisá-los estatisticamente são, portanto, uma parte essencial dos testes (BACHMAN, 2004, p. 3).

Desta feita, justificamos o interesse em trabalharmos com métodos quantitativos para as análises de dados. Entendemos que a interpretação dos escores da prova oral do Celpe-Bras, por meio de inferências estatísticas, são um passo importante para chegarmos a algumas conclusões sobre o comportamento avaliativo dos sujeitos que os atribuem e, consequentemente, sobre a confiabilidade. É nesse sentido que reafirmamos nesta investigação o estabelecimento necessário de um diálogo entre a Linguística Aplicada e a Estatística.

Trata-se de uma pesquisa de **abordagem quantitativa** que, segundo Dörnyei (2007), envolve procedimentos de coleta de dados cujos resultados são principalmente dados numéricos e analisados por um software estatístico, como, por exemplo, o *Statistical Package for the Social Sciences* (SPSS).

De acordo com as categorizações tratadas por Pradanov e Freitas (2013, p. 51), esta pesquisa tem características de **natureza aplicada**, pois visa gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos e envolve verdades e interesses

locais. Quanto aos objetivos, é **descritiva**, pois visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis. Quanto aos procedimentos, trata-se de um **estudo de caso**, pois será analisado um caso específico: o comportamento de avaliadores da prova oral do Celpe-Bras e sua possível interferência na confiabilidade dos resultados do exame, a partir de um estudo longitudinal. O estudo longitudinal desta pesquisa é caracterizado pelos dados relativos a sete edições consecutivas do exame, das quais acreditamos que os avaliadores, em sua maioria, sejam os mesmos. De acordo com Pradanov e Freitas (2013, p. 62), o estudo de caso tem o objetivo de explorar situações da vida real, cujos limites não estejam claramente definidos, além de descrever a situação do contexto em que a investigação está sendo feita.

4.2 Coleta dos dados

Para a realização desta pesquisa, foram solicitados ao Inep os dados da prova oral de sete edições do exame Celpe-Bras, com as seguintes informações: notas atribuídas na primeira e na segunda instâncias, pelo observador (considerando-se todos os critérios da grade de avaliação) e pelo entrevistador; notas atribuídas na terceira instância; notas finais da prova oral. Foi disponibilizado um banco de dados constando: seis edições com as notas atribuídas na primeira instância e as notas finais da prova oral; uma edição completa, com as todas as notas atribuídas nas três instâncias e as notas finais da prova oral. Para manutenção do sigilo das informações, todas as sete edições estão identificadas apenas por números cardinais (1 a 7).

Para a liberação dos dados[36], foi atendida a norma que traz o Guia do Examinando – *versão simplificada* (BRASIL, 2013d, p. 10). A solicitação dos dados deu-se em duas etapas: (i) cadastro da demanda no Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC[37]); (ii) envio de documentação dos pesquisadores envolvidos e termos de compromisso e sigilo à Diretoria da Educação Básica (DAEB).

Conforme tratado por Marôco (2014), dada a impossibilidade de realizar análises a partir da *população teórica*, que seria a totalidade das notas de todos os examinandos e de

---

[36] Registramos o nosso agradecimento ao Inep, pela disponibilização dos dados.

[37] De acordo com informações disponíveis no endereço eletrônico https://esic.cgu.gov.br/sistema/site/index.html (acesso em 11 de março de 2017), "o Sistema Eletrônico do Serviço de Informações ao Cidadão (e-SIC) permite que qualquer pessoa, física ou jurídica, encaminhe pedidos de acesso à informação, acompanhe o prazo e receba a resposta da solicitação realizada para órgãos e entidades do Executivo Federal. O cidadão ainda pode entrar com recursos e apresentar reclamações sem burocracia".

todas as edições até então efetivadas, foi necessário definir o grupo de estudo, ou seja, a *população de estudo*, que são as notas dos examinandos das sete edições em análise. Para a edição completa, chamada de Edição 5, os dados foram divididos em quatro amostras para melhor descrição e análise (Amostras A, B, C e D).

Considerando-se as três instâncias de avaliação, a figura a seguir detalha as amostras da Edição 5. Ressaltamos que no Capítulo II detalhamos cada uma das instâncias e sugerimos que suas denominações fossem alteradas para, apenas, *primeira*, *segunda* ou *terceira* instância, sendo que essas novas denominações são as utilizadas no Capítulo V (Análise dos dados e discussão dos resultados).

Figura 5 - Amostras da edição 5

Exami-nando	1ª instância posto aplicador		2ª instância compatibilização		3ª instância nota de consenso
	Observador Σ	Entrevistador	Observador Σ	Entrevistador	
1	4,79	5,00			
6	2,50	3,00			
89	5,00	5,00			
112	3,58	3,00			
197	4,54	3,00	4,08	4,00	
440	3,58	2,00	4,50	3,00	3,00
3.200	4,92	3,00	4,71	3,00	4,00

Amostra A
Amostra D
Amostra B
Amostra C

Legenda

Amostra A = provas avaliadas em 1ª instância = população de estudo

Amostra B = provas cuja avaliação em 1ª instância resultou em discrepância significativa

Amostra C = provas cuja avaliação em 2ª instância resultou em discrepância significativa

Amostra D = provas avaliadas em 1ª instância e que não apresentaram discrepância significativa

Fonte: elaborado pela autora, 2018.

A quantidade de provas que compõe cada uma das amostras consta do quadro a seguir.

Quadro 10 - População de estudo

Edições	Quantidade de provas por amostra			
	A	**B**	**C**	**D**
1	3.456			
2	4.163			
3	4.513			
4	4.448			
5	4.585	733	130	3.852
6	4.709			
7	3.957			
Total	29.831			

Fonte: elaborado pela autora, 2018.

Em resumo, a Edição 5 contém notas atribuídas nas três instâncias de avaliação e estão distribuídas em quatro amostras: a Amostra A contém as notas (4.585) de todos os examinandos avaliados na primeira instância; a Amostra B contém as notas (733) que foram consideradas discrepantes na primeira instância e avaliadas em segunda; a Amostra C contém as notas (130) que foram consideradas discrepantes na segunda instância e avaliadas em terceira e a Amostra D contém as notas (3.852) que não foram consideradas discrepantes na primeira instância. As demais edições, por fim, contêm as notas apenas da Amostra A e as notas finais da prova oral. A população de estudo, portanto, envolve notas de 29.831 examinandos.

4.3 Métodos e técnicas de exploração e análise dos dados

Para a análise dos dados, foi utilizado o *software* **SPSS *Statistics***[38], versão 23, sendo que os resultados são apresentados em três etapas, a saber:

(1) análise exploratória dos dados, em que é detalhado o comportamento avaliativo dos atribuidores de nota a partir dos níveis de proficiência adotados pelo exame Celpe-Bras, das notas finais e do grau de similitude das avaliações.

[38] O **IBM SPSS Statistics** ou, simplesmente, **SPSS Statistics**, é, por tradição, o programa de eleição dos cientistas das ciências sociais. Este programa é produzido e comercializado pela 'SPSS, an IBM Company', originária de Chicago nos EUA (MARÔCO, 2014, p. 65).

(2) inferências estatísticas, em que são mostradas medidas de tendência central (média), de dispersão (desvio padrão, valores mínimos e máximos e o coeficiente de variação de Pearson), tanto em relação às notas dos avaliadores em geral quanto às notas de cada um dos critérios da grade analítica. Além disso, foram utilizados:

- tabulação cruzada de frequência e porcentagem de duas ou mais variáveis tomadas juntas (LEVIN; FOX; FORDE, 2012, p. 439). Por exemplo, foi possível realizar uma tabulação cruzada das variáveis *adequação lexical* e *adequação gramatical* para verificar o percentual de ocorrência da maneira como os examinandos foram avaliados (com notas iguais, maiores ou menores) na primeira e na segunda instâncias. *Tabulação cruzada* refere-se ao cruzamento das variáveis, em linhas e colunas;

- testes de hipótese, que permitem avaliar a validade (ou não) de uma afirmação sobre determinada característica da população, usando para isso os dados de uma amostra retirada dessa população (PINHEIRO et al, 2015, p. 262). Antes mesmo de realizarmos os testes de hipótese, foi necessário verificar se há distribuição normal das amostras. Segundo Marôco (2014), a distribuição normal, também chamada de distribuição gaussiana, é a função de densidade de probabilidade provavelmente mais importante no processo de inferência estatística.

(3) estimativa da confiabilidade, em que foram considerados:

3.1 Uma análise preliminar aos estudos da confiabilidade, para a verificação da dimensionalidade da escala via Análise dos Componentes Principais (ACP), conforme sugerem Scholtes, Terwee e Poolman (2011). As variáveis selecionadas para a ACP foram os critérios avaliados pelo observador, partindo do pressuposto de que eles medem o construto *desempenho oral dos examinandos*. A nota atribuída pelo entrevistador não foi considerada, tendo em vista ser única.

Para analisar a aplicabilidade da ACP nos dados, foram considerados o teste de esfericidade de Bartlett: $p<0,001$, conforme sinaliza Maroco (2014, p. 500), e o Kaiser-Meyer-Olkin (KMO), conforme valores descritos no Quadro 11, a seguir.

Quadro 11 - Valores de referência do Kaiser-Meyer-Olkin (KMO)

Valor de KMO	Recomendação relativamente à ACP
]0.9; 1.0]	Excelente
]0.8; 0.9]	Boa
]0.7; 0.8]	Média
]0.6; 0.7]	Medíocre
]0.5; 0.6]	Mau, mas ainda aceitável
≤ 0.5	Inaceitável

Fonte: adaptado de SHARMA, 1996 apud MAROCO, 2014, p. 477.

3.2 O cálculo do coeficiente *Alfa de Cronbach*, para verificar a consistência interna dos itens da escala. Os valores de referência do coeficiente estão descritos no Quadro 7, no Capítulo 3, e reapresentado a seguir.

Valor do coeficiente *α Cronbach*	Interpretação
<0.6	Confiabilidade inaceitável
0.7	Confiabilidade baixa
0.8 a 0.9	Confiabilidade moderada a elevada
>0.9	Confiabilidade elevada

Fonte: MARÔCO; GARCIA-MARQUES, 2006, p. 73 (adaptado).

3.3 O cálculo do coeficiente *Kappa*, para verificar o nível de concordância entre os avaliadores, considerando o quanto dessa concordância é devida ao acaso. Os valores de referência do coeficiente *Kappa* estão descritos no Quadro 8, no Capítulo 3, e reapresentado a seguir.

Valor de referência do *Kappa*	Interpretação
<.40	Pobre
.40-.75	Satisfatório a bom
>.75	Excelente

Fonte: FLEISS (1981 apud FONSECA; SILVA; SILVA, 2007, p. 85).

A partir dessas três etapas, a pergunta de pesquisa *o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste?* pode ser respondida.

4.4 *Outliers x* Discrepância: distinções conceituais

Tendo em vista que esta tese estabelece um diálogo entre as áreas Estatística e Linguística Aplicada, cumpre-nos esclarecer uma distinção entre os termos *observações discrepantes ou outliers*, utilizado nos estudos estatísticos, e *discrepância ou notas*

discrepantes, utilizado quando da explicação do compósito de notas do desempenho dos examinandos do Celpe-Bras. O primeiro termo, conforme explicam Pinheiro *et al* (2015), refere-se a valores que estão muito afastados, para mais ou para menos, de outros valores em um banco de dados e, devido a isso, podem interferir no resultado das análises estatísticas.

> Assim sendo, é útil que tenhamos disponível um critério de detecção de observações discrepantes. Uma vez detectada a presença de uma observação discrepante, poderá ser tomada a decisão de repetir aquele experimento ou meramente expurgar aquele dado da amostra (ou até mesmo mantê-lo, se for encontrada uma explicação plausível para aquela discrepância) (PINHEIRO et al, 2015, p. 26).

O segundo termo, por sua vez, *discrepância ou notas discrepantes*, refere-se, nesta pesquisa, à diferença (*gap*) entre notas atribuídas por avaliadores do exame Celpe-Bras, diferença esta que pode levar à necessidade de reavaliação da prova do examinando. Chamamos de *discrepância significativa* quando, por exemplo, a diferença entre as notas atribuídas pelo AI e pelo AO for igual ou maior que 1,5 ponto, numa escala de 0 a 5[39].

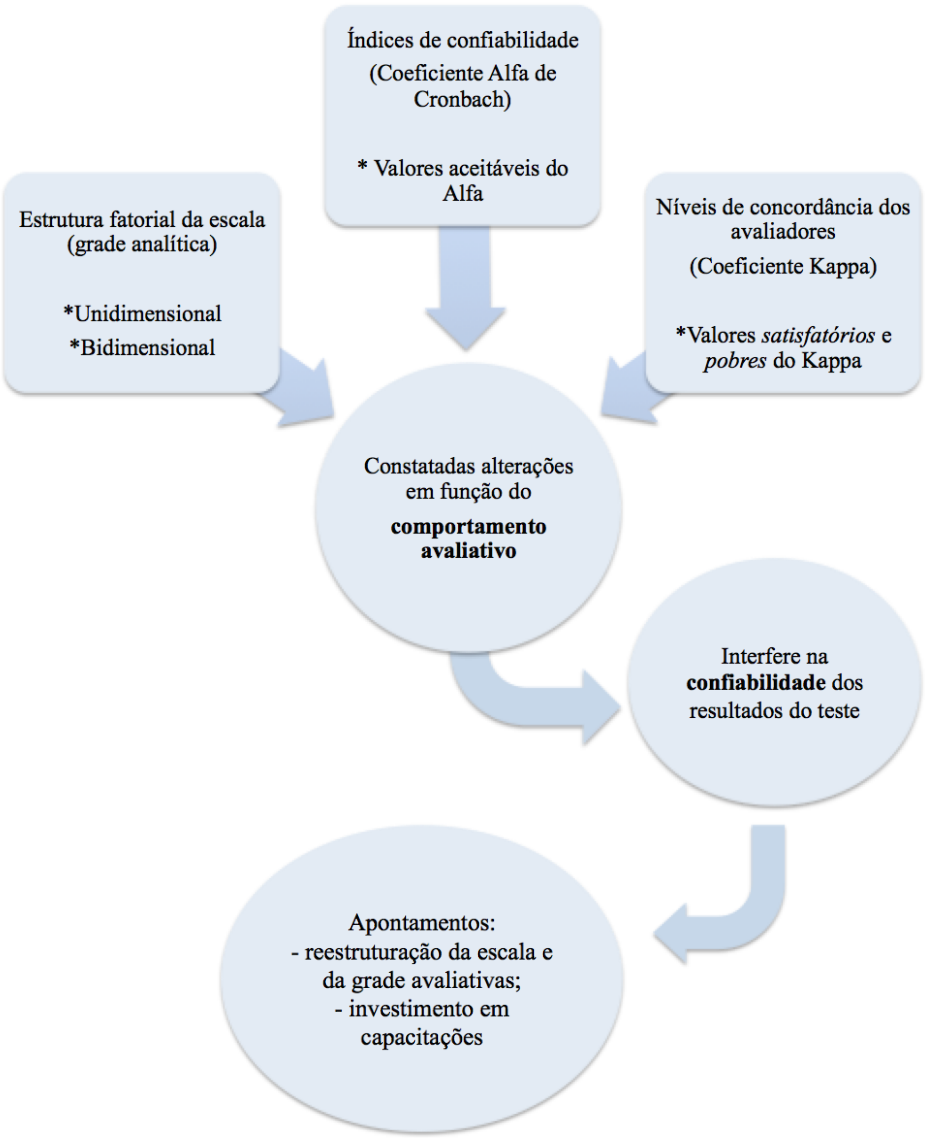
As notas discrepantes, portanto, caracterizam-se como um fator de alerta na avaliação da proficiência do examinando, tendo em vista que podem sinalizar um possível problema da avaliação e que merece atenção para a garantia de escores confiáveis. Elas, então, não são excluídas ou desconsideradas, mas, em algumas situações, levam à reavaliação da prova do examinando. Do mesmo modo, as notas discrepantes que compõem o banco de dados desta tese (embora o termo tenha o mesmo sentido de *valores afastados*, como os *outliers*) não são, num primeiro momento, consideradas *outliers*, devido ao fato de que são discrepantes na comparação entre os escores atribuídos por dois avaliadores (ou pelas duas partes, PE e PO) ao desempenho de cada examinando e não em comparação aos outros valores dos outros examinandos, constantes do banco de dados.

Tendo mostrado, neste capítulo, as características da pesquisa e o seu percurso metodológico, o próximo, que está dividido em quatro partes, é dedicado à apresentação e análise dos dados, bem como à discussão dos resultados.

---

[39] Os casos de *discrepância significativa* são apresentados no Capítulo 2, em que é explicitado o processo de avaliação do exame.

CAPÍTULO 5
ANÁLISE DOS DADOS E DISCUSSÃO DOS RESULTADOS



5 ANÁLISE DOS DADOS E DISCUSSÃO DOS RESULTADOS

Este capítulo é dedicado à apresentação e análise dos dados, bem como à discussão dos resultados. As análises têm como suporte os pressupostos da confiabilidade, com base em estudos da Psicometria, Estatística e Linguística Aplicada.

O capítulo está dividido em quatro partes, a saber:

Parte I - análise exploratória dos dados

Esta parte trata de um conjunto de técnicas de tratamento de dados que, sem implicar fundamentação matemática mais rigorosa, nos ajuda a fazer uma sondagem do terreno, ou seja, tomar um primeiro contato com a informação disponível (PINHEIRO *et al*, 2015, p. 2).

Parte II – inferências estatísticas

São apresentadas algumas informações estatísticas da população de estudo, bem como realizados testes de hipóteses.

Parte III - estimativa da confiabilidade

Para discussões sobre a confiabilidade dos resultados das sete edições do exame, são apresentados (i) uma análise preliminar aos estudos da confiabilidade, pela utilização da Análise dos Componentes Principais (ACP), que mostra a dimensionalidade da escala avaliativa; (ii) os índices de confiabilidade, por meio do cálculo do coeficiente *Alfa de Cronbach*, e (iii) o nível de concordância entre os avaliadores, por meio do cálculo do coeficiente *Kappa*.

Parte IV - discussão dos resultados

Estabelece-se um diálogo entre os resultados apresentados nas três primeiras partes.

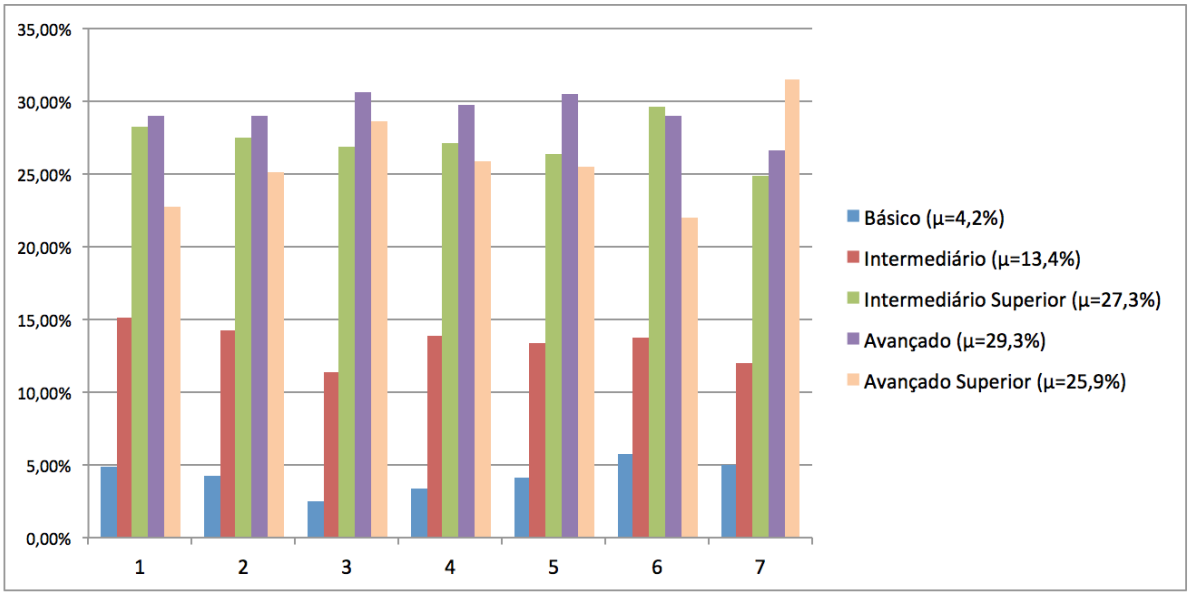
PARTE I – ANÁLISE EXPLORATÓRIA DOS DADOS

5.1 Os níveis de proficiência oral da população de estudo

Iniciamos a apresentação dos dados com base nos níveis de proficiência oral dos examinandos. Os gráficos a seguir partem de uma visão geral sobre a população de estudo (sete edições) até chegar às particularidades de cada amostra da edição 5, revelando, assim, indícios do comportamento dos avaliadores da prova oral do exame Celpe-Bras.

No processo avaliativo das interações face a face, tanto observador quanto entrevistador fazem a avaliação do desempenho oral dos examinandos com base em grades (uma analítica e outra holística – ANEXOS A a C) e atribuem notas que definem os níveis de proficiência. No Gráfico 2, é mostrado o percentual dos níveis de proficiência atribuídos aos examinandos em cada uma das edições (e a média populacional (μ)), no que se refere à **nota final da prova oral**, ou seja, ao último resultado da avaliação do desempenho oral dos examinandos, após todas as análises de discrepâncias.

Gráfico 2 - Níveis de proficiência da população de estudo, por edição



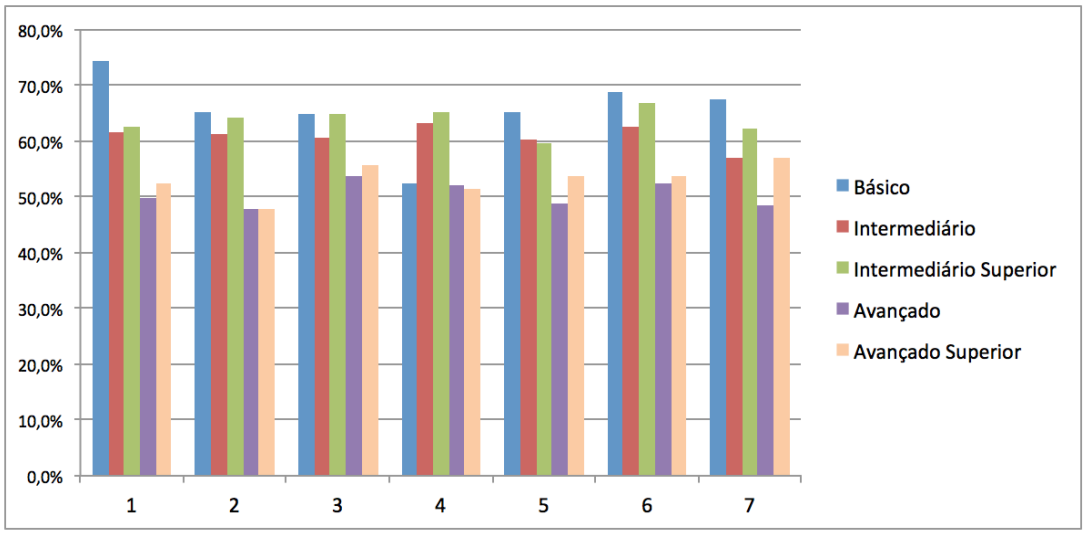
Fonte: elaborado pela autora, 2018.
Nota: - gráfico elaborado com base no Apêndice 1.1;
- N=29.831.

Nas sete edições, nota-se que, em média, 4,2% dos examinandos foram classificados no nível Básico e, a grande maioria, em níveis que permitem a certificação. Cerca de 80% dos examinandos, em cada edição, foram classificados nos níveis *Intermediário Superior*,

Avançado e *Avançado Superior*, tendo maior destaque o nível *Avançado*. Isso pode sinalizar para uma boa preparação dos examinandos, por um lado, e boa capacitação dos seus professores, por outro.

Para detalhar sobre o processo avaliativo, o Gráfico 3, a seguir, mostra o percentual em que os mesmos examinandos foram avaliados nos mesmos níveis tanto pelo observador quanto pelo entrevistador, na primeira instância, nas sete edições.

Gráfico 3 - Percentual de concordância de avaliação em 1ª instância, por nível de proficiência e por edição



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado com base nas tabulações cruzadas apresentadas no Apêndice 1.2;
- N=29.831.

O nível que apresenta maior percentual de concordância é o *Básico* (6 em 7 edições). Os níveis mais altos (*Avançado* e *Avançado Superior*) lideram os menores percentuais de concordância, ao longo das edições, o que significa que os avaliadores entram mais em desacordo na medida em que os examinandos são mais proficientes. Esse dado pode sinalizar que, quanto mais proficientes são os examinandos, mais subjetivos são os escores a eles atribuídos por avaliadores distintos.

Os maiores percentuais de concordância verificados foram:

- *Básico*: 74,5% (edição 1);
- *Intermediário*: 63,2% (edição 4);
- *Intermediário Superior*: 67% (edição 6);
- *Avançado*: 53,8% (edição 3) e
- *Avançado Superior*: 57,1% (edição 7).

Como se pode notar, com exceção do nível *Básico* da 1ª edição analisada, nenhum outro nível em nenhuma outra edição atingiu 68% de concordância entre os avaliadores da primeira instância. Esse resultado indica para uma necessidade de revisão dos descritores das grades de avaliação e consequente capacitação dos avaliadores.

Ressaltamos que os percentuais apresentados não levam em conta nenhuma análise estatística. Na parte III deste capítulo, é mostrado, por meio do cálculo do coeficiente *Kappa*, o quanto dessa concordância é devida ao acaso.

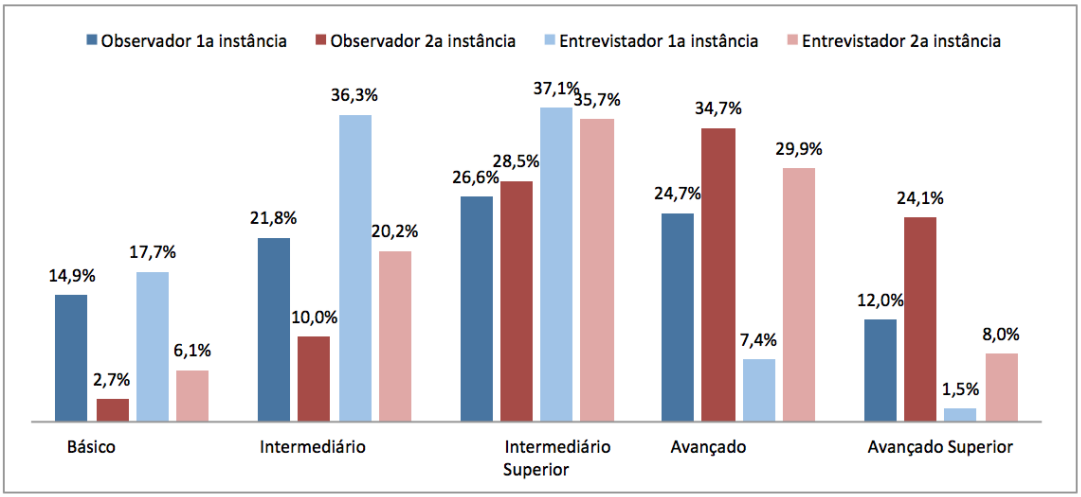
5.2 Os níveis de proficiência oral da edição 5

5.2.1 Na visão dos avaliadores

Como tratado no capítulo 4, a edição 5 do *corpus* desta pesquisa é composta pelas notas atribuídas nas três instâncias de avaliação. Nesta seção, são feitas considerações acerca das Amostras B e C de referida edição, ou seja, das provas cujas notas atribuídas pelo observador e pelo entrevistador foram consideradas discrepantes e, por conseguinte, reavaliadas.

Do total de provas avaliado na primeira instância (4.585), 733 delas tiveram notas consideradas discrepantes e é esse o número de provas que compõe a **Amostra B**. O Gráfico 4, a seguir, mostra o nível de proficiência dos examinandos tanto na visão dos avaliadores da primeira quanto da segunda instâncias.

Gráfico 4 - Edição 5: Amostra B - níveis de proficiência na visão dos avaliadores da 1ª e 2ª instâncias



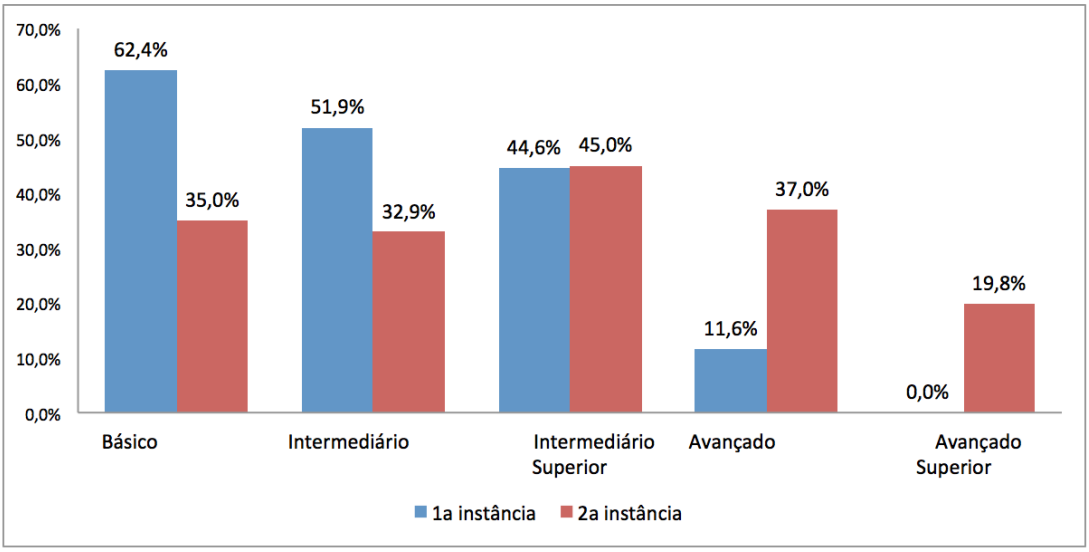
Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas nos Apêndices 1.3 e 1.4;
- N=733.

No Gráfico 4, verifica-se o comportamento dos avaliadores tanto na primeira quanto na segunda instâncias de avaliação da Amostra B. No que se refere aos avaliadores da primeira instância, nota-se que o entrevistador tende a dar menos notas relativas aos níveis *Avançado* e *Avançado Superior* do que o observador. Esse comportamento mantém-se em relação aos avaliadores da segunda instância.

Estabelecendo uma comparação entre as duas instâncias de avaliação, verifica-se uma diminuição na quantidade de examinandos que inicialmente obtiveram os níveis *Básico* e *Intermediário*. Ou seja, os examinandos foram avaliados como mais proficientes na segunda do que na primeira instância.

Esses resultados não mostram, porém, qual o percentual em que os mesmos examinandos foram avaliados nos mesmos níveis pelo observador e pelo entrevistador, na primeira e segunda instâncias. Para a verificação dessa ocorrência, foram feitas tabulações cruzadas, cujos resultados estão resumidos no Gráfico 5, a seguir.

Gráfico 5 - Edição 5: Amostra B - percentual de concordância entre avaliadores (AI e AO), por instância



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas nos Apêndices 1.3 e 1.4. Os percentuais correspondentes estão destacados em negrito, na diagonal das tabelas;
- N=733.

Os maiores percentuais de concordância entre observador e entrevistador apresentam-se na primeira instância: nos níveis *Básico* e *Intermediário*; na segunda instância: nos níveis *Intermediário Superior* e *Avançado*, ou seja, os avaliadores da primeira instância concordam mais entre si apenas nos níveis mais baixos de proficiência. Merecem destaque também as seguintes ocorrências:

- nível *Intermediário Superior*: os percentuais de concordância são equilibrados entre as duas instâncias;
- nível *Avançado Superior*: não apresenta concordância na avaliação realizada em primeira instância;
- todos os níveis: o maior percentual de concordância, nas duas instâncias, não atinge 63%.
- segunda instância de avaliação: embora os avaliadores da segunda instância estejam entre os mais experientes do quadro de colaboradores, comparando-se a 1ª e a 2ª instâncias, os percentuais de concordância entre eles não atingem 46%, em nenhum dos níveis de proficiência. Uma possível justificativa para isso é que as interações que são reavaliadas, devido à existência de discrepâncias, apresentam fatos imprevistos e imprevisíveis que não são contemplados (e é difícil que sejam) nas grades. Alguns desses fatos podem ser melhor avaliados na primeira instância, já que os avaliadores encontram-se face a face com o candidato.

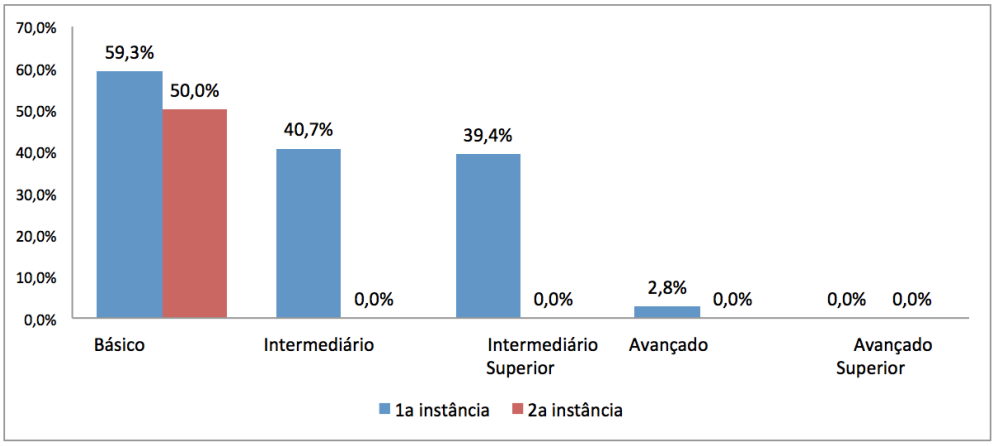
Esses resultados apontam para duas possíveis interpretações: (i) a necessidade de revisão dos descritores das grades de avaliação e consequente capacitação dos avaliadores; (ii) a maior complexidade de avaliação quando o examinando encontra-se em níveis mais altos de proficiência do examinando, o que, de certa forma, remete ao item (i). Além disso, os resultados permitem remeter ao que He e Young (1998) consideram uma ameaça à confiabilidade: quando dois avaliadores diferentes julgam, de forma diferente, a habilidade oral de um mesmo sujeito.

Na parte III deste capítulo, é apresentado o quanto desses percentuais de concordância é devido ao acaso.

Os dados apresentados nos Gráficos 4 e 5 referem-se às 733 provas que compõem a Amostra B. Na avaliação dessas provas em segunda instância, 130 novas discrepâncias significativas foram geradas, sendo este o total que compõe a **Amostra C** e sobre a qual são apresentados os resultados a seguir. Essa amostra, portanto, contém as avaliações feitas em primeira, segunda e terceira instâncias. Na terceira, é atribuída apenas uma nota, com base na grade holística, sendo esta a nota final do desempenho oral do examinando.

O Gráfico 6, a seguir, mostra o percentual em que os mesmos examinandos foram avaliados nos mesmos níveis pelo observador e pelo entrevistador, na primeira e segunda instâncias, considerando a Amostra C.

Gráfico 6 - Edição 5: Amostra C - percentual de concordância entre avaliadores, por instância

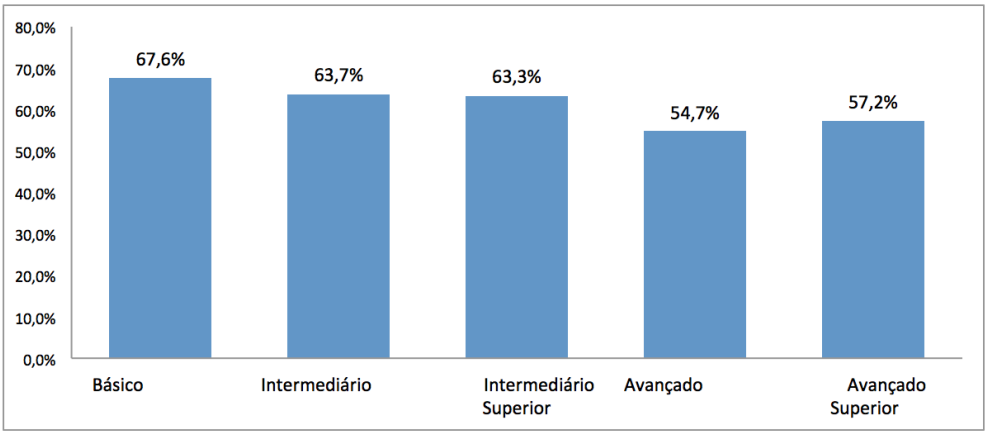


Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas nos Apêndices 1.5 e 1.6. Os percentuais correspondentes estão destacados em negrito, na diagonal das tabelas;
- N=130.

O Gráfico 6 mostra que há mais percentuais de concordância entre os avaliadores da primeira instância do que os da segunda, ainda que sejam baixos. Na segunda instância, apenas há registro de concordância (50%) quando os examinandos são avaliados no nível *Básico*. As informações constantes desse gráfico justificam o motivo de as provas terem sido encaminhadas para análise em terceira instância.

E como se apresentam os percentuais de concordância dos avaliadores nos casos em que as notas por eles atribuídas não foram discrepantes? Os resultados podem ser visualizados no Gráfico 7, a seguir.

Gráfico 7 - Edição 5: Amostra D - percentual de concordância entre avaliadores, por nível



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabulação cruzada apresentada no Apêndice 1.7. Os percentuais correspondentes estão destacados em negrito, na diagonal das tabelas;
- N=3.582.

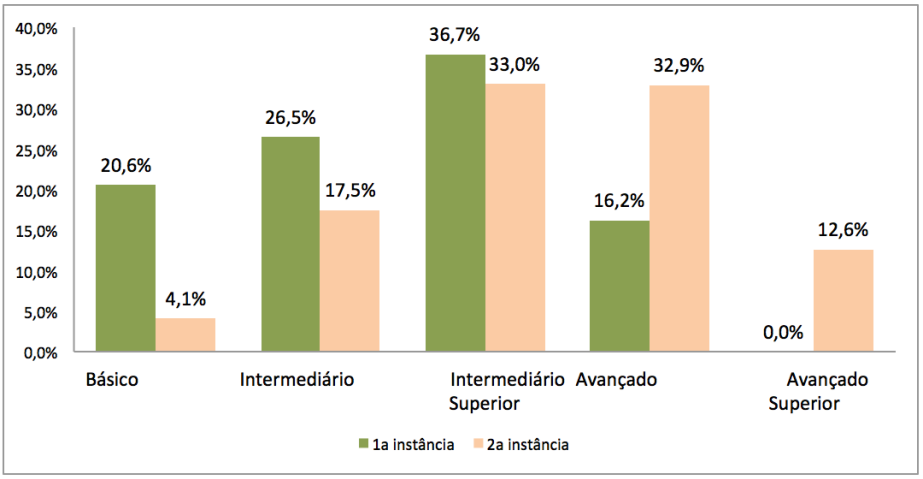
O Gráfico 7 mostra o percentual de concordância entre observador e entrevistador na avaliação em 1ª instância, nos casos em que as notas não foram consideradas discrepantes. O maior percentual é relativo ao nível *Básico*; exceto esse nível, nenhum outro atingiu 65% de concordância. Os níveis *Avançado* e *Avançado Superior* são os que apresentam menores percentuais, o que pode reforçar a hipótese de que, quanto mais proficiente é o examinando, mais nuances são apresentadas em seu desempenho, tornando mais complexa a sua avaliação. Na parte III deste capítulo, é apresentado o quanto desses percentuais é devido ao acaso.

5.2.2 Por notas finais

Tendo estabelecidas comparações entre os níveis de proficiência na visão dos avaliadores, são apresentadas, a seguir, comparações relativas às notas finais. Ressaltamos que a nota final é a média aritmética simples das notas atribuídas pelo observador e pelo entrevistador.

O Gráfico 8, a seguir, mostra os níveis de proficiência a partir das notas finais tanto da primeira quanto da segunda instâncias, no que se refere à **Amostra B**.

Gráfico 8 - Edição 5: Amostra B - níveis de proficiência com base nas notas finais da 1ª e 2ª instâncias



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabulação cruzada apresentada no Apêndice 1.8;
- N=733.

Verifica-se que, na 2ª instância, houve diminuição da quantidade de examinandos que obtiveram os níveis *Básico*, *Intermediário* e *Intermediário Superior*, sendo essa ocorrência mais acentuada no primeiro. Ou seja, muitos examinandos mudaram de nível (de menor para

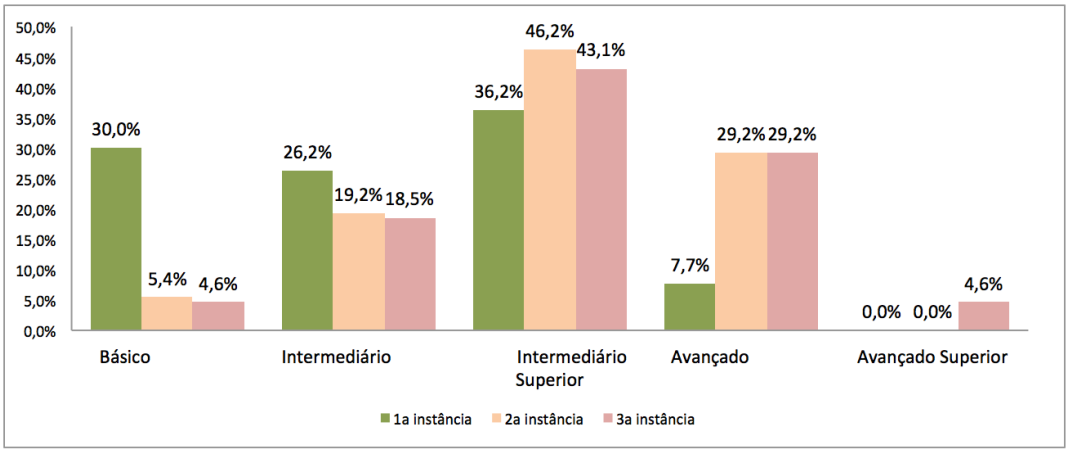
maior) na segunda avaliação, o que significa que o *recurso de ofício* considerado pelo INEP tem sido positivo para os examinandos (examinandos que não seria certificados, por exemplo, são reavaliados em níveis que permitem essa certificação). Por outro lado, significa também que os avaliadores da segunda instância têm entendimento diferenciado da grade de avaliação. Um dado deve ser levado em conta no tocante a essa diferença de entendimento: o tempo destinado à avaliação pode ser maior em segunda instância, tendo em vista que os avaliadores têm acesso à gravação da entrevista e não estão limitados aos 20 minutos de aplicação da prova, como ocorre na primeira instância.

Esses resultados não são suficientes para afirmar se a avaliação realizada em segunda instância é melhor ou pior do que a primeira, no entanto, na Parte III deste capítulo, são feitas análises sobre a consistência dessas avaliações.

Estabelecendo comparação entre os Gráficos 4 e 8, neste não há examinandos considerados do nível Avançado Superior pelos avaliadores da primeira instância, como há naquele. Isso é devido ao fato de que, ao se calcular a média simples das notas atribuídas pelos avaliadores, os examinandos não permaneceram nesse nível.

O Gráfico 9, a seguir, mostra os níveis de proficiência a partir das notas finais da primeira, segunda e terceira instâncias, no que se refere à **Amostra C**.

Gráfico 9 - Edição 5: Amostra C - níveis de proficiência com base nas notas finais da 1ª, 2ª e 3ª instâncias



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas nos Apêndices 1.9 e 1.10;
- N=130.

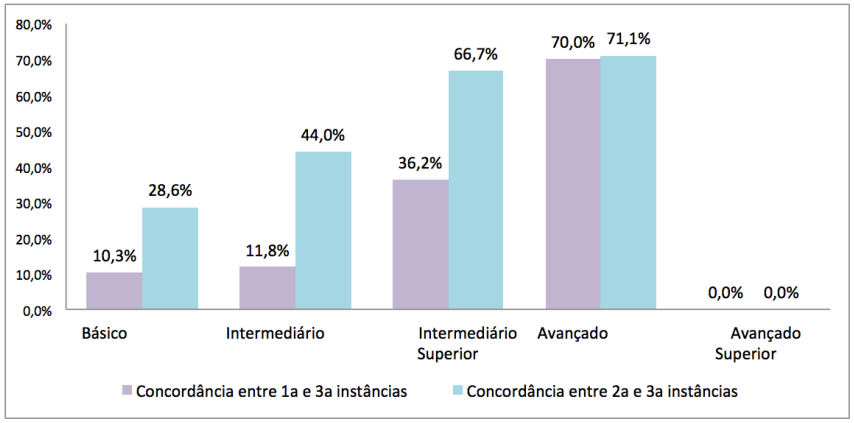
O Gráfico 9 contém informações semelhantes ao Gráfico 8, no sentido de que, tanto na segunda quanto na terceira instâncias, houve diminuição da quantidade de examinandos que obtiveram inicialmente os níveis *Básico*, *Intermediário*, sendo essa ocorrência mais acentuada no primeiro. Nota-se, também, que houve candidatos avaliados no nível *Avançado Superior*

na terceira instância, o que não ocorreu nas duas primeiras. Esse gráfico mostra que os examinandos são considerados mais proficientes na medida em que as reavaliações acontecem. Como os juízes da terceira instância utilizam apenas a grade holística para atribuírem uma única nota ao desempenho dos examinandos, estariam eles mais preocupados com a justeza da nota e também conscientes com a possível eliminação de candidatos?

O nível *Básico*, aquele que não certifica os examinandos, tende a ser mais frequente na avaliação realizada em primeira instância e esse nível é, de certa forma, diluído nas demais instâncias, fazendo com que alguns candidatos sejam certificados. Ou seja, das 130 provas que compõem a Amostra C, havia, na primeira avaliação, 30% de examinandos avaliados em nível *Básico*. Na segunda, esse percentual caiu para 5,4% e, na terceira, para 4,6%. Isso quer dizer que os 25,4% (30% - 4,6%) inicialmente de nível *Básico* foram avaliados, em terceira instância, em níveis que permitem certificação. É possível observar, com isso, como ocorreu a flutuação dos níveis de proficiência ao longo das avaliações. A partir dessa constatação, surge o seguinte questionamento: isso reflete algum problema de entendimento da grade ou do construto do exame, por parte dos avaliadores?

Além disso, surge outro questionamento com base nas instâncias de avaliação. Se o avaliador da terceira instância atribui uma nota única, sendo ela soberana às demais (da primeira e da segunda instâncias), inferimos que ele seja um sujeito com mais experiência do que a maioria dos outros avaliadores. Nesse sentido, questionamos: qual a relação (de concordância ou discordância) a nota da terceira instância apresenta com as atribuídas na primeira ou na segunda? A resposta a esse questionamento está apresentada no Gráfico 10, a seguir.

Gráfico 10 - Edição 5: Amostra C - percentual de concordância entre as 3 instâncias avaliativas



Fonte: elaborado pela autora, 2018.

Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas nos Apêndices 1.9 e 1.10. Os percentuais correspondentes estão destacados em negrito, na diagonal das tabelas;
- N=130.

O Gráfico 10 mostra os percentuais em que os mesmos examinandos foram avaliados nos mesmos níveis de proficiência: com base nas notas finais da primeira instância em relação à terceira e nas da segunda em relação à terceira.

O nível que apresenta percentual mais alto e mais equilibrado é o *Avançado*, ou seja, para avaliação dos examinandos da Amostra C, parece não ter havido muita dificuldade quanto à interpretação desse nível, mas o oposto ocorre com os demais, com destaque para o *Avançado Superior* (0% de concordância).

O referido gráfico também mostra que as notas atribuídas na terceira instância convergem mais em relação às atribuídas na segunda do que na primeira instância, em especial nos níveis *Intermediário Superior* e *Avançado*. Na Parte III deste capítulo, é mostrado o quanto desses percentuais é devido ao acaso.

Os resultados até aqui apresentados tratam de níveis de proficiência, com base na visão dos avaliadores e nas notas finais e como esses níveis foram sofrendo alterações ao longo das três instâncias de avaliação, o que revela, de certa forma, comportamentos diferentes na forma de avaliar.

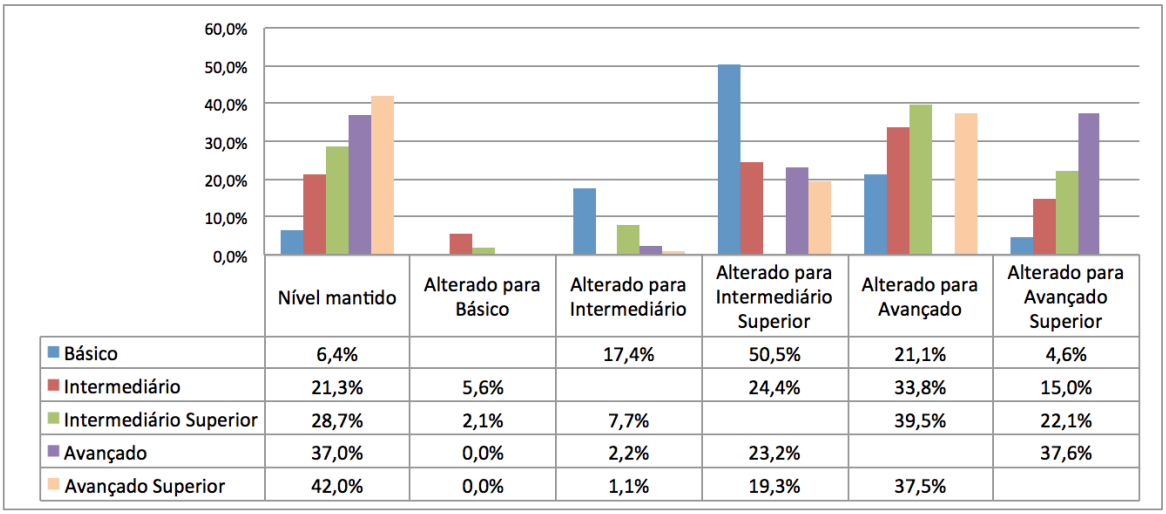
Os gráficos 4 a 10 mostraram os percentuais dos níveis de proficiência atribuídos pelos avaliadores e a partir do resultado das notas finais, bem como os percentuais de concordância das avaliações. A seção seguinte contém, além dos percentuais de concordância, os percentuais de divergência. Por exemplo: quando 60 % dos examinandos foram avaliados no nível *Básico* pelos dois avaliadores em determinada instância, os outros 40% foram avaliados em quais níveis?

5.2.3 Manutenção ou alteração de níveis

Na edição 5, a **Amostra B** permite comparar as avaliações realizadas em primeira e segunda instâncias. Os gráficos a seguir mostram o percentual em que os níveis de proficiência foram mantidos ou alterados na avaliação em segunda instância em relação à primeira, considerando: as notas atribuídas pelo observador, pelo entrevistador e as notas finais.

O Gráfico 11 trata dos níveis de proficiência atribuídos pelos **observadores** da primeira e da segunda instâncias, mostrando os percentuais de manutenção e/ou alteração desses níveis, o que permite comparar o entendimento que esses avaliadores têm dos critérios da grade analítica.

Gráfico 11 - Edição 5: Amostra B - comparação dos níveis de proficiência entre observadores (1ª e 2ª instâncias)



	Nível mantido	Alterado para Básico	Alterado para Intermediário	Alterado para Intermediário Superior	Alterado para Avançado	Alterado para Avançado Superior
Básico	6,4%		17,4%	50,5%	21,1%	4,6%
Intermediário	21,3%	5,6%		24,4%	33,8%	15,0%
Intermediário Superior	28,7%	2,1%	7,7%		39,5%	22,1%
Avançado	37,0%	0,0%	2,2%	23,2%		37,6%
Avançado Superior	42,0%	0,0%	1,1%	19,3%	37,5%	

Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabulação cruzada apresentada no Apêndice 1.11;
- N=733.

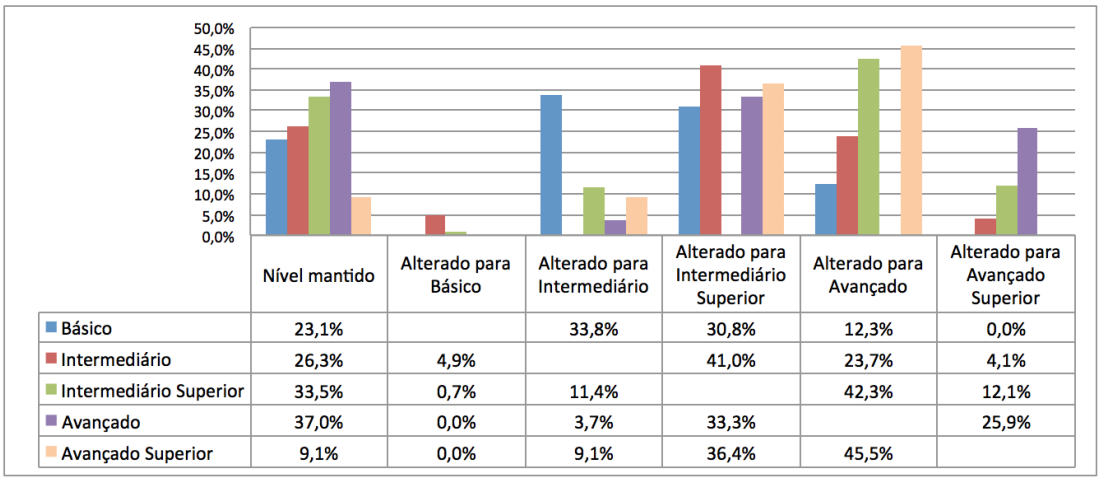
6,4% dos examinandos que não seriam certificados permaneceram no nível *Básico*. Os demais 93,6% foram reavaliados em níveis que permitem a certificação. Ressalta-se que 76,2% dos examinandos foram avaliados em dois ou mais níveis superiores ao *Básico*. Quanto aos níveis *Intermediário* e *Intermediário Superior*, verifica-se que a maioria foi alterada para níveis mais altos. No *Avançado*, houve um equilíbrio no percentual de examinandos que mantiveram-se no nível (37%) e que mudaram-se para o *Avançado Superior*

(37,6%). Já com relação ao *Avançado Superior*, a maioria dos examinandos mudou-se para níveis inferiores (*Avançado*, 37,5%, *Intermediário Superior*, 19,3% e *Intermediário*, 1,1%).

Esses dados revelam que, em geral, os examinandos são considerados mais proficientes pelos observadores da segunda do que da primeira instância de avaliação, com exceção dos inicialmente avaliados em *Avançado Superior*. Esse resultado demonstra um possível desequilíbrio avaliativo entre os observadores das duas instâncias.

Enquanto o Gráfico 11 trata das avaliações realizadas pelos observadores, o Gráfico 12, a seguir, trata das realizadas pelos **entrevistadores**, ou seja, é possível estabelecer comparação quanto ao entendimento que os entrevistadores tanto da primeira quanto da segunda instâncias têm dos critérios da grade holística de avaliação.

Gráfico 12 – Edição 5: Amostra B - comparação dos níveis de proficiência entre entrevistadores (1ª e 2ª instâncias)



	Nível mantido	Alterado para Básico	Alterado para Intermediário	Alterado para Intermediário Superior	Alterado para Avançado	Alterado para Avançado Superior
Básico	23,1%		33,8%	30,8%	12,3%	0,0%
Intermediário	26,3%	4,9%		41,0%	23,7%	4,1%
Intermediário Superior	33,5%	0,7%	11,4%		42,3%	12,1%
Avançado	37,0%	0,0%	3,7%	33,3%		25,9%
Avançado Superior	9,1%	0,0%	9,1%	36,4%	45,5%	

Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabulação cruzada apresentada no Apêndice 1.12;
- N=733.

23,1% dos examinandos que não seriam certificados permaneceram no nível *Básico* e os outros 76,9% foram avaliados em níveis que permitem a certificação. Ao contrário do que ocorreu na avaliação dos observadores (GRÁFICO 11), nenhum examinando inicialmente de nível *Básico* foi avaliado no *Avançado Superior*. Na visão dos entrevistadores, os examinandos *avançam* mais de um nível em proporções menores do que ocorreu na avaliação dos observadores.

Quanto aos níveis *Intermediário* e *Intermediário Superior*, verifica-se a mesma ocorrência da avaliação dos observadores: a maior parte (68,8% de *Intermediário* e 54,4% de *Intermediário Superior*) foi alterada para níveis mais altos. Quanto ao nível *Avançado*, 37%

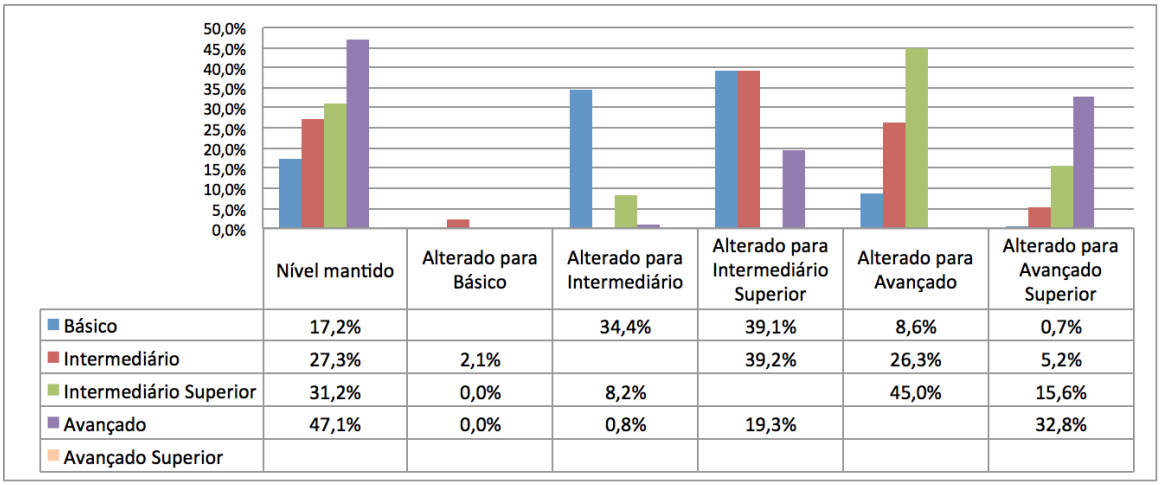
mantiveram-se no mesmo nível, a mesma proporção diminuiu de nível (37%, sendo 3,7% para *Intermediário* e 33,3% para *Intermediário Superior*) e os demais foram avaliados no nível *Avançado Superior*. No que toca a este nível, a maioria dos examinandos foi avaliada em níveis mais baixos, tendo apenas 9% mantidos em *Avançado Superior*, fato que também ocorreu na avaliação dos observadores.

Esses resultados são similares aos do Gráfico 11, pois revelam que, em geral, os examinandos são considerados mais proficientes pelos entrevistadores da segunda do que da primeira instância de avaliação, com exceção dos inicialmente avaliados em *Avançado Superior*.

De forma a estabelecer comparações entre os avaliadores (GRÁFICOS 11 e 12), no que se refere aos níveis que apresentaram menores percentuais de manutenção, verificam-se: *Básico* (6,4%), na visão dos observadores, e *Avançado Superior* (9,1%), na visão dos entrevistadores. Ou seja, não há muito equilíbrio avaliativo quando se trata dos níveis mais baixos e mais altos de proficiência adotados pelo exame.

Enquanto os Gráficos 11 e 12 mostram a manutenção e/ou alteração dos níveis de proficiência na visão dos avaliadores, o Gráfico 13, a seguir, mostra essa comparação em relação às **notas finais**, ou seja, ao resultado da média aritmética simples entre as notas atribuídas pelos avaliadores (AI e AO).

Gráfico 13 – Edição 5: Amostra B - comparação dos níveis de proficiência entre notas finais (1ª e 2ª instâncias)



	Nível mantido	Alterado para Básico	Alterado para Intermediário	Alterado para Intermediário Superior	Alterado para Avançado	Alterado para Avançado Superior
Básico	17,2%		34,4%	39,1%	8,6%	0,7%
Intermediário	27,3%	2,1%		39,2%	26,3%	5,2%
Intermediário Superior	31,2%	0,0%	8,2%		45,0%	15,6%
Avançado	47,1%	0,0%	0,8%	19,3%		32,8%
Avançado Superior						

Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabulação cruzada apresentada no Apêndice 1.13;
- N=733.

A maioria dos examinandos inicialmente avaliados nos níveis *Básico*, *Intermediário* e *Intermediário Superior* obtiveram níveis mais altos na segunda instância. Merece destaque os percentuais do nível *Básico*: 82,8% dos examinandos que, incialmente, não seriam certificados, foram classificados em níveis que permitem certificação. Trata-se, então, de um resultado preocupante, tendo em vista que a *não certificação* pode acarretar problemas de cunho pessoal e profissional aos examinandos. Isso remete ao que Bachman (1990) questiona: *o que é mais problemático: certificar um examinando que não é proficiente, ou o contrário? Ou as duas situações?* Daí a importância de se ter o processo de reanálise de provas cujas notas são discrepantes.

Quanto ao nível *Avançado*, quase a metade (47,1%) manteve-se no mesmo nível e uma parcela considerável (32,8%) foi reavaliada em *Avançado Superior*.

Constata-se, portanto, que a avaliação realizada em segunda instância tende a considerar os examinandos como mais proficientes do que a realizada em primeira, tanto considerando os avaliadores individuais (AI e AO) quanto notas finais. Ou seja, nota-se certo desequilíbrio avaliativo. Esse resultado permite que façamos o seguinte questionamento: os avaliadores da segunda instância estariam sendo mais preocupados com o impacto de resultados negativos na vida dos examinandos e, com isso, sendo mais lenientes?

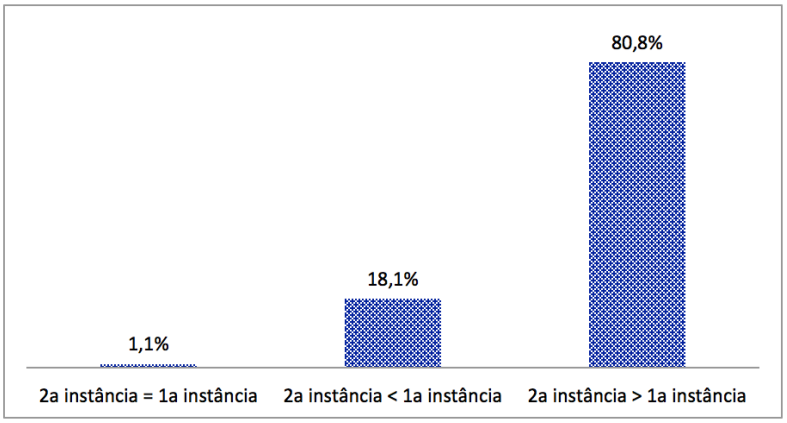
5.2.4 Grau de similitude das avaliações na 1ª e 2ª instâncias

Após terem sido feitas considerações sobre os níveis de proficiência do desempenho oral dos examinandos, esta seção vem tratar dos graus de similitude das avaliações com relação às notas (i) finais, (ii) do entrevistador, (iii) do observador e (iv) de cada um dos critérios da grade analítica. Para a exploração dos dados, foram geradas frequências das notas e, posteriormente, feita a categorização delas em:

1) 1ª instancia = 2ª instância;
2) 1ª instância > 2ª instancia;
3) 1ª instância < 2ª instancia.

Dessa forma, é possível visualizar o comportamento avaliativo. É o que mostram os Gráficos 14 a 22, a seguir.

Gráfico 14 - Edição 5: Amostra B - grau de similitude das avaliações (notas finais)

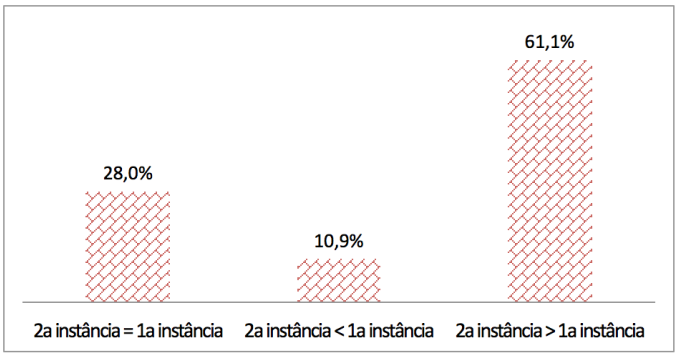


Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir da tabela do Apêndice 1.14;
- N=733.

O Gráfico 14 mostra a comparação das **notas finais**, considerando-se a Amostra B. Nota-se que na segunda instância de avaliação, 80,8% das notas foram maiores do que as atribuídas na primeira instância. Apenas em 1,1% as notas permaneceram as mesmas. Isso dialoga com os Gráficos 11 a 13, que mostram como os níveis de proficiência da primeira avaliação foram sofrendo alterações, muitas vezes para níveis superiores, na segunda instância. Dialoga, também, com a ideia de que a segunda instância trata-se de um processo que causa impacto positivo para os examinandos, tendo em vista que o desempenho oral deles pode ser reavaliado e considerado melhor, ou seja, alguns examinandos, ao terem suas provas reavaliadas, foram considerados mais proficientes.

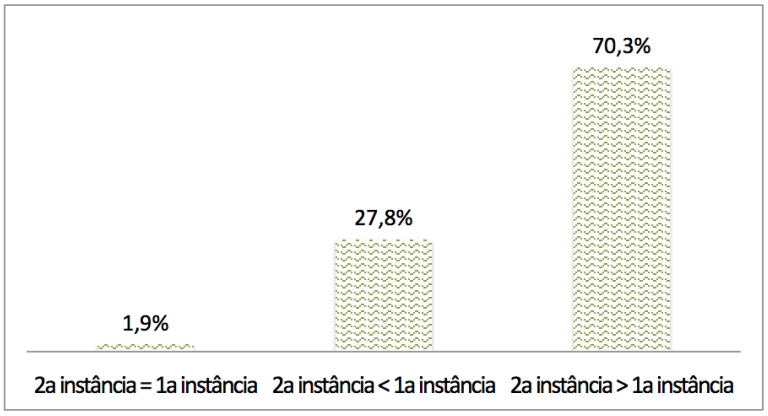
Os Gráfico 15 e 16, a seguir, mostram os graus de similitude das avaliações com base nas notas atribuídas pelos **entrevistadores** (entrevistador da 2ª instância x entrevistador da 1ª) e pelos **observadores** (observador da 2ª instância x observador da 1ª).

Gráfico 15 - Edição 5: Amostra B - grau de similitude das avaliações (entrevistadores)



Fonte: elaborados pela autora, 2018.
Notas: - gráfico elaborado a partir das tabelas dos Apêndices 1.15;
- N=733.

Gráfico 16 - Edição 5: Amostra B - grau de similitude das avaliações (observadores)



Fonte: elaborados pela autora, 2018.
Notas: - gráfico elaborado a partir das tabelas dos Apêndices 1.16;
- N=733.

Em ambos os casos, os avaliadores da segunda instância atribuíram notas mais altas ao desempenho oral dos examinandos. Quanto à manutenção da avaliação, nota-se que os observadores têm um percentual bem mais baixo (1,9%) do que os entrevistadores (28%). De forma geral, os Gráficos 15 e 16 mostram a diferença do comportamento avaliativo entre as instâncias.

No tocante à avaliação feita pelo observador, são seis os critérios considerados para se chegar a sua nota final: **compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia**. Os Gráficos 17 a 22, a seguir, mostram o comportamento dos observadores frente a cada um desses critérios.

Gráfico 17 - Edição 5: Amostra B - grau de similitude em Compreensão

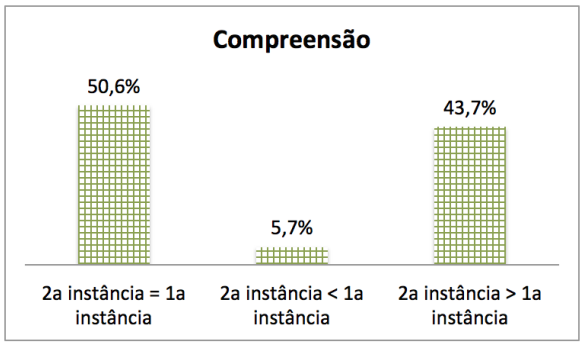


Gráfico 18 - Edição 5: Amostra B - grau de similitude em Competência Interacional

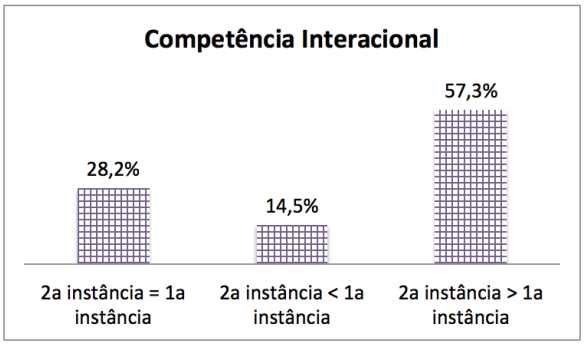


Gráfico 19 - Edição 5: Amostra B - grau de similitude em Fluência

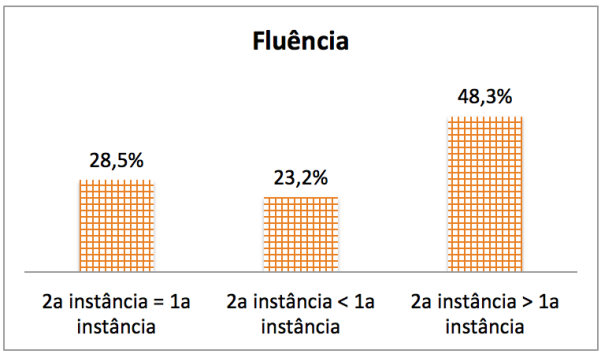


Gráfico 20 - Edição 5: Amostra B - grau de similitude em Adequação Lexical

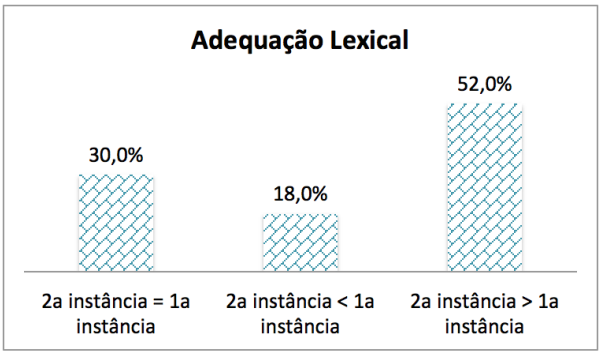


Gráfico 21 - Edição 5: Amostra B - grau de similitude em Adequação Gramatical

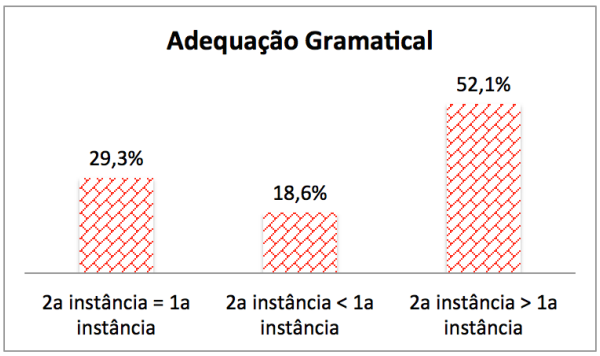
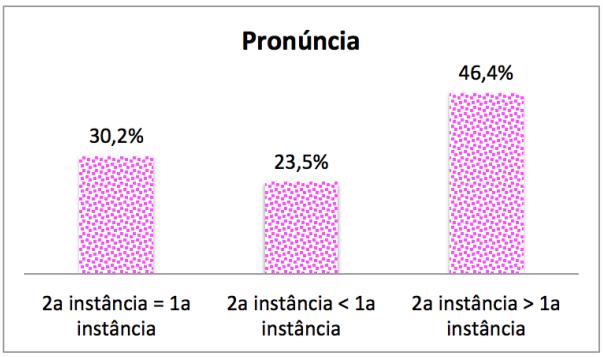


Gráfico 22 - Edição 5: Amostra B - grau de similitude em Pronúncia



Fonte: elaborados pela autora, 2018.
Notas: - gráficos elaborados a partir das tabelas do Apêndice 1.17;
- N=733, cada.

Com exceção do critério de *Compreensão*, todos os outros foram avaliados com notas mais altas na segunda instância, o que reforça o resultado de que os examinandos são considerados mais proficientes na segunda instância do que na primeira.

Compreensão é o critério que apresenta maior percentual de igualdade de avaliações nas duas instâncias, o que significa maior convergência de interpretação por parte dos observadores. O contrário ocorreu com os demais critérios (no intervalo entre 28,2% e 30,2%), o que pode indicar que os descritores da grade não estão bem detalhados.

Competência Interacional e *Fluência* são os critérios com menor grau de similitude entre os avaliadores da primeira e da segunda instâncias.

Adequação Lexical e *Adequação Gramatical* são os critérios que dividem o mesmo peso na grade (42% do total da nota do observador) e apresentam percentuais similares de convergência e divergência na avaliação. Esse equilíbrio nos percentuais nos leva ao seguinte questionamento: se o eixo léxico-gramática parece ser avaliado da mesma maneira pelos observadores, seria possível unificar esses dois critérios em um único? A resposta a esse questionamento será apresentada na Parte II, por meio de inferências estatísticas. Se, por um lado, há esse equilíbrio nos percentuais, por outro, verifica-se que, na casa dos 52%, as notas foram mais altas na segunda instância, o que significa dizer que há diferença de entendimento da grade entre os observadores das duas instâncias. Ou seja, apesar de os critérios parecerem medir da mesma forma o desempenho dos examinandos, os avaliadores das duas instâncias não chegam a um consenso quanto ao que deve ser avaliado.

Os resultados, especialmente os que dizem respeito aos critérios da grade analítica, permitem que façamos algumas considerações sobre essa grade. Como já tratado, o desempenho oral do examinando é avaliado e mensurado a partir de seis critérios, com base em uma grade holística e outra analítica (Anexos A a C). Observando como os descritores da grade analítica estão dispostos, percebe-se que todos os critérios, com exceção *fluência*, são inicialmente categorizados pelo conceito que os representa e, posteriormente, por eventuais problemas que podem ocorrer na habilidade. Significa dizer que a descrição da escala de avaliação não segue o mesmo padrão em todos os descritores. A seguir, são dados exemplos com base na nota 5, sendo que os conceitos estão destacados em itálico.

- **Compreensão**: *compreensão do fluxo natural da fala*. Rara necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala.
- **Competência interacional**: *apresenta muita desenvoltura e autonomia, contribuindo muito para o desenvolvimento da conversa*. Quando necessário, faz uso de

estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos).

- **Adequação lexical**: *vocabulário amplo e adequado* para discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Raras interferências de outras línguas.
- **Adequação gramatical**: *uso de variedade ampla de estruturas*. Raras inadequações na utilização de estruturas.
- **Pronúncia**: *pronúncia (sons, ritmo e entonação adequada).*

Do contrário, **fluência** é assim categorizado: pausas e hesitações para organização do pensamento e, eventualmente, para resolver algum problema de construção linguística, *sem interrupções no fluxo da conversa*. Ou seja, a conceituação de *fluência* apresenta-se ao final do descritor.

Diante disso, sugerimos a inversão da apresentação desse descritor, de forma que o seu conceito apareça no início, para que a escala de avaliação mantenha-se uniforme, em todos os critérios (apresentação do conceito e, posteriormente, de eventuais problemas).

Nesta Parte I deste capítulo, foi feita uma análise exploratória dos dados, iniciando pela população de estudo e depois tratando da edição 5. Inicialmente, foram explorados os níveis de proficiência oral dos examinandos, tanto com base nas avaliações feitas pelo observador e pelo entrevistador quanto com base nas notas finais. Foram apresentados os percentuais de concordância nas avaliações, como os níveis de proficiência foram sofrendo alterações ao longo das instâncias e também os graus de similitude dessas avaliações.

Em resumo, os resultados apontam para:

- a maioria dos examinandos foi avaliada, ao longo das edições, em níveis de proficiência que permitem a certificação, o que significa que eles estão se preparando bem para a prova e/ou os professores estão cada vez mais se capacitando para prepararem esse público;
- os examinandos são considerados mais proficiências à medida que são feitas as reavaliações do seu desempenho oral;
- as notas atribuídas na terceira instância de avaliação convergem mais em relação às atribuídas na segunda do que na primeira;
- muitos examinandos inicialmente avaliados no nível *Básico*, aquele que não permite certificação, foram reavaliados em níveis mais altos, o que indica que o processo de análise de discrepâncias é positivo para os examinandos;

- quanto aos critérios da grade analítica, *Compreensão* é o que apresenta maior percentual de convergência nas avaliações em primeira e segunda instâncias. *Adequação Lexical* e *Adequação Gramatical* parecem medir da mesma forma o desempenho dos examinandos e, para sanar essa dúvida, um teste de hipótese será apresentado na Parte II deste capítulo;
- os percentuais de divergência apontam para a necessidade de revisão dos descritores da grade;
- há comportamentos avaliativos diferenciados ao longo das instâncias, no que se refere à edição 5. Essa diferença de comportamento pode ocorrer por inúmeros fatores, como, por exemplo, entendimento diferenciado do construto do exame, entendimento e operacionalização da grade de avaliação ou até mesmo em função do momento de avaliação, pois, na primeira instância (no posto aplicador), não é possível ouvir a gravação da interação, como é feito nas demais instâncias. Por outro lado, os avaliadores do posto podem interpretar outros recursos da comunicação, que não o verbal. Além desses fatores, entram também em cena as próprias concepções que os avaliadores têm sobre língua, linguagem, proficiência, interação, entre outras, que devem ser alinhadas nos eventos de capacitação.

Por fim, trazemos novamente à tona o problema de pesquisa: *o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste?* A resposta a esse questionamento está na Parte III deste capítulo, em que são mostrados os resultados das estimativas da confiabilidade.

PARTE II – INFERÊNCIAS ESTATÍSTICAS

Nesta parte, são apresentadas descrições estatísticas relativas à população de estudo e à edição 5, com vistas a explorar mais o comportamento avaliativo, a partir de inferências estatísticas.

5.3 Características da população de estudo: notas finais, do observador e do entrevistador

No Apêndice 2.1.1, são apresentados: medidas de tendência central (média), de dispersão (desvio padrão, mínimo e máximo) e o coeficiente de variação de *Pearson*, que permitem comparar as notas atribuídas ao longo das edições, considerando-se a avaliação em primeira instância.

No que diz respeito às notas **mínimas** e **máximas**, os avaliadores atribuíram notas que flutuam em toda a escala: de 0 a 5, isso significa dizer que, na população de estudo, há examinandos com desempenho do mais baixo ao mais alto nível de proficiência oral, o que é algo positivo, pois todas as notas da escala servem para classificar os diferentes desempenhos.

Quanto às notas finais, verifica-se que a menor **média** foi da edição 1 (3,442) e, a maior, da edição 7 (3,714). Não há uma progressão constante das médias, o que não permite afirmar que os examinandos são mais proficientes na medida em que aconteciam as edições, até mesmo porque são examinandos diferentes, em sua grande maioria. O que pôde ser constatado, na Parte I deste capítulo, é que os examinandos estão se preparando / sendo preparados bem para o exame, pelo fato de ter havido alto percentual de níveis que permitem a certificação.

Já com relação às notas atribuídas pelo avaliadores, nota-se que as médias dos observadores são sempre maiores do que as do entrevistador. Observar as médias puramente como estão calculadas não nos dão suporte para fazer afirmação consistente, tendo em vista que, para cada critério, há o desvio padrão, uma medida da variabilidade que reflete o desvio típico da média (LEVIN, FOX, FORDE, 2012, p. 435). Portanto, para relativizar esses dados, foi calculado o **coeficiente de variação de *Pearson***, que mostra o quão volátil é a média (quanto menor o coeficiente, mais estável é a média). Os valores desse coeficiente mostram que as médias dos observadores são mais estáveis, ou seja, variam menos do que as dos entrevistadores. Se há essa variação de média e estabilidade, surge uma inquietação: as notas atribuídas pelos avaliadores deveriam ter pesos distintos na nota final da prova oral, a serem calculados a cada edição?

5.4 Características da população de estudo: critérios da grade analítica

Assim como feito com as notas finais, as atribuídas pelo observador e pelo entrevistador, tem-se, no Apêndice 2.1.2, as descrições estatísticas dos critérios da grade analítica.

Em todos os critérios, há examinandos avaliados com nota mínima (0) e máxima (5). Com relação às médias, nota-se que *Compreensão* é o que apresenta maiores médias: é o único com médias acima de 4. Isso significa que, ao longo das edições, os examinandos foram avaliados com notas altas nesse critério. É, também, o critério que apresenta maior estabilidade de média (menor coeficiente de variação de *Pearson*). Esses resultados podem ser justificáveis pelo próprio objeto analisado, a interação, ou seja, compreensão é condição *sine qua non* para que ocorra a interação entre os sujeitos e é a partir disso que as relações se estabelecem.

Adequação Lexical e *Adequação Gramatical*, por outro lado, têm as menores médias e menor estabilidade de médias (maior coeficiente de variação de Pearson).

5.5 Características da Edição 5

No Apêndice 2.2.1, é apresentada a estatística descritiva da nota do observador, do entrevistador e de cada um dos critérios analíticos, no que diz respeito à 1ª e 2ª instâncias de avaliação da **Amostra B**.

Quanto às notas dos avaliadores, as médias do observador são mais altas e mais estáveis do que as do entrevistador, o que sinaliza um comportamento avaliativo diferente. Nota-se, também, que as médias da 2ª instância são maiores e mais estáveis do que as da 1ª.

Já com relação aos critérios da grade analítica, nota-se:

- as médias do critério de *Compreensão* são maiores e mais estáveis, nas duas instâncias avaliativas;
- *Compreensão* é o único critério que tem a amplitude de notas alterada (de 0 a 5, na primeira instância, para 2 a 5, na segunda instância). Isso significa que os examinandos foram considerados mais proficientes nesse critério na 2ª instância do que na primeira;
- *Adequação Lexical* e *Adequação Gramatical* são os critérios que apresentam menores médias e menor estabilidade delas, nas duas instâncias avaliativas;

- todos os critérios, em segunda instância, apresentam médias maiores;
- as médias da segunda instância são mais estáveis do que as da primeira.

No que se refere aos resultados relativos às *Adequações Lexical* e *Gramatical*, é interessante estabelecer comparações com os apresentados na Parte I. Nos Gráficos 20 e 21, há informação de que, em 52% e 52,1% dos casos, nos critérios de *Adequação Gramatical* e *Adequação Lexical*, respectivamente, as notas da segunda instância foram mais altas do que as da primeira. No que se refere a notas mais altas na primeira, tem-se 18% para *Adequação Lexical* e 18,6% para *Adequação Gramatical*. Ou seja, o percentual de discordância entre os avaliadores das duas primeiras instâncias, para esses critérios, varia entre 70% (18% + 52%) e 70,7% (18,6% + 52,1%). Esse dado, aliado ao fato de que esses dois critérios são os que apresentam menor estabilidade, dá suporte para afirmar que eles são operacionalizados de forma diferente nas instâncias de avaliação.

Por fim, no Apêndice 2.2.2, é apresentada a estatística descritiva relativa à **Amostra C**, que permite comparar os dados das três instâncias de avaliação. Esclarecemos que, como na 3ª instância de avaliação é atribuída apenas uma nota ao desempenho oral do examinando, nota esta considerada a nota final da prova oral, a comparação que se mostra nesse apêndice leva em conta as notas finais de cada uma das instâncias.

Os dados mostram que a média das notas aumenta ao longo das instâncias de avaliação, tendo variações nos valores do desvio padrão. Como forma de relativizar os dados, o coeficiente de variação de *Pearson* indica que as médias mais estáveis são as da segunda instância de avaliação, o que confirma o já observado na Amostra B. Isso não significa que a avaliação da segunda instância seja melhor ou pior, mas apenas que varia menos.

5.5.1 Comparação entre Adequação Lexical e Adequação Gramatical

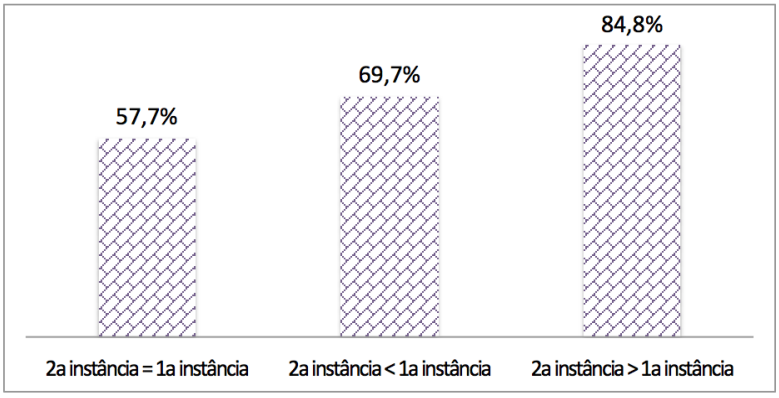
As adequações lexical e gramatical são os critérios que dividem o mesmo peso na grade analítica (42% do total da nota do observador) e apresentam algumas semelhanças. Nos resultados apontados a partir dos Gráficos 20 e 21, os dois critérios apresentaram percentuais similares de convergência (cerca de 30%) e divergência (cerca de 70%) na avaliação em primeira e segunda instâncias. Diante disso, fez-se o seguinte questionamento: *se o eixo léxico-gramática parece ser avaliado da mesma maneira pelos observadores, seria possível unificar esses dois critérios em um único?*

Nas seções 5.4 e 5.5, foi mostrado que são os critérios que apresentam as menores médias, tanto nas sete edições (seção 5.4), quanto na Edição 5, Amostra B (seção 5.5), bem

como são os que têm as médias menos estáveis (maiores coeficientes de variação de *Pearson*). Portanto, ratificamos a pertinência do questionamento.

O fato de ter havido essas semelhanças não nos permite responder a essa pergunta sem que sejam estabelecidas comparações estatísticas entre esses critérios. É preciso identificar, por exemplo, qual o percentual em que os mesmos examinandos foram avaliados da mesma maneira neles, além de realizar teste específico. Para isso, primeiramente foi feita uma tabulação cruzada, tomando como referência a Amostra B, constituída pelas 733 interações que apresentaram notas discrepantes, para verificar o percentual de concordância nas avaliações em primeira e segunda instâncias. Os resultados estão apresentados no Gráfico 23, a seguir.

Gráfico 23 - Edição 5: Amostra B: percentual de concordância entre *Adequação Lexical* e *Adequação Gramatical*



Fonte: elaborado pela autora, 2018.
Notas: - gráfico elaborado a partir das tabulações cruzadas apresentadas no Apêndice 2.4. Os percentuais correspondentes estão destacados em negrito, na diagonal da tabela;
- N=733.

Nota-se que o maior percentual (84,8%) refere-se aos casos em que os mesmos examinandos foram avaliados com nota mais alta na segunda instância no eixo léxico-gramática, o que significa que os avaliadores não compartilham o mesmo entendimento dos critérios.

Posteriormente, foi testada a normalidade das amostras (Adequação Lexical e Adequação Gramatical, Amostra B), cujos resultados (Apêndice 2.3.1) apontam que elas não seguem uma distribuição normal: em primeira instância (*KS*(Adequação Lexical – N=733)=0,194; p<0,001); (*KS*(Adequação Gramatical – N=733)=0,177; p<0,001); em segunda instância (*KS*(Adequação Lexical – N=733)=0,207; p<0,001); (*KS*(Adequação Gramatical – N=733)=0,199; p<0,001).

Por fim, foi realizado um teste para verificar se as medianas dos dois critérios são iguais. O resultado (Apêndice 2.4) mostra que as medianas apresentam diferenças estatisticamente significativas ($\chi^2(4)=508{,}378$; $p<0{,}001$), o que significa dizer que eles não podem formar um único descritor na grade de avaliação analítica, justificando, assim, a necessidade de serem avaliados separadamente.

5.5.2 Comparação dos critérios da grade analítica

Interessa-nos, também, saber se a mediana dos critérios da grade analítica pode ser, estatisticamente, considerada igual nas duas instâncias. Primeiramente, foi testada a normalidade das amostras (critérios, por instância), cujos resultados, apresentados no Apêndice 2.3, mostram que elas não seguem uma distribuição normal:

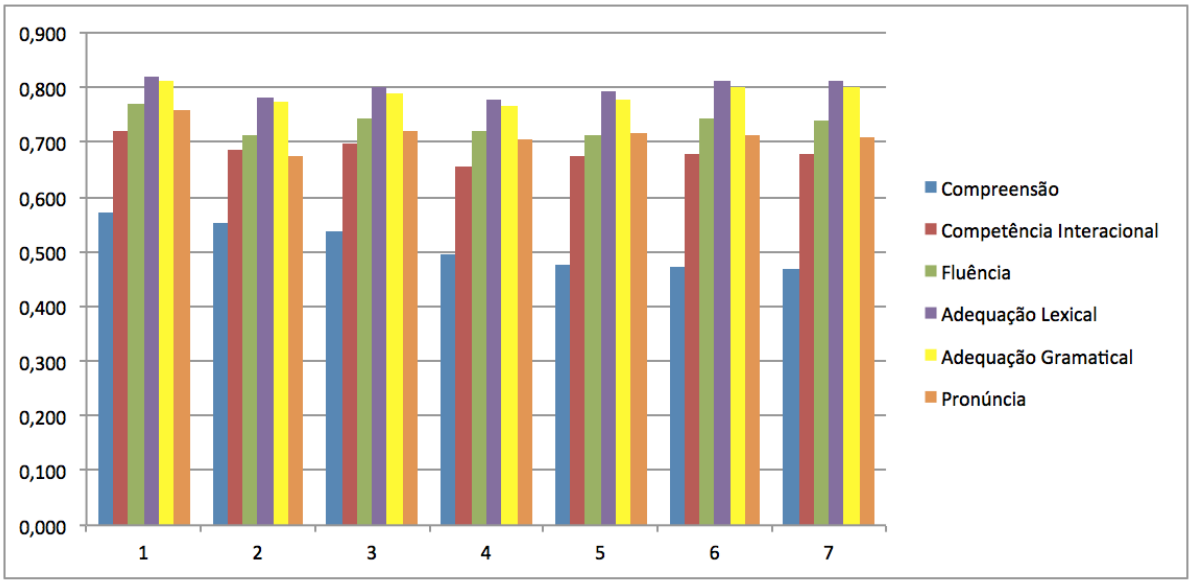
- em primeira instância: (*KS*(Compreensão – N=733)=0,294; p<001); (*KS*(Competência Interacional – N=733)=0,187; p<0,001); (*KS*(Fluência – N=733)=0,164; p<0,001); (*KS*(Adequação Lexical – N=733)=0,194; p<0,001); (*KS*(Adequação Gramatical – N=733)=0,177; p<0,001); (*KS*(Pronúncia – N=733)=0,175; p<0,001);
- em segunda instância: (*KS*(Compreensão – N=733)=0,494; p<001); (*KS*(Competência Interacional – N=733)=0,248; p<0,001); (*KS*(Fluência – N=733)=0,193; p<0,001); (*KS*(Adequação Lexical – N=733)=0,207; p<0,001); (*KS*(Adequação Gramatical – N=733)=0,199; p<0,001); (*KS*(Pronúncia – N=733)=0,197; p<0,001).

Em segundo lugar, foi analisado, par a par, cada critério de avaliação da grade analítica, levando-se em conta as instâncias de avaliação (ex: compreensão na 1ª instância x compreensão na 2ª instância, e assim por diante). Os resultados do teste (Apêndice 2.5) mostram que as medianas dos critérios apresentam diferenças estatisticamente significativas: (Wilcoxon – N=733 (Compreensão=14,186; p<0,001); (Competência Interacional=13,865; p<0,001); (Fluência=8,430; p<0,001); (Adequação Lexical=11,293; p<0,001); (Adequação Gramatical=11,688; p<0,001) e (Pronúncia=8,359; p<0,001)). Isso significa que o comportamento dos avaliadores, nas duas instâncias, é diferente, seja pela concepção de avaliação que cada um carrega, pelo entendimento que eles têm dos critérios da grade, pelas nuances que são apresentadas nas interações face a face ou pela dificuldade (ou facilidade) na operacionalização da grade etc.

5.6 Correlações entre as notas do observador e do entrevistador

No Apêndice 2.6, são apresentadas as correlações entre as notas do observador (cada um dos critérios analíticos) e do entrevistador, por edição, considerando-se a avaliação feita em primeira instância. O cálculo dessas correlações verifica se os critérios da grade analítica produzem resultados consistentes e se são concorrentes com a nota do entrevistador, ou seja, se a prova tem capacidade de avaliar o construto que objetiva avaliar: proficiência oral. Todas as correlações foram positivas e os resultados estão resumidos no Gráfico 24, a seguir.

Gráfico 24 - Correlação entre as notas do observador e do entrevistador - população de estudo



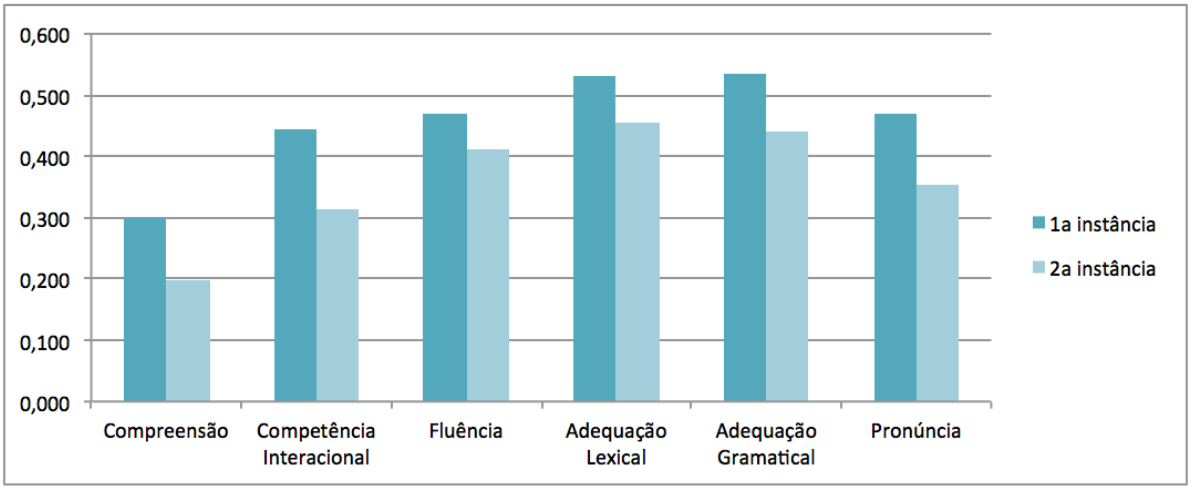
Fonte: elaborado pela autora, 2018.
Nota: gráfico elaborado a partir das correlações apresentadas no Apêndice 2.6.

Ao longo das sete edições analisadas, nota-se que o critério *Compreensão* é o que apresenta menor correlação com a nota do entrevistador. Por outro lado, os que apresentam maior correlação são *Adequação Lexical* e *Adequação Gramatical*, resultado que dialoga com os apresentados por Ferreira (2018). Dito de outra maneira, esses são os critérios que mais explicam a nota do entrevistador e, aquele, o que menos explica.

É possível constatar, também, que o comportamento na atribuição de notas (pelo valor das correlações) apresenta-se equilibrado de uma edição para outra, o que significa que, ao longo dos anos, os critérios de correção são entendidos da mesma forma. Já que há esse equilíbrio, é preciso que os *stakeholders*, ao fazerem alterações nas grades, as apresentem com clareza para todos os envolvidos no processo de aplicação e avaliação.

Como forma de comparar as duas primeiras instâncias avaliativas, há, no Apêndice 2.7, as correlações da Edição 5, no que toca à Amostra B, ou seja, as notas atribuídas em primeira e segunda instâncias. Os resultados estão resumidos no Gráfico 25, a seguir.

Gráfico 25 - Correlação entre as notas do observador e do entrevistador - Edição 5, Amostra B



Fonte: elaborado pela autora, 2018.
Nota: gráfico elaborado a partir das correlações apresentadas no Apêndice 2.7.

Nota-se que, na primeira instância, o critério que apresenta a menor correlação com a nota do entrevistador é *Compreensão* e, os maiores, *Adequação Lexical* e *Adequação Gramatical*, assim como já apontado no Gráfico 24. Na segunda instância, os critérios também mantêm essa ordem.

Um dado relevante mostrado nesse gráfico é que, em segunda instância, todas as correlações são menores do que em primeira. Isso significa dizer que, mesmo apresentando discrepância significativa, as notas atribuídas pelo observador em primeira instância são mais correlacionadas com a nota do entrevistador do que as atribuídas em segunda instância. Dito de outra forma, a instância responsável por reanalisar as provas cujas notas apresentam problemas (discrepância significativa) na primeira avaliação é a que contém menor índice de correlação das notas entre observador e entrevistador.

5.7 Análise das discrepâncias

Conforme apresentado no Capítulo 2, as provas orais são avaliadas no posto aplicador (1ª instância) e, existindo caso significativo de discrepância, elas são reavaliadas em outras instâncias, até que se chegue à nota final do desempenho do examinando. Essa reavaliação

pode ocorrer pela discrepância existente tanto em relação às notas atribuídas pelos avaliadores da prova oral, quanto em relação às notas entre as partes escrita e oral.

Ressaltamos que as notas discrepantes são uma variável importante para a análise de qualquer teste de proficiência. A partir delas, constroem-se as amostras, por instância de avaliação ou outra variável, e comparam-se os comportamentos avaliativos nas diversas instâncias. Além disso, são um objeto que sinaliza a existência de possíveis problemas interpretativos da grade e, consequentemente, do que o exame se propõe a medir.

Levando em consideração que apenas a Edição 5 contém as avaliações nas três instâncias, as análises relativas às notas discrepantes são dessa edição.

A tabela a seguir contém as quantidades e os tipos de discrepâncias geradas/analisadas, revelando, assim, como elas são tratadas neste trabalho.

Tabela 1 - Quantificação das discrepâncias significativas - Edição 5

Discrepâncias geradas no(s) processo(s) de avaliação						
1. Discrepâncias de notas atribuídas por AO e AI (≥ 1,50)	**Localidade dos postos**				**Total**	
	Brasil		**Exterior**			
	Quant.	%	Quant.	%	Quant.	%
Base de cálculo: p r o v a s o r a i s a p l i c a d a s	1.761	38,41%	2.824	61,59%	4.585	
1.1 Discrepâncias geradas na 1ª instância (posto aplicador)	114	6,47%	117	4,14%	231	5,04%
1.1.1 Casos compatibilizados	114	100,00%	116	99,15%	230	99,57%
1.1.1.1 Discrepâncias geradas na 2ª instância (compatibilização)	22	19,30%	20	17,24%	42	18,26%
1.1.1.1.1 Casos analisados na 3ª instância (nota de consenso)	22	100,00%	20	100,00%	42	100,00%

2. Discrepâncias de notas atribuídas por AO e AI (< 1,50) que foram compatibilizadas	**Localidade dos postos**				**Total**	
	Brasil		**Exterior**			
	Quant.	%	Quant.	%	Quant.	%
Base de cálculo: p r o v a s o r a i s a p l i c a d a s	1.761	38,41%	2.824	61,59%	4.585	
2.1 Discrepâncias de < 1,5 ponto geradas na 1ª instância (posto aplicador) e que foram compatibilizadas	143	8,12%	360	12,75%	503	10,97%
2.1.1 Discrepâncias geradas na 2ª instância (compatibilização)	27	18,88%	61	16,94%	88	17,50%
2.1.1.1 Casos analisados na 3ª instância (nota de consenso)	27	100,00%	61	100,00%	88	100,00%

Resumo	
Total de provas aplicadas / avaliadas na 1ª instância	4.585
Total de provas avaliadas na 2ª instância (itens 1.1.1 + 2.1)	733
Total de provas avaliadas na 3ª instância (itens 1.1.1.1.1 + 2.1.1.1)	130
Total de provas com nota discrepante aplicadas no Brasil (itens 1.1 + 2.1)	257
Total de provas com nota discrepante aplicadas no exterior (itens 1.1 + 2.1)	477
Total de provas com notas discrepantes (Brasil + exterior)	734
Total de provas com notas discrepantes reavaliadas (itens 1.1.1+2.1)	733

Legenda: **AO**: Avaliador Observador **AI**: Avaliador Interlocutor

Fonte: elaborado pela autora, 2018.

A primeira parte da Tabela 1 mostra a quantidade de discrepâncias, cujo valor tenha sido igual ou maior que 1,50 ponto, comparando-se as notas atribuídas pelos avaliadores da prova oral (AO e AI). Esse tipo de discrepância representa 5,04% do total de provas aplicadas.

Estabelecendo uma comparação entre a localidade dos postos, nota-se que o percentual das discrepâncias **(≥ 1,50)** geradas na 1ª instância foi maior nas provas aplicadas no Brasil (6,47%) do que nas aplicadas no exterior (4,14%). Isso também é observado em relação às discrepâncias geradas na 2ª, sendo 19,30% e 17,24%, para Brasil e exterior, respectivamente. Ou seja, parece haver mais ruídos nas interações realizadas em postos brasileiros. Esse resultado nos leva à seguinte reflexão: o fator de os avaliadores estarem no exterior faz com que eles se sintam *guardiões* da língua portuguesa e, com isso, são mais rigorosos na aplicação dos critérios avaliativos? Para discutir sobre essa hipótese, seria necessário ouvir algumas gravações de interações e analisar as notas atribuídas pelos avaliadores, de forma que seja possível estabelecer relação entre a gestão da interlocução com as notas, como fez Brown (2005).

Já a segunda parte da tabela trata de provas cuja diferença das notas atribuídas pelos avaliadores tenha sido menor que 1,50 ponto, o que representa 10,97% do total de provas aplicadas. Inicialmente, conforme apontado no Capítulo 2, os casos de diferença de nota menor que 1,50 não são considerados discrepantes. Inferimos, então, que os 503 casos aqui apresentados referem-se à discrepância entre as partes escrita e oral.

Comparando-se a localidade dos postos, nota-se que o percentual de discrepâncias **(< 1,50)** geradas na primeira instância foi maior nas provas aplicadas no exterior (12,75%) do que nas aplicadas no Brasil (8,12%), sendo um resultado inverso do observado na primeira parte da tabela. Já com relação às discrepâncias geradas na segunda instância, o percentual continuou sendo maior nas provas aplicadas no Brasil (18,88%) do que nas aplicadas no exterior (16,94%).

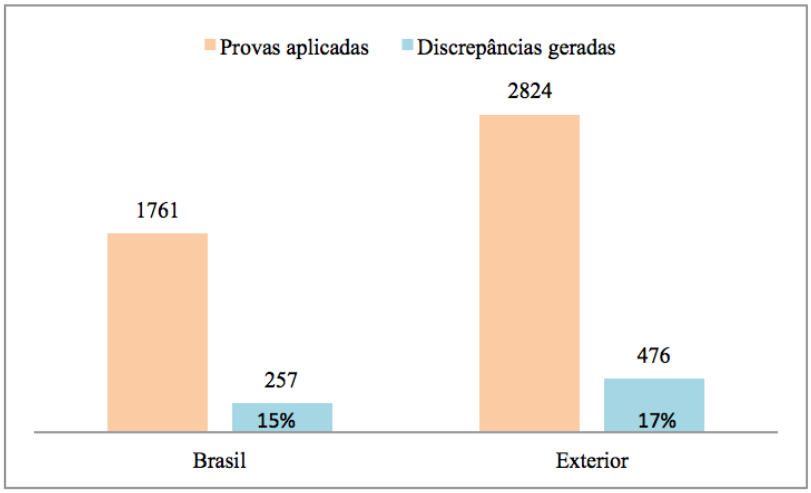
Os dois tipos de discrepância (**<1,50** e **≥ 1,50**) totalizam, então, 16,01% das provas aplicadas na Edição 5.

Comparando-se as instâncias avaliativas, nota-se que a 2ª contém mais avaliações com notas discrepantes do que a 1ª. Das discrepâncias de **≥ 1,50**, 5,04% foram gerados na 1ª instância e, 18,26%, na 2ª. Das discrepâncias **<1,50**, 10,97% foram gerados na 1ª instância e, 17,50%, na 2ª. Ou seja, a instância responsável por resolver os problemas de divergência avaliativa gerados na 1ª é a que provoca maiores reflexões no que diz respeito à consistência da avaliação. Esse dado, aliado à constatação da existência de comportamento diferenciado

dos avaliadores da 2ª instância (atribuem notas mais altas e com médias mais estáveis), fazem com que surja um novo questionamento: quais nuances da interação e da grade avaliativa fazem com que o comportamento avaliativo seja distinto de uma instância para a outra, a ponto de serem geradas mais discrepâncias entre os avaliadores da segunda? Esse, portanto, configura-se um importante objeto de investigação, não contemplado nesta tese.

Do total de discrepâncias geradas, 734 (**<1,50** e **≥ 1,50**), sendo 257 de provas aplicadas no Brasil e 477 de provas aplicadas no exterior, apenas 1 não foi reavaliada. Nesse caso específico, o examinando obteve o nível *Avançado* (nota 4,17), na visão do AO, e nível *Intermediário* (nota 2) na visão do AI. Gerou-se, então, uma discrepância de 2,17, o que deveria ter sido, à primeira vista, levada para a avaliação em segunda instância. As provas que foram reavaliadas devido à existência de discrepância significativa somam, portanto, 733 e estão distribuídas conforme mostra o Gráfico 26, a seguir.

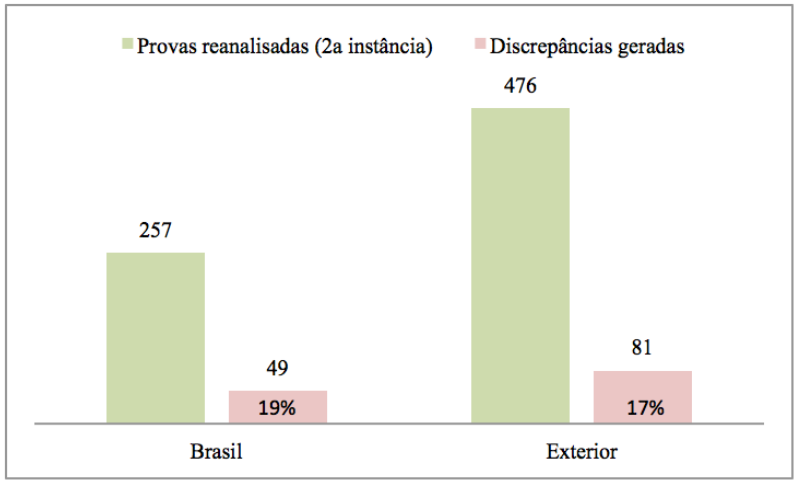
Gráfico 26 - Discrepâncias geradas na 1ª instância *versus* posto aplicador



Fonte: elaborado pela autora, 2018.

Nota-se que o percentual de discrepâncias é maior nas provas aplicadas no exterior (17%) do que nas aplicadas no Brasil (15%), ou seja, parece que há mais ruídos, na visão dos avaliadores da primeira instância, nas interações dos examinandos que fizeram prova no exterior. Dessas 733 provas que foram reavaliadas, em 130 houve novamente discrepância de 1,50 ponto ou mais e, por isso, foram reavaliadas em terceira instância. É o que mostra o Gráfico 27, a seguir.

Gráfico 27 - Discrepâncias geradas na 2ª instância *versus* posto aplicador



Fonte: elaborado pela autora, 2018.

Das 130 discrepâncias geradas na segunda instância de avaliação, 19% referem-se a provas aplicadas no Brasil e 17% a provas aplicadas no exterior, ou seja, parece haver mais ruídos, na visão dos avaliadores da segunda instância, nas interações dos examinandos que fizeram prova no Brasil.

A seguir, é apresentado um teste de hipótese sobre o contexto de aplicação das provas do exame (Brasil e exterior) e sobre as discrepâncias.

5.7.1 Testes de hipótese das discrepâncias

Na Tabela 1, consta que, das 4585 provas orais aplicadas (38,41% no Brasil e 61,59% no exterior), foram geradas e analisadas 16,01% de discrepâncias significativas. Das discrepâncias ≥ 1,50 ponto, 6,47% foram de provas aplicadas no Brasil e 4,14% de provas aplicadas no exterior. Já com relação às discrepâncias < 1,50 ponto, 8,12% foram de provas aplicadas no Brasil e 12,75% de provas aplicadas no exterior. Portanto, há diferenças entre os contextos de aplicação *Brasil* e *exterior* e, diante disso, fizemos um teste de hipótese para verificar se essas diferenças são estatisticamente significativas.

As amostras não têm distribuição normal ($KS(230)=0,343$; $p<0,05$; $KS(503)=0,451$; $p<0,05$), conforme Apêndice 2.3.2, e os resultados do teste de hipótese estão apresentados nos Apêndices 2.8.1 e 2.8.2.

Tanto com relação às discrepâncias de ≥ 1,50 quanto às <1,50, os resultados dos testes ((χ^2(1)=12,743; p<0,001) e ((χ^2(1)=23,780; p<0,001), respectivamente, mostram que existem diferenças estatisticamente significativas entre os grupos *Brasil* e *Exterior*. É plausível, portanto, a preocupação de pesquisadores, como Coura-Sobrinho (2014) e Costa (2015), em relação a esses contextos, ou seja, os contextos de aplicação das provas carregam características que devem ser analisadas para, inclusive, serem exploradas nos eventos de capacitação.

Nesta Parte II do capítulo, foram apresentadas as características estatísticas do teste, iniciando pela população de estudo e depois tratando da edição 5. Inicialmente, uma estatística descritiva mostrou algumas medidas, como média, desvio padrão, valores mínimos e máximos e o coeficiente de variação de *Pearson*, tanto das notas finais, do observador e do entrevistador quanto dos critérios da grade analítica.

Posteriormente, dadas algumas semelhanças entre os critérios *Adequação Lexical* e *Adequação Gramatical*, foi realizado um teste para verificar se eles podem formar um único descritor, como forma de responder ao questionamento que se apresenta desde a Parte I.

Por fim, foram comparadas as medianas dos critérios da grade analítica, da edição 5, e analisados os casos de notas discrepantes.

Em resumo, os resultados apontam para:

- as médias das notas do observador são sempre maiores e mais estáveis do que as do entrevistador;
- as médias do critério *Compreensão* são sempre maiores e mais estáveis;
- as médias de todos os critérios avaliados na segunda instância são maiores e mais estáveis do que as da primeira, em se tratando da Amostra B da edição 5;
- as médias dos critérios *Adequação Lexical* e *Adequação Gramatical* são sempre menores e menos estáveis. Se o coeficiente de variação de *Pearson* é o resultado do desvio padrão dividido pela média, tem-se que *Adequação Lexical* e *Adequação Gramatical* são os critérios mais dispersos, resultado esse também apresentado no item 5.4. Quanto mais disperso o dado, menor a precisão da medida e, consequentemente, mais difícil de se fazer inferências a respeito dele;
- comparações estabelecidas pela Amostra C da Edição 5 mostram que a média das notas aumenta ao longo das instâncias e as da segunda são mais estáveis do que as das demais;

- um teste mostrou um alto percentual em que os mesmos examinandos foram avaliados com notas mais altas nos critérios *Adequação Lexical* e *Adequação Gramatical*, na segunda instâncias da Amostra B da Edição 5, o que sinaliza para um entendimento diferenciado dos descritores da grade, por parte dos observadores. Além disso, devido à existência de diferenças estatisticamente significativas entre as medianas desses critérios, eles não podem formar um único descritor;
- outro teste apontou para diferenças estatisticamente significativas entre as medianas dos critérios da grade analítica, considerando-se a primeira e a segunda instâncias de avaliação da Amostra B da Edição 5, ou seja, constata-se variabilidade de comportamento avaliativo entre as instâncias;
- os coeficientes de correlação mostram que *Compreensão* é o critério que menos se correlaciona com a nota do entrevistador. O contrário ocorre com *Adequação Lexical* e *Adequação Gramatical*. Isso mostra que o eixo léxico-gramática é o que mais pesa na avaliação tanto do observador quanto do entrevistador;
- os coeficientes de correlação da avaliação em segunda instância são menores do que da avaliação em primeira. Isso revela que os avaliadores responsáveis por resolver os problemas de notas gerados na primeira instância atribuem notas menos correlacionadas;
- se o critério *Compreensão* parece não ser eficiente para diferenciar um candidato ruim de um muito bom (é o critério que possui maiores médias e que menos se correlaciona com a nota do entrevistador), ele poderia ser excluído da grade? A resposta a essa pergunta está na Parte III deste capítulo, em que são mostrados os valores do coeficiente *Alfa de Cronbach*;
- no que se refere às discrepâncias, uma análise mostrou que há diferenças estatisticamente significativas entre os contextos de aplicação das provas: Brasil e exterior;
- os avaliadores da segunda instância são responsáveis por gerar maior percentual de discrepância (do tipo $\geq 1{,}50$), do que os da primeira.

PARTE III - ESTIMATIVA DA CONFIABILIDADE

Esta parte do Capítulo V apresenta (i) uma análise preliminar ao estudo da confiabilidade, mostrando a dimensionalidade da escala avaliativa, por meio da Análise dos Componentes Principais (ACP); (ii) os índices de confiabilidade, por meio do cálculo do coeficiente *Alfa de Cronbach*, e (iii) o nível de concordância entre os avaliadores, por meio do cálculo do coeficiente *Kappa*.

5.8 Estrutura fatorial dos itens da escala: grade de avaliação analítica

No Apêndice 3.1, são apresentados os resultados da ACP relativa às notas atribuídas pelo observador a cada um dos critérios da grade analítica, por edição, ou seja, têm-se os resultados da **população de estudo**. Os seguintes resultados indicam a aplicabilidade da técnica de Análise dos Componentes Principais: valores do teste de esfericidade de Bartlett ($p<0,001$) mostram que as variáveis estão correlacionadas significativamente e do Kaiser-Meyer-Olkin – KMO (0,882 – 0,897, considerando-se as sete edições) indicam que a recomendação face à ACP varia de *boa* a *excelente*.

No que se refere às comunalidades, ou seja, ao peso fatorial, nota-se que as maiores estão entre *Adequação Lexical* e *Fluência* e entre *Adequação Lexical* e *Adequação Gramatical*, portanto essas são as variáveis que têm maior poder de explicação da nota final do observador. Os menores valores da comunalidade, entretanto, são de *Compreensão*, que variam entre 0,478 a 0,639. Verificam-se dois valores abaixo de 0,5, que é o mínimo aceitável: 0,478 (edição 6) e 0,492 (edição 7). Portanto, *Compreensão* é o critério que tem o menor poder de explicação da nota do observador, como também apontado por Ferreira (2018).

O método de extração *Análise do Componente Principal*, em todas as edições, mostra resultados que apontam para a unidimensionalidade da escala, ou seja, foi extraído apenas um componente, o que significa dizer que apenas uma dimensão está sendo avaliada. Os critérios avaliados pelo observador em primeira instância, portanto, avaliam uma única dimensão, ou seja, um único construto: proficiência oral. Esse resultado corrobora com o apresentado por Ferreira (2018), de que os dados analisados revelam uma escala unidimensional.

Para cada edição analisada, foi extraído apenas um componente (composto pelos seis critérios da grade analítica), sendo que ele explica entre 71,864% (edição 6) e 79,916% (edição 2) da variância total, ou seja, contém a maior parte da informação do construto que

está sendo mensurado. Outro dado relevante é que, na escala de seis itens (os seis critérios), três são responsáveis por explicar mais de 90% da variância total.

No Apêndice 3.2, por sua vez, são apresentados os resultados da ACP relativos à **Edição 5**, Amostra B, ou seja, estabelece-se uma comparação entre a avaliação realizada em primeira e segunda instâncias. Vimos anteriormente que, ao longo das edições, na avaliação em primeira instância, a escala mostra-se unidimensional. *E em segunda instância, isso se mantém?*

Os seguintes resultados indicam a aplicabilidade da técnica de Análise dos Componentes Principais: valores do teste de esfericidade de Bartlett ($p<0,001$, para ambas as instâncias) indicam que as variáveis estão correlacionadas significativamente e os do KMO (0,878 e 0,831, na primeira e segunda instâncias, respectivamente) indicam que a recomendação face à ACP é *boa*.

No que se refere às comunalidades (pesos fatoriais), nota-se que as maiores estão entre *Adequação Lexical* e *Adequação Gramatical*, portanto essas são as variáveis que têm maior poder de explicação da nota final do observador. Os menores valores da comunalidade, entretanto, são de *Compreensão*. Os valores das comunalidades em segunda instância são maiores do que em primeira, com exceção do critério *Fluência*.

Em primeira instância, o método de extração *Análise do Componente Principal* aponta para a unidimensionalidade da escala. O componente extraído explica 74,094% da variância total; três são responsáveis por explicar mais de 90%, como também apontado no resultado anterior (da população de estudo).

No entanto, em segunda instância, os resultados apontam para uma escala bidimensional, ou seja, foram extraídos dois componentes, o que significa que há dois construtos sendo avaliados. Após a rotação, um dos componentes é responsável por explicar 48,075% da variância total e, o outro, 31,345%, ou seja, os dois juntos explicam 79,421% da variância total. Os componentes extraídos (método de extração: Análise de Componente Principal; método de rotação: Varimax com normalização de Kaiser) são:

Componente 1	**Componente 2**
Fluência	Compreensão
Adequação Lexical	Competência Interacional
Adequação Gramatical	
Pronúncia	

Nota-se com clareza a diferença tanto entre as instâncias (uma uni e a outra bidimensional) quanto com relação aos eixos avaliativos propostos pela grade (eixo 1: compreensão, competência interacional e fluência, que são responsáveis por 50% da nota total do observador; eixo 2: adequação lexical e adequação gramatical, responsáveis por 42% da nota, e eixo 3: pronúncia, responsável por 8% da nota).

Esse resultado dialoga com os apresentados nas Partes I e II deste capítulo, em que foi apontado comportamento avaliativo diferenciado ao longo das instâncias avaliativas. Portanto, a constatação que fica e que gera uma nova hipótese de pesquisa é: **a dimensionalidade da escala varia na medida em que varia o comportamento avaliativo**. Trata-se de uma nova hipótese porque o banco de dados analisado contém a segunda instância de avaliação apenas em uma edição, o que não permite estabelecer comparações dos resultados ao longo do tempo.

A alteração na dimensionalidade da escala é, portanto, o primeiro indício de que o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados de um teste.

Outro dado relevante a ser destacado é que, na Edição 5, Amostra B, segunda instância (Apêndice 3.2.2), na matriz de correlações, foi verificada correlação quase nula entre: Compreensão e Adequação Lexical (0,332), Adequação Gramatical (0,262) e Pronúncia (0,195); e entre: Competência Interacional e Pronúncia (0,299), ou seja, esses critérios (compreensão e as adequações lexical e gramatical; compreensão e pronúncia; competência interacional e pronúncia) estão pouco correlacionados entre si. Já na primeira instância (Apêndice 3.2.1), essa *quase nulidade* não foi verificada, o que mostra, mais uma vez, divergência no comportamento avaliativo nas duas primeiras instâncias.

5.9 Índices de confiabilidade dos resultados do exame Celpe-Bras: grade analítica

Tendo sido avaliada a dimensionalidade da escala, esta seção trata de verificar o quão confiáveis são os resultados da avaliação feita por meio dessa escala, em se tratando das notas atribuídas pelo observador a cada um dos critérios da grade analítica.

No que se refere à **população de estudo**, os índices de confiabilidade são os constantes do Apêndice 3.3. Os valores do coeficiente *Alfa de Cronbach*, distribuídos entre 0,921 e 0,948, indicam confiabilidade elevada em todas as edições. Isso significa que os itens

da escala estão suficientemente correlacionados entre si, ou seja, possuem elevada consistência interna.

No referido apêndice, nas tabelas relativas às "estatísticas de item total", na última coluna, há a informação de como ficaria o valor do coeficiente *Alfa* caso fosse retirado algum dos critérios em análise. Apenas o critério *Compreensão*, se retirado, promoveria uma ligeira melhora no *Alfa*, mas ainda assim este permaneceria como "confiabilidade elevada". Se, por exemplo, o valor inicial do *Alfa* estivesse como "moderado" e fosse alterado para "elevado" caso retirado esse critério, justificaria retirá-lo da análise e gerar novo teste de confiabilidade. Ou seja, no *corpus* analisado, excluir o critério *Compreensão* não faz com que melhore significativamente a consistência interna dos itens da escala, o que justifica mantê-lo na grade.

Esse resultado responde negativamente à pergunta deixada na Parte II deste capítulo, qual seja: *se o critério Compreensão parece não ser eficiente para diferenciar um candidato ruim de um muito bom (é o critério que possui maiores médias e que menos se correlaciona com a nota do entrevistador), ele poderia ser excluído da grade?* Portanto, embora pareça ser um critério com pouca eficiência na distinção entre os examinandos, ele é peça importante na escala para avaliação da proficiência oral. Se, por um lado, tem-se um critério pouco eficiente, mas, por outro, há a necessidade de sua manutenção na grade, talvez seja necessário um ajuste nos pesos dos critérios, como propõe Ferreira (2018). Nas normas vigentes, *Compreensão* é responsável, juntamente com *Competência Interacional* e *Fluência*, por 50% da nota da avaliação feita pelo observador.

No Apêndice 3.4, por sua vez, constam os índices de confiabilidade relativos à **Edição 5**, Amostra B, sendo possível comparar a primeira e a segunda instâncias.

Em primeira instância, o coeficiente *Alfa* tem valor de 0,929, indicando confiabilidade elevada. Já em segunda instância, o coeficiente *Alfa* tem valor de 0,875, indicando confiabilidade moderada, ou seja, a avaliação em primeira instância apresenta índice maior de confiabilidade dos resultados, em se tratando das notas atribuídas pelo observador. Dito de outra maneira, a mudança de comportamento avaliativo é responsável por alteração nos índices de confiabilidade.

Considerando que, na seção anterior, foram apresentados resultados que apontam para a bidimensionalidade da escala quando se trata da segunda instância avaliativa, da Edição 5, foi necessário verificar a consistência interna dos itens de cada um dos componentes extraídos. É o que consta do Apêndice 3.5.

Para o Componente 1 (Fluência, Adequação Lexical, Adequação Gramatical, e Pronúncia), o valor do *coeficiente* Alfa é de 0,901, o que indica confiabilidade elevada; já para o Componente 2 (Compreensão e Competência Interacional), o valor é de 0,543, o que indica confiabilidade inaceitável. Esses resultados apontam para uma não aplicabilidade de dois componentes na escala, tendo em vista que um deles não possui índice de confiabilidade aceitável.

Parece um pouco contraditório dizer que a instância responsável por resolver os problemas de notas discrepantes é a que mostra menor índice de confiabilidade dos resultados. Esse é, então, o segundo indício de que o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados de um teste.

5.10 Níveis de concordância entre os avaliadores

Depois de verificada a dimensionalidade da escala e analisados os índices de confiabilidade, resta averiguar o percentual de concordância entre os avaliadores, levando em consideração o quanto dessa concordância é devida ao acaso.

Tendo em vista que as notas atribuídas pelo observador não são arredondadas, como as do entrevistador, todas elas foram transformadas em níveis, de acordo com estabelecido no Quadro 3, no Capítulo II, para possibilitar o cálculo dos graus de concordância por nível de proficiência.

No que tange à população de estudo, os resultados estão resumidos na tabela a seguir.

Tabela 2 - Valores do coeficiente *Kappa*: população de estudo

Edição	Valor do *Kappa*
1	0,446
2	0,403
3	0,453
4	0,442
5	0,414
6	0,448
7	0,414

Fonte: elaborada pela autora, 2018.
Nota: tabela elaborada a partir do Apêndice 3.6.

Como se observa, todas as sete edições apresentam valor *satisfatório* do coeficiente *Kappa*, ainda que baixo, quando são comparados os níveis de proficiência atribuídos pelo observador e pelo entrevistador.

Quanto à Edição 5, Amostra B, os resultados estão resumidos na tabela seguinte.

Tabela 3 - Valores do coeficiente *Kappa*: Edição 5 - Amostra B

Instância avaliativa	Valor do Kappa
1ª	0,166
2ª	0,133

Fonte: elaborada pela autora, 2018.
Nota: tabela elaborada a partir do Apêndice 3.7.

Não se observa valor aceitável para o coeficiente de *Kappa* para nenhuma das instâncias, sendo os dois valores considerados *pobres*. A ocorrência de um *Kappa* baixo na primeira instância da Amostra B é de certa forma justificável, pois é nela que se situam todas as notas consideradas discrepantes. Ou seja, se há discrepância significativa, é de se esperar um nível de concordância baixo.

Por outro lado, é de se esperar maior concordância na segunda instância, pois é ela a responsável por resolver os problemas de discrepância. Mas os resultados mostram que é na segunda instância que se situa o menor valor de concordância. Esse é, portanto, o terceiro indício de que o comportamento avaliativo pode ser considerado um fator responsável por erro de mensuração que interfere na confiabilidade dos resultados do teste.

Considerando-se que os avaliadores da segunda instância são mais experientes do que os da primeira, esse resultado nos leva a refletir sobre uma conclusão a que chega Barnwell (1986) em sua pesquisa, de que um treinamento mínimo dos avaliadores é capaz de promover níveis altos de concordância entre eles, na maioria dos casos. Nos dados analisados nesta tese, isso não foi confirmado.

A Amostra B é composta pelas provas cujas notas foram consideradas discrepantes. E *como é o nível de concordância dos avaliadores da Amostra D, aquela em que constam as provas sem notas discrepantes?* O Apêndice 3.7.3 mostra o resultado de valor do coeficiente *Kappa* a 0,450, ou seja, *satisfatório*. Significa que, mesmo quando não há discrepância significativa entre as notas atribuídas pelo observador e pelo entrevistador, o valor do coeficiente é similar aos apresentados na Tabela 2, ou seja, não é muito maior. Esse resultado sinaliza para a necessidade de se intensificar estudos relativos aos níveis de proficiência, como, por exemplo, detalhar melhor os descritores das grades de avaliação, mostrando a

diferença entre “alguns problemas”, “muitos problemas” e “problemas sérios”; entre “contribuindo pouco para o desenvolvimento da conversa” e “raramente contribuindo”; entre “poucas interferências de outras línguas” e “algumas interferências de outras línguas”.

Em resumo, os resultados apontam para:

- no que se refere à Análise dos Componentes Principais:
    - as variáveis com maior poder de explicação da nota final do observador são: *Adequação Lexical*, *Adequação Gramatical* e *Fluência*;
    - a variável com menor poder de explicação é *Compreensão*. Em duas edições, 6 e 7, *Compreensão* apresentou valor de comunalidade (peso fatorial) abaixo do aceitável
    - na avaliação realizada em primeira instância, tanto em relação à população de estudo quanto à Edição 5, Amostra B, foi constatada a unidimensionalidade da escala;
    - na avaliação realizada em segunda instância (Edição 5, Amostra B), foi constatada a bidimensionalidade.
- no que se refere aos índices de confiabilidade (coeficiente *Alfa de Cronbach*):
    - na população de estudo, foram verificados que os itens da grade analítica possuem elevada consistência interna, indicando confiabilidade elevada em todas as edições;
    - uma análise acerca do critério *Compreensão* revela que, se ele fosse excluído da grade de avaliação, não promoveria uma melhora significativa na consistência interna dos itens, ou seja, o critério revela-se peça importante na avaliação da proficiência oral;
    - na Edição 5, Amostra B, foi verificada confiabilidade elevada, na avaliação realizada em primeira instância, e confiabilidade moderada, na avaliação da segunda instância, ou seja, a avaliação em primeira instância revela resultados mais confiáveis;
    - considerando que a Análise dos Componentes Principais verificou a existência de dois componentes na segunda instância (Amostra B, edição 5), ou seja, escala bidimensional, foi verificada a consistência interna de cada um deles: o primeiro (fluência, adequação lexical, adequação gramatical e pronúncia) revela confiabilidade elevada, mas o segundo (compreensão e competência interacional), por outro lado, revela confiabilidade inaceitável;

- no que se refere aos níveis de concordância entre os avaliadores (Coeficiente *Kappa*):
    - nas sete edições, foram verificados valores do coeficiente *satisfatórios*, ainda que baixos;
    - na Edição 5, na Amostra B, a segunda instância avaliativa revelou valor de coeficiente menor (e *pobre*, inclusive) do que a primeira instância, ou seja, a instância responsável por dirimir os problemas de notas discrepantes demonstra menos concordância entre seus avaliadores do que a própria instância geradora das discrepâncias. Trata-se de um resultado preocupante, pois, se o Inep adota o recurso de ofício, na certeza de que as reavaliações dirimirão os problemas avaliativos advindos da primeira instância, a segunda deveria apresentar níveis de concordância aceitáveis;
    - na Edição 5, a Amostra D também revela valor *satisfatório* de coeficiente, mas ainda assim, baixo. Isso significa dizer que, mesmo quando não há divergência significativa na avaliação realizada pelo observador e pelo entrevistador, não se observam grandes concordâncias entre eles, o que indica a necessidade de revisão dos procedimentos para identificação de discrepâncias.

PARTE IV – DISCUSSÃO DOS RESULTADOS

Os resultados apresentados ao longo deste capítulo tratam do processo de mensuração que inclui três traços distintos, na visão de Bachman (1990): o desempenho oral dos examinandos foi *quantificado*, por meio da atribuição de notas, observando-se *características*, que dizem respeito aos descritores Compreensão, Competência Interacional, Fluência, Adequação Lexical, Adequação Gramatical e Pronúncia. E tudo é feito tendo como base *regras e procedimentos*, que dizem respeito aos fatores que devem nortear o processo de quantificação que, no caso da prova oral do exame Celpe-Bras, refere-se às grades de avaliação, que devem refletir o construto do exame.

Nesta pesquisa, os avaliadores são considerados uma variável de suma importância no processo de mensuração e que deve ser analisada, devido à subjetividade que se instaura na avaliação da proficiência oral. Daí a necessidade de abordar o comportamento avaliativo, entendido, nesta pesquisa, como a maneira como as notas são atribuídas pelos diferentes avaliadores e nas diferentes instâncias de avaliação, materializando-se na observação da consistência *inter-rater*, nos termos de Bachman (1990) e Moskal e Leydens (2000).

O comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste? Para responder à pergunta de pesquisa, o objetivo geral foi analisar a maneira pela qual a confiabilidade tem a ver com o comportamento avaliativo, a partir de dados de sete edições consecutivas do exame. Para isso, foi adotada uma metodologia quantitativa de análise de dados, via *software* SPSS, que permitiu a visualização do processo de mensuração do exame de uma maneira a considerar os diferentes atores e etapas envolvidos.

Os resultados das análises foram apresentados por partes, considerando a especificidade de cada um deles. Na Parte I, foi feita uma análise exploratória dos dados, com o objetivo de mostrar um panorama do que compõe o *corpus* da pesquisa e de materializar a conceituação dada a *comportamento avaliativo*. Na Parte II, foram mostradas as características estatísticas do teste, no que diz respeito às medidas de tendência central e dispersão, bem como resultados de testes de hipótese. A Parte III, por fim, foi dedicada à estimativa de confiabilidade, em que, primeiramente, foi feita uma análise preliminar ao estudo da confiabilidade, por meio da utilização da técnica de Análise dos Componentes Principais; posteriormente, foram verificados os índices de consistência interna, via coeficiente *Alfa de Cronbach*, e os níveis de concordância dos avaliadores, via coeficiente *Kappa*.

No que tange ao desempenho final dos examinandos na prova oral do exame, verifica-se que, nas sete edições analisadas, a grande maioria foi avaliada em níveis que permitem certificação, tendo maior ocorrência no nível *Avançado*. Algumas inferências podem ser feitas a partir desse resultado, tais como: (i) a maior divulgação do exame e dos seus procedimentos de avaliação pode proporcionar aos candidatos preparação mais eficaz; (ii) os candidatos estão se preparando mais para a prova; (iii) a crescente procura pelos cursos de capacitação docente para atuar na área de PLE pode proporcionar melhor capacitação aos candidatos e (iv) a melhoria no ensino de PLE, devido ao efeito retroativo do próprio exame.

Com relação à avaliação realizada pelo observador e pelo entrevistador, foi verificado um baixo percentual de concordância, conforme consta do Gráfico 3, que mostra os percentuais em que os mesmos examinandos foram classificados nos mesmos níveis, considerando-se a população de estudo. O nível que apresenta maior percentual de concordância é o *Básico*, ou seja, os avaliadores demonstram maior convergência na avaliação quando os examinandos não são proficientes. Os maiores percentuais de divergência estão concentrados nos níveis *Avançado* e *Avançado Superior*, o que indica que, quanto mais proficiente é o examinando, mais difícil é avaliá-lo.

No que se refere à Edição 5, Amostra B, os resultados do Gráfico 5 mostram que os avaliadores concordam mais na primeira instância nos níveis *Básico* e *Intermediário*; em segunda instância, nos níveis *Intermediário Superior* e *Avançado*. Ou seja, os avaliadores da primeira instância concordam mais entre si quando os examinandos são menos proficientes. Dois outros resultados merecem destaque: (i) em todos os níveis de proficiência, os percentuais de concordância, nas duas instâncias, não atingem 63%; (ii) com relação à segunda instância, os avaliadores são, acredita-se, os mais experientes do quadro de colaboradores, mas os percentuais de concordância entre eles não atingem 46% em nenhum dos níveis de proficiência.

Esses resultados apontam para a existência de variabilidade de comportamento avaliativo, o que pode ser diminuído se os descritores forem mais bem detalhados na grade. Essa variabilidade também é constatada quando são comparados os níveis de proficiência atribuídos aos examinandos pelos avaliadores da primeira e segunda instâncias, da Edição 5: na Amostra B, os examinandos são considerados mais proficientes na segunda instância do que na primeira; na Amostra C, eles são mais proficientes na terceira instância. Ou seja, na medida em que são reavaliados, os examinandos são considerados mais proficientes.

A Amostra C também revelou dois outros resultados relevantes: (i) quanto ao percentual em que os mesmos examinandos foram avaliados nos mesmos níveis pelo

observador e pelo entrevistador, o Gráfico 6 mostrou que há mais registros de concordância na primeira instância, ainda que em percentuais baixos; (ii) na segunda instância, por sua vez, há registro de concordância apenas no nível *Básico*, o que mostra que os descritores das grades devem ser mais bem definidos. Esses resultados justificam a necessidade de as provas terem sido encaminhadas para avaliação em terceira instância.

Essa necessidade, portanto, pode indicar que as 130 interações apresentaram maior ruído na comunicação e talvez isso seja um dos fatores para a baixa ou nenhuma concordância entre os avaliadores. Ou seja, as provas que compõem a Amostra C revelam ser um importante objeto de futuras pesquisas para identificar as causas desse ruído.

Com base nessa Amostra C, também foi possível identificar os níveis de concordância nas três instâncias avaliativas. Inferimos que o avaliador da terceira instância seja mais experiente do que os demais, tendo em vista que a nota atribuída por ele é soberana a todas as anteriores. Dessa forma, foi verificado com qual instância os avaliadores da terceira convergem mais na avaliação. Os resultados do Gráfico 10 apontam que há maior concordância com a segunda instância, entretanto não há registro de concordância no nível *Avançado Superior* com nenhuma instância. Para esse nível, portanto, há total discordância de interpretação do descritor da grade, por parte dos avaliadores da terceira instância em relação à primeira e à segunda, nas 130 interações que compõem a Amostra C. Os resultados refletem uma necessidade de se intensificar reflexões acerca dos critérios de avaliação nos eventos de capacitação.

Ainda com relação aos percentuais de concordância de avaliação, na Amostra D, aquela em que não contém notas com discrepância significativa, nenhum nível atingiu 65% de concordância, com exceção do *Básico*. A partir desse resultado, surge uma inquietação: se não foram apresentados percentuais altos de concordância entre os avaliadores, mesmo nas interações sem discrepância significativa, acreditamos que os valores para se considerar discrepância ($\geq$1,50, entre avaliadores, ou >1,50, entre as partes escrita e oral) e as amplitudes dos níveis de proficiência (de 0 a 1,99 = Básico; de 2,00 a 2,75 = Intermediário, e assim por diante) deveriam ser revistos e reformulados.

Nas reavaliações que são feitas das provas cujas notas apresentaram discrepância significativa, muitos examinandos inicialmente avaliados no nível *Básico*, aquele que não permite certificação, foram reavaliados em níveis mais altos, o que indica que o processo de análise de discrepâncias tem impacto positivo para os examinandos. Pelo Gráfico 13, tem-se

que uma parcela representativa (39%) foi reavaliada no nível Intermediário Superior. Se tomarmos como base, por exemplo, os médicos estrangeiros que precisavam[40] passar pelo exame para conseguirem a revalidação de seu diploma no Brasil, esse dado é de grande relevância para tais profissionais. Esse resultado, portanto, nos remete ao questionamento feito por Bachman (1990): *o que é mais problemático: certificar um examinando que não é proficiente, ou o contrário? Ou as duas situações?* Isso nos permite refletir sobre o impacto que os resultados de um teste têm sobre a vida das pessoas que a ele se submetem, daí a necessidade de terem resultados confiáveis e válidos.

Uma interpretação diferenciada dos descritores da grade pode levar a atribuição diferenciada de notas, interferência no resultado do teste e, consequentemente, na confiabilidade. Ou seja: quando dois avaliadores avaliam de forma diferente o desempenho de um mesmo sujeito, com base em um mesmo construto e a partir de grades de avaliação responsáveis por minimizar certas subjetividades avaliativas, essa variabilidade avaliativa pode gerar escores não confiáveis. Portanto, além de interferir negativamente na confiabilidade dos resultados do exame, pode ter consequências importantes na vida dos sujeitos que a ele se submetem.

Uma análise do grau de similitude das avaliações (Gráficos 14 a 16) também revelou que as avaliações em segunda instância consideram os examinandos como mais proficientes do que em primeira, o que reflete uma interpretação diferenciada dos descritores da grade.

Passando a tratar dos critérios que compõem a grade analítica do Celpe-Bras, os Gráficos 17 a 22 mostram que os critérios *Competência Interacional* e *Fluência* foram os que apresentaram menores percentuais (28,2 e 28,5%, respectivamente) de convergência de interpretação entre os avaliadores da primeira e segunda instâncias, o que pode sinalizar para uma necessidade de melhor detalhamento desses descritores na grade de avaliação. Quanto à *Competência Interacional,* o resultado corrobora, de certa maneira, com o apontado por Niederauer (2014), ao relatar uma pesquisa exploratória realizada por Santos e Niederauer (2004 apud Niederauer, 2014). Ao analisar a grade de avaliação analítica do exame Celpe-Bras, a autora (2014) destaca que o desempenho na Competência Interacional é avaliado em termos de: (a) desenvoltura e autonomia; (b) contribuição para o desenvolvimento da conversa; (c) uso de respostas breves e (d) uso de estratégias para resolver problemas lexicais, gramaticais e/ou fonológicos, sendo essas estratégias o foco do artigo e entendidas

[40] Embora esteja suspensa a obrigatoriedade da certificação para os médicos, ela perdurou até janeiro de 2016.

por “estratégias comunicativas”, que são descritas na grade como *reformulação*, *paráfrase* e *correções*, como se pode verificar nos Anexos B e C.

A pesquisa desenvolvida por Santos e Niederauer (2004 apud Niederauer, 2014) teve como objetivo investigar a relação entre Estratégia Comunicativa (EC) e nível de proficiência. Para isso, foram analisadas 20 entrevistas orais, como simulação do Celpe-Bras, de alunos de um programa de ensino e pesquisa em português para estrangeiros. O uso de EC foi analisado a partir da variável nível de proficiência, comparando as provas orais de 5 alunos de cada um dos níveis Avançado II, Avançado I, Intermediário e Iniciante II[41]. Para análise e classificação das EC, as pesquisadoras utilizaram a taxonomia de Dörnyei e Scott (1997 apud Niederauer, 2014), que abrange 33 estratégias comunicativas.

Um dos resultados que a pesquisa apontou foi que, quanto menor o nível de proficiência, maior a diversidade de estratégias a que os examinandos lançam mão durante a interação, ou seja, o que caracteriza o nível de proficiência em língua estrangeira não é o uso de EC, mas o “não uso” de determinadas EC. Consequentemente, quanto maior o nível de proficiência, menor foi a necessidade de se utilizar EC.

As EC são recursos a que os falantes recorrem para resolver problemas lexicais, gramaticais e/ou fonológicos durante uma interação, ou seja, quanto maior o nível de proficiência do falante, menos problemas ele terá e, consequentemente, menor a sua necessidade de utilização de EC. Considerando isso e os resultados que a pesquisa apontou, as pesquisadoras propuseram uma inversão na granulação desse descritor na grade de avaliação analítica do Celpe-Bras. Ou seja, quanto menor o nível de proficiência, maior a necessidade de uso de EC e não o contrário, como previsto na grade.

Por fim, as pesquisadoras consideram que as diferentes funções que cada EC tem na interação e seu potencial para indicar o nível de proficiência em língua adicional permitem entender que a forma como essas estratégias são operacionalizadas na grade de avaliação analítica do Celpe-Bras não fazem com que se aproveite seu potencial avaliativo.

Talvez o (não) entendimento dessas *estratégias comunicativas* por parte dos avaliadores, ou a falta de melhor detalhamento da grade do que se entende por *interação*, sejam o motivo de o descritor *Competência Interacional*, nesta tese, ter sido o que apresentou menor percentual de convergência de interpretação. Isso nos leva a afirmar que ele precisa ser reformulado, assim como mais discutido nos eventos de capacitação.

---

[41] Níveis utilizados no curso de PLE de que trata a pesquisa de Niederauer (2014).

O critério *Compreensão*, por outro lado, é o que apresenta maior percentual (50,6%) de convergência nas avaliações em primeira e segunda instâncias.

Adequação Lexical e *Adequação Gramatical* apresentam percentuais similares de convergência e divergência na avaliação (são os que apresentam maior percentual de divergência) e parecem, com isso, medir da mesma forma o desempenho dos examinandos. Dado isso, foi colocado o seguinte questionamento: *se o eixo léxico-gramática parece ser avaliado da mesma maneira pelos observadores, seria possível unificar esses dois critérios em um único?* A resposta é apresentada mais adiante.

De uma forma geral, os percentuais de divergência apontam para a necessidade de revisão dos descritores da grade. Os resultados até aqui discutidos mostram evidências empíricas da existência de variabilidade de comportamento avaliativo. Isso pode ocorrer por inúmeros fatores, como, por exemplo, entendimento diferenciado do construto do exame, entendimento e operacionalização da grade de avaliação, a própria grade, ou até mesmo em função do momento de avaliação, pois, na primeira instância (no posto aplicador), não é possível ouvir a gravação da interação, como é feito nas demais instâncias. Por outro lado, os avaliadores do posto podem interpretar outros recursos da comunicação além do verbal. Além desses fatores, entram também em cena as próprias concepções que os avaliadores têm sobre língua, linguagem, proficiência, interação, entre outras, que devem ser alinhadas nos eventos de capacitação.

Essas evidências empíricas também foram mostradas por meio das características estatísticas do teste. Em resumo, os resultados apontam que:

- as médias das notas do observador são sempre maiores e mais estáveis do que as do entrevistador;
- as médias do critério *Compreensão* são sempre maiores e mais estáveis. Parece que os examinandos estão demonstrando um alto nível de compreensão do que é proposto na interação face a face. Por outro lado, podemos afirmar que a *Compreensão* é um critério que subjaz ao próprio contrato de comunicação, ou seja, se não houver compreensão, o contrato está fadado ao fracasso e, consequentemente, não há interlocução. Ao analisar as médias altas desse critério, é possível inferir que ele não é o responsável por discriminar um candidato fraco de um mediano ou forte no que se refere à proficiência oral;
- na avaliação em segunda instância, as médias de todos os critérios avaliados são maiores e mais estáveis do que na primeira, em se tratando da Amostra B da edição 5;
- a média das notas aumenta ao longo das instâncias e as da segunda são mais estáveis do que as das demais, em se tratando da Amostra C da Edição 5;

- *Adequação Lexical* e *Adequação Gramatical*, na Amostra B da Edição 5, apresentam sempre menores médias e menos estáveis, ou seja, menor estabilidade indica que são os critérios mais dispersos. Quanto mais disperso o dado, menor a precisão da medida e, consequentemente, mais difícil de se fazer inferências a respeito dele. Há duas possibilidades de análise: ou são os avaliadores que variam muito na atribuição de notas ou a própria população que apresenta essa alta variabilidade de proficiência em léxico e gramática. Isso, aliado ao fato de que são critérios com alto percentual de divergência de interpretações, sinaliza que eles devem ser mais bem explicitados na grade de avaliação e ser objeto de estudo e discussão nos eventos de capacitação dos avaliadores;
- outro resultado relevante de *Adequação Lexical* e *Adequação Gramatical* é o de que há um alto percentual (84,8%) em que os mesmos examinandos foram avaliados com notas maiores na segunda instância, no que se refere à Amostra B da Edição 5. Significa dizer que os observadores da primeira e segunda instâncias demonstram ter visão heterogênea desses critérios e, consequentemente, variabilidade do comportamento avaliativo. Dada a existência de diferenças estatisticamente significativas entre as medianas desses critérios, eles não podem ser unificados, o que justifica a avaliação ser feita separadamente.

Espera-se que os itens de uma escala de avaliação sejam correlacionados suficientemente para medirem determinado construto. Considerando que a nota final do desempenho oral do examinando é a média aritmética simples das notas atribuídas pelo entrevistador e observador, foram verificadas as correlações entre elas (os critérios do observador x a nota do entrevistador). Essa correlação, apresentada nos Gráficos 24 e 25, mostra se as notas do observador são concorrentes com a do entrevistador. *Compreensão* é o critério que menos se correlaciona com a nota do entrevistador. Esse resultado é de certa forma justificável: as médias das notas desse critério são sempre mais altas e mais estáveis; a compreensão é o ponto inicial para que qualquer interlocução tenha sucesso e, portanto, para que um examinando possa demonstrar habilidades nos demais critérios, ele precisa, primeiramente, compreender o conteúdo informacional do Elemento Provocador e o que está sendo dito na interação. Portanto, não parece ser um critério eficiente para diferenciar um candidato ruim de um bom. Então, *ele poderia ser excluído da grade de avaliação?* Uma análise do coeficiente *Alfa de Cronbach* indica que, se excluído da grade, não promove melhora significativa na confiabilidade, ou seja, o critério que deve ser mantido.

Adequação Lexical e *Adequação Gramatical*, por outro lado, mostram maiores correlações com a nota do entrevistador. Isso mostra que o eixo léxico-gramática é o que mais pesa na avaliação tanto do observador quanto do entrevistador. Esse resultado pode

levar a discussões teóricas no campo do ensino de línguas. Fala-se muito, atualmente, num ensino voltado mais para a interação e menos para o foco na forma, a depender dos objetivos específicos de cada curso. Embora haja perspectivas de ensino que defendam isso, os resultados da avaliação oral do exame apontam para uma maior cobrança de habilidades em léxico e gramática. O resultado dessa maior valorização desses critérios dialoga com a pesquisa de Schoffen (2009), ao analisar a parte escrita do exame Celpe-Bras. Segundo a pesquisadora, embora as tarefas e as grades de avaliação apresentem os aspectos constituintes do gênero do discurso, os pontos de corte entre os níveis de proficiência são definidos pela recuperação de informações do texto-base e as adequações lexical e gramatical.

Outro resultado relevante é o de que há mais correlação entre as notas atribuídas em primeira instância do que em segunda, ou seja, os avaliadores responsáveis por resolver os problemas das discrepâncias atribuem notas menos correlacionadas.

Passando a comentar sobre as discrepâncias e, levando em consideração as informações mostradas na Tabela 1, podemos estabelecer um paralelo entre o que dizem os Editais sobre a análise de provas com notas discrepantes, conforme mostrado no Capítulo 3, e o que de fato foi realizado no processo de avaliação. É previsto que, na avaliação das interações face a face, todas as vezes que houver discrepância **de ≥ 1,50** entre as notas atribuídas pelos dois avaliadores, essas interações serão reavaliadas por outra instância. Os dados nos mostraram que, das 231 discrepâncias dessa natureza (114 de provas aplicadas no Brasil e 117 de provas aplicadas no exterior), foram reavaliadas 230.

Permanecendo discrepância de **≥ 1,50** no processo de reavaliação (2ª instância: compatibilização), as interações devem ser novamente avaliadas em 3ª instância (nota de consenso). Da reavaliação dessas 230 interações, foram geradas 44 discrepâncias (22 de provas aplicadas no Brasil e 20 de provas aplicadas no exterior), sendo que todas elas foram reavaliadas em 3ª instância.

Além das discrepâncias de **≥ 1,50**, o sistema registrou outras (503 casos, com diferença de nota **< 1,50**) que, ao nosso ver, referem-se à discrepância existente entre as partes escrita e oral. Esse tipo de discrepância é um indicativo de um desequilíbrio de habilidades escrita e oral do examinando, o que foi objeto de pesquisa de Costa (2015); entretanto, tendo em vista que não termos informações sobre as notas da prova escrita, não é possível estabelecermos comparações relativas ao desempenho global do examinando. Mas o fato é que, das 503 interações que foram reavaliadas em segunda instância devido a discrepâncias dessa natureza, as notas atribuídas geraram 88 novas discrepâncias de **≥ 1,50**. Todas as 88 foram novamente avaliadas na 3ª instância.

Ou seja, somando-se todas as discrepâncias de **≥ 1,50**, dá-se o total de 363 (231 + 44 + 88), sendo que 362 foram resolvidas. Esse resultado revela que o processo de *triagem* das interações face a face com notas discrepantes pode ser considerado justo, na medida em que foram adotados os procedimentos para reavaliação de provas conforme prevê o edital. Ou seja, está sendo cumprida a decisão da Administração Pública, via Inep, em apresentar recurso de ofício, o que demonstra o seu compromisso ético. No entanto, há que se destacar que o percentual de discrepâncias (do tipo ≥ 1,50, ou seja, entre AI e AO) é maior na segunda instância do que na primeira, o que significa dizer que, embora seja justa e necessária a atitude de recorrer de ofício, os avaliadores da segunda instância têm apresentado maior divergência de interpretação da grade, o que, em alguns casos, justificou a reavaliação das provas em terceira instância.

Ainda sobre as provas com notas discrepantes, uma análise mostrou que há diferenças estatisticamente significativas entre os contextos de aplicação das provas: Brasil e exterior. É plausível a preocupação com o contexto de aplicação do Celpe-Bras, por parte de alguns pesquisadores, a exemplo de Coura-Sobrinho (2014) e Costa (2015). Dessa forma, entendemos que o contexto de aplicação configura-se uma *variável* importante de análise de um teste de proficiência, tendo em vista uma gama de temas que podem surgir para discussões acadêmicas: o entendimento que os avaliadores têm sobre o construto do exame, a diferença da condução das interações face a face e, consequentemente, da avaliação, o entendimento que os avaliadores têm dos critérios da grade, entre outras nuances que permeiam tanto a condução quanto a avaliação da interação face a face.

Trazemos novamente o questionamento de Bachman (1990): *o quanto do desempenho individual em um teste está relacionado ao erro de mensuração ou a outros fatores, além da habilidade linguística que se quer medir?* Retomamos também o problema de pesquisa: *o comportamento avaliativo pode ser considerado uma variável responsável por erro de mensuração que interfere no desempenho oral dos examinandos, além da habilidade linguística que se quer medir?* Aliado a isso, entendemos que o posicionamento de He e Young (1998) seja fundamental para tratarmos da confiabilidade: uma maneira em que as entrevistas podem não ser confiáveis é quando dois avaliadores diferentes julgam, de forma diferente, a habilidade oral de um mesmo sujeito, o que é considerado uma ameaça à confiabilidade dos resultados do teste.

Como se sabe, todo processo de mensuração carrega consigo uma quantidade de erro, aleatório ou sistemático, nos termos de Marôco e Garcia-Marques (2006), sendo que o erro é inversamente proporcional à confiabilidade dos resultados: quanto menos erros, mais

confiabilidade. Portanto, tendo sido constatada a existência de variabilidade de comportamento avaliativo, resta saber se isso pode ser considerado uma variável responsável por erro de mensuração. Foi este, então, um dos principais focos da pesquisa: estimar a confiabilidade dos resultados da prova oral do exame Celpe-Bras e estabelecer relação com o comportamento avaliativo.

Primeiramente, foi feita a Análise dos Componentes Principais, para verificar a dimensionalidade da escala de avaliação. Nas avaliações realizadas em primeira instância, tanto com relação às sete edições quanto à Amostra B da Edição 5, foi constatada uma escala unidimensional. Ou seja, os itens da escala medem um único construto: a proficiência oral. Por outro lado, na avaliação realizada em segunda instância (Amostra B, Edição 5), foi constatada uma escala bidimensional, o que indica que há dois construtos sendo avaliados. Portanto, essa constatação gera uma nova hipótese de pesquisa: a dimensionalidade da escala varia na medida em que varia o comportamento avaliativo.

Outros dois resultados da ACP são relevantes: as variáveis com maior poder de explicação da nota final do observador são: *Adequação Lexical*, *Adequação Gramatical* e *Fluência*. Mais uma vez, léxico e gramática destacam-se entre os critérios que mais pesam na avaliação da proficiência oral. Por outro lado, a variável com menor poder de explicação é *Compreensão* e, inclusive, em duas edições (6 e 7) mostrou valores abaixo do aceitável, indicando que esse descritor deveria ser excluído da grade. Com relação a essa possível exclusão, fazemos considerações a seguir.

Posteriormente à análise da dimensionalidade da escala, foram verificados os índices de consistência interna, via coeficiente *Alfa de Cronbach*.

Nas sete edições analisadas, a avaliação em primeira instância revela que os itens da grade analítica possuem *elevada* consistência interna, indicando confiabilidade elevada. Essa consistência revela-se, inclusive, com o critério *Compreensão* que, se fosse excluído da escala, não promoveria uma melhora significativa na consistência interna dos itens. Ou seja, mesmo apresentando baixo ou inaceitável poder de explicação, como apontado na ACP, *Compreensão* mostra-se uma peça importante na avaliação do construto proficiência oral.

Já com relação à Edição 5, Amostra B, a avaliação realizada em primeira instância revela confiabilidade *elevada*. No entanto, a segunda instância revela confiabilidade *moderada*. Dito de outra maneira, a variabilidade do comportamento avaliativo faz com que diminua a confiabilidade dos resultados do teste. A avaliação em primeira instância revela resultados mais confiáveis do que a segunda.

A análise da consistência interna também levou em conta os dois componentes (duas dimensões) verificados na ACP, na avaliação em segunda instância da Amostra B da Edição 5. O primeiro componente revela confiabilidade *elevada*, mas o segundo, por sua vez, revela confiabilidade *inaceitável*. Ou seja, não há aplicabilidade de dois componentes na escala.

Para finalizar as estimativas de confiabilidade dos resultados do exame, foram verificados os níveis de concordância entre os avaliadores e o quanto dessa concordância é devida ao acaso. Nas sete edições analisadas, foram verificados, ainda que baixos, valores *satisfatórios* do coeficiente *Kappa*.

Na Edição 5, Amostra B, é possível comparar os níveis de concordância na primeira e na segunda instâncias da avaliação das provas que apresentaram discrepância significativa. Nas duas, não foram verificados valores aceitáveis do coeficiente, o que significa dizer que as concordâncias foram mais devido ao acaso (0,166, para a primeira, e 0,133, para a segunda). É aceitável a ocorrência de um coeficiente baixo para a primeira instância dessa Amostra, tendo em vista que nela estão todas as interações cujas notas foram consideradas discrepantes. Entretanto, não é aceitável que a segunda instância apresentasse um coeficiente baixo (e menor, inclusive). É esta a instância responsável por reavaliar as provas e, no entanto, apresenta menor concordância entre seus avaliadores do que a própria instância geradora das discrepâncias.

Ainda com relação à Edição 5, a Amostra D também mostra um valor *satisfatório* do coeficiente Kappa, ainda que baixo. Nessa amostra, estão todas as provas que não tiveram notas com discrepância significativa. Ainda assim, o nível de concordância entre os avaliadores não foi alto.

Em resumo, os valores do coeficiente *Kappa* denotam que a avaliação da proficiência oral, ao longo das edições analisadas, tem demonstrado certo desequilíbrio interpretativo da grade ou, olhando por outro ângulo, revelam que a escala de avaliação não tem sido eficiente, pois os avaliadores não demonstraram grau de concordância muito maior do que o esperado pelo acaso, ainda que os valores tenham sido, em sua maioria, *satisfatórios*.

Um fato interessante a ser apontado é em relação à capacitação que o Inep vem promovendo para os aplicadores e avaliadores do exame. O banco de dados desta pesquisa contém três edições realizadas após a implementação dos eventos de capacitação on-line, em que todos os avaliadores têm que participar e demonstrar rendimento satisfatório. Entretanto, não foram observadas melhoras em relação ao nível de concordância entre os avaliadores. Ou seja, o formato dessas capacitações merece ser avaliado.

Três indícios revelam que o comportamento avaliativo tem relação direta com a confiabilidade dos resultados do exame Celpe-Bras, especialmente em relação à Edição 5, que contém as três instâncias avaliativas:

1- a dimensionalidade da escala varia na medida em que varia o comportamento avaliativo;
2- há menos consistência interna dos itens da escala na avaliação feita em segunda instância;
3- o grau de concordância entre os avaliadores é menor na segunda instância.

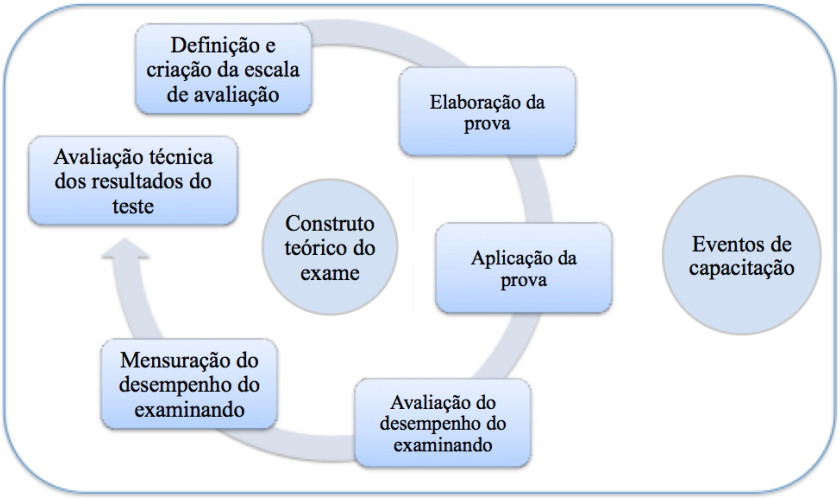
Todos esses resultados nos permitem retomar a pergunta de pesquisa desta tese e afirmar, não mais de forma modalizada, mas a partir de uma constatação estatística: **o comportamento avaliativo é considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados de um teste**. Nesse sentido, a variabilidade de comportamento avaliativo é, nos termos de Murphy e Davidshofer (2005), um fator que contribui para a inconsistência da medida, ou seja, é uma característica que pode afetar os resultados do teste, mas que não tem relação com o atributo que está sendo medido.

Reconhecemos que a variabilidade do comportamento avaliativo pode sofrer interferência da própria condução da interação face a face, ou seja, da maneira como o entrevistador gerencia a prova. Como discutido no capítulo 2, alguns pesquisadores, como Sakamori (2006), Furtoso (2011b), Bottura (2014), Coura-Sobrinho (2014) e Costa (2015), já comprovaram que há variabilidade na gestão da prova oral e que isso pode interferir no desempenho dos examinandos. Portanto, instala-se mais um campo para futuras pesquisas: estabelecer comparações entre a maneira como a interação face a face é conduzida e o comportamento avaliativo dos sujeitos responsáveis por atribuir notas ao desempenho dos examinandos.

Retomando o processo de aplicação da parte oral do Celpe-Bras, para que esta seja bem conduzida e avaliada, é importante que os avaliadores levem em conta o construto do exame, além de permitirem que os examinandos sejam submetidos às mesmas condições ao longo da interação, o que remete à isonomia do processo.

Nesse sentido, é interessante destacar que, para alcançar resultados confiáveis, o construto do exame deve ser refletido em seis fases significativas, desde a definição e criação da escala de avaliação até a análise técnica dos resultados do teste, conforme mostra a Figura 6, a seguir.

Figura 6 - Processo cíclico para a confiabilidade dos resultados da parte oral do exame Celpe-Bras[42]



Fonte: elaborado pela autora, 2018.

Definido o construto teórico do exame, este precisa ser refletido nas ações do processo de avaliação da proficiência oral dos examinandos, a saber:

1ª - a definição e criação da escala de avaliação, em que os critérios a serem avaliados devem ser bem detalhados de forma a diminuir a subjetividade de interpretações. É nessa fase também que são definidos os pesos de cada um dos critérios;

2ª – a elaboração da prova, de forma que o seu conteúdo seja capaz de avaliar o que se pretende;

3ª – a aplicação da prova, com isonomia e maior padronização possível;

4ª – a avaliação do desempenho do examinando por parte do entrevistador e do observador, ação esta que já inicia a seguinte, relativa ao processo de mensuração;

5ª – a mensuração do desempenho do examinando, que envolve a análise de notas discrepantes até que se chegue à nota final, e

6ª – a análise técnica dos resultados do teste, que deve ser feita ao fim de todo o processo de mensuração, para que os resultados sejam avaliados estatisticamente, com vistas a verificar a sua confiabilidade. A partir dessa fase, os administradores do exame podem avaliar se é ou não necessária a redefinição ou readaptação de alguma fase anterior.

[42] Esse processo também pode ser aplicado à parte escrita do exame.

Trata-se, portanto, de um processo cíclico, em que, a cada edição do exame, deve haver certa *padronização* dos procedimentos adotados. Essa padronização vem sendo buscada pelo INEP para as fases 2 a 4, por meio dos eventos de capacitação *on-line*.

Reconhecemos que esses eventos de capacitação de aplicadores e avaliadores sejam importantes para a padronização dos procedimentos que envolvem todo o processo, o que contribui para a confiabilidade dos resultados do exame. Entretanto, notamos que, nos três eventos (para as edições 2017/1, 2016/2 e 2016/1), dos quais participamos, a maneira como o sistema adota para corrigir as respostas dos participantes não simula uma situação real de avaliação. Por exemplo: para a avaliação dos áudios das interações face a face, os participantes devem atribuir uma nota (a cada critério), de acordo com a grade de avaliação, e essa nota, para ser considerada correta, tem de ser igual à nota já atribuída por especialistas àquela interação. Ou seja, não se leva em consideração uma possível *margem de erro*, como ocorre numa situação real, em que as notas são comparadas e somente tidas com *problemáticas* quando apresentam discrepância significativa. Se um participante não *acerta* a maioria das notas, ele terá seu desempenho baixo na capacitação, mesmo que suas notas não tenham tido discrepância significativa, e, por consequência, correrá o risco de não ser selecionado para a aplicação.

Entendemos, então, que os eventos de capacitação devam simular situações reais de avaliação, para que façam sentido para os seus participantes. Da mesma maneira, acreditamos que, dada a importância do significado de *discrepância significativa* e o que ela pode impactar no resultado do desempenho examinando, a capacitação deva promover reflexões aprofundadas sobre as diferenças dos níveis de proficiência.

Refletindo sobre o processo cíclico para a confiabilidade na parte oral do exame, podemos afirmar que é preciso que os sujeitos que dele participam possuam letramento em avaliação de línguas, ou seja, não basta apenas entender sobre o exame, mas saber aplicar o seu construto em todas essas fases, focando no que se espera na e da avaliação dos examinandos. Assim como afirma Scaramucci (2014), esse letramento envolve vários protagonistas, os chamados *stakeholders* (os alunos, professores, elaboradores de políticas, administradores) e são distintos os níveis de conhecimento de cada um desses sujeitos. Assim, conforme aponta a pesquisadora, é necessário que cada um deles seja competente no conhecimento, nas habilidades e nas suas atividades, a depender dos papéis e das suas responsabilidades.

A reflexão sobre esse processo cíclico que propomos representa, nos termos de Vianna (2003), um ponto crítico que se apresenta no cenário da avaliação em larga escala: a avaliação da própria avaliação e, simultaneamente, a auto-avaliação de seus procedimentos, para que seja possível rever as ações e propor novas outras à luz da experiência acumulada.

CONSIDERAÇÕES FINAIS

Com foco na prova oral do Celpe-Bras, esta pesquisa objetivou analisar de que maneira o comportamento avaliativo tem relação com a confiabilidade dos resultados do exame, vista como uma propriedade psicométrica que todos os testes devem buscar.

Sendo os sujeitos avaliadores considerados peça fundamental no processo de mensuração do desempenho oral dos examinandos, o problema de pesquisa surgiu com a seguinte inquietação: *o comportamento avaliativo pode ser considerado uma fonte de erro de mensuração que interfere na confiabilidade dos resultados do teste?*

Para Bachman (1990), uma das preocupações no desenvolvimento e no uso de testes de língua é identificar potenciais fontes de erros em uma dada medida de habilidades e também minimizar o efeito desses erros, com os quais devemos nos preocupar, tendo em vista que qualquer teste de desempenho é afetado por outros fatores além das habilidades que queremos medir. Estabelece-se, idealmente, uma relação de causa e efeito: ao minimizarmos os efeitos dos fatores de erro, minimizamos o erro de mensuração e maximizamos a confiabilidade dos resultados do teste.

> Quando aumentamos a confiabilidade de nossas medições, nós também satisfazemos uma condição necessária para a validade: para um escore de um teste ser considerado válido, ele precisa ser confiável. (...) A preocupação com a confiabilidade e a validade pode ser vista como a condução de dois objetivos complementares no desenho e no desenvolvimento de testes: minimizar os efeitos de erro de mensuração e maximizar os efeitos das habilidades linguísticas que queremos medir (BACHMAN, 1990, p. 160-161).

Nesse sentido, identificar se o comportamento avaliativo é ou não uma fonte de erro de mensuração permite que decisões sejam tomadas pelos administradores do exame, em prol da garantia de resultados confiáveis. Para tanto, motivados pela afirmação de Bachman (2004, p. 3), de que *a compreensão da natureza dos dados quantitativos e como analisá-los estatisticamente são um parte essencial dos testes*, propusemos uma metodologia quantitativa de análise, que levou em conta dados de sete edições consecutivas do exame Celpe-Bras, envolvendo notas de 29.831 candidatos.

No transcurso das análises, constatamos a existência de variabilidade no comportamento avaliativo e também o fato de essa variabilidade interferir negativamente na confiabilidade dos resultados do teste, caracterizando-se, portanto, uma fonte de erro de mensuração. O desafio para os *stakeholders*, portanto, recai sobre a necessidade de fazer com que essa variabilidade não se torne um erro sistemático, pois o instrumento com erro

sistemático é um instrumento com validade reduzida (MARÔCO; GARCIA-MARQUES, 2006, p. 67).

Como afirmam Bachman (1990), Moskal e Leydens (2000) e Brown e Abeywickrama (2010), a variabilidade no processo avaliativo pode comprometer a consistência da avaliação, apesar de reconhecerem que todo processo avaliativo é de natureza subjetiva e que, portanto, pode ser influenciado por erro humano em função de entendimentos enviesados. Para evitar que essa subjetividade não cause impacto indesejável nos resultados do teste, os autores sugerem que haja uma grade de avaliação, com critérios bem definidos, para que os avaliadores saibam o quê e como devem avaliar.

A subjetividade na entrevista oral é claramente marcada, tendo em vista que é na e pela linguagem que é feita a avaliação do desempenho de candidatos. Como afirma Schoffen (2009), a subjetividade está sempre presente nas relações humanas e não é possível excluí-la de nenhum processo de avaliação. Diante disso, conhecer o limite entre a subjetividade e a confiabilidade configura-se como grande desafio para a área da avaliação linguística.

Os resultados apresentados nesta tese direcionam para a necessidade de maior detalhamento da grade de avaliação e, consequentemente, mais investimento em capacitação de avaliadores, para que o comportamento de avaliação não seja uma ameaça à confiabilidade, tendo em vista a significativa diferença de interpretação dos descritores da grade. Ou seja, são necessárias ações para se melhorar o grau de confiabilidade dos resultados do Celpe-Bras.

As duas ações: maior detalhamento da grade e investimento em capacitação podem impactar de forma positiva a *praticidade* do exame. Quanto mais houver uma interpretação equilibrada da grade e, consequentemente, mais alinhados os avaliadores estiverem, menos notas discrepantes serão geradas, menos tempo e recursos financeiros serão gastos no processo de reanálise de provas. Então, há uma convergência de melhoria da praticidade e da confiabilidade.

Afirmamos ser necessária a revisão dos descritores da grade, com base na análise da relação entre o comportamento dos avaliadores e o desempenho do examinando, o qual é marcado por flutuações ao longo da interação face a face e essas flutuações podem ser decorrentes do tópico em discussão (se é mais ou menos familiar ao examinando, se é mais fácil ou mais difícil), do estado físico e mental (cansaço, ansiedade, medo etc.), do comportamento do entrevistador na condução da interação, da (as)simetria enunciativa e de outros fatores externos à própria habilidade que está sendo medida. Dada essa flutuação do desempenho, o examinando pode demonstrar níveis de proficiência variados ao longo dos 20

minutos de prova. Aliado a isso, os resultados desta pesquisa sinalizam que há variabilidade de comportamento avaliativo entre observadores e entrevistadores e entre as instâncias de avaliação. Diante disso, colocamos em discussão dois posicionamentos:

1. seria mais justa se a avaliação do desempenho oral fosse feita considerando cada parte da interação (a cada cinco minutos)? Ou seja, haveria resultados mais confiáveis se os observadores atribuíssem notas para o desempenho do examinando no quebra-gelo e nos três tópicos discutidos a partir dos elementos provocadores?
2. haveria resultados mais confiáveis se a grade de avaliação contivesse sub-habilidades a serem avaliadas separadamente e que, *somadas*, comporiam a totalidade do critério que se quer avaliar?

Esses dois posicionamentos servem para reflexão daquilo que o exame se propõe a avaliar e da maneira como é feita essa avaliação, tendo em vista a existência de flutuação do desempenho tanto do examinando quanto do avaliador.

Retomando o que sinaliza Rosa Becker (2010), de que a avaliação não é um fim em si mesmo, mas um instrumento que pode ser útil para corrigir rumos e delinear o futuro, as discussões apresentadas nesta tese sugerem duas propostas que podem servir de base para outros estudos e também reflexões por parte dos administradores do exame. São elas:

a. Criação de um programa de capacitação contínua

Criação de um programa de capacitação de avaliadores, para que sejam minimizadas algumas variáveis do processo de aplicação do exame e mensuração do desempenho dos candidatos.

Essa preocupação com a capacitação de avaliadores também consta dos trabalhos desenvolvidos por Schoffen (2009) e Cândido (2015). Schoffen (2009), ao tratar sobre a parte escrita do Celpe-Bras, afirma que um treinamento uniforme de todos os avaliadores, bem como pesquisas que tratem do funcionamento da grade de avaliação possibilitam que eles compreendam melhor essas grades e o próprio exame.

Cândido (2015), por sua vez, ao pesquisar a atuação dos entrevistadores, afirma que deva haver capacitações constantes para que a validade e a confiabilidade não sejam ameaçadas, uma vez que podem alinhar o trabalho da equipe, tornando suas atuações mais semelhantes entre si e não criar condições diferentes para o desempenho dos examinandos (CÂNDIDO, 2015, p. 99). A autora reforça a necessidade de investimento em capacitações,

mesmo que os examinadores sejam experientes na aplicação do exame, tendo em vista que algumas dificuldades são inerentes à própria avaliação de proficiência oral.

b. Elaboração de nova grade avaliativa

Elaboração de uma grade de avaliação que permita a atribuição de notas a cada uma das sub-habilidades avaliadas. Por exemplo, o critério *adequação gramatical* (conforme grade apresentada no Anexo B) prevê para:

- a nota 5: uso de variedade ampla de estruturas. Raras inadequações na utilização de estruturas;
- a nota 4: uso de variedade ampla de estruturas. Poucas inadequações na utilização de estruturas complexas e raras inadequações no uso de estruturas básicas;
- a nota 3: uso de variedade de estruturas. Algumas inadequações na utilização de estruturas complexas e poucas inadequações no uso de estruturas básicas;
- a nota 2: uso de variedade limitada de estruturas. Inadequações mais frequentes tanto na utilização de estruturas complexas quanto nas básicas;
- a nota 1: uso de variedade limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas e
- a nota 0: uso de variedade bastante limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas, comprometendo a interação.

Ou seja, há mais de uma habilidade sendo avaliada no critério: o uso de estruturas linguísticas (básicas e complexas) e o uso adequado dessas estruturas. O avaliador, com isso, precisa analisar essas habilidades e atribuir uma única nota ao desempenho em *Adequação Gramatical*. A atribuição de notas separadamente poderia, inclusive, esclarecer melhor a distinção entre *raras*, *poucas* e *algumas* inadequações. Levando em consideração o fato de que o critério *Adequação Gramatical* foi um dos que apresentou maior divergência na avaliação, tal distinção serviria também para dirimir esse desequilíbrio avaliativo.

A preocupação com a atribuição de uma única nota para cada critério da grade analítica também se estende à avaliação feita pelo entrevistador, que atribui uma única nota ao desempenho global do examinando. Ou seja, o AI precisa avaliar a interação face a face e sistematizar o desempenho do examinando em uma nota única. Se o processo de avaliação já carrega um caráter de subjetividade, essa subjetividade pode ser ainda maior nesse momento de avaliar o global, sem que seja possível levar em conta as flutuações de desempenho que ocorrem ao longo dos vinte minutos de prova, flutuações essas marcadas, inclusive, pelas

várias habilidades avaliadas, que não se desenvolvem no mesmo ritmo. Por isso, há a necessidade de estudos acerca da pertinência de o entrevistador fazer também uma avaliação analítica e não holística. Outro fator motivador dessa proposição é a própria composição da nota do desempenho oral do examinando: na avaliação holística (de AI), todos os critérios têm pesos iguais, o que não ocorre na analítica (AO).

A proposta de uma nova grade avaliativa tem a ver com a própria definição de construto que, de acordo com Fulcher (2003) e Costa (2011), é considerado latente quando não pode ser observado diretamente. Portanto, a proficiência oral é um construto latente, assim como todos os critérios constantes das grades de avaliação do Celpe-Bras: compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia. Nessa linha de raciocínio, esses critérios não avaliam as habilidades que suas designações carregam, pois, para cada um deles, é atribuída apenas uma nota, mesmo neles havendo várias sub-habilidades. Significa dizer que o critério de *compreensão* não avalia compreensão, o critério de *fluência* não avalia fluência, e assim por diante. O que eles avaliam é o construto *proficiência oral*. Talvez esse seja um dos motivos para que a variabilidade do comportamento avaliativo, mostrada por desequilíbrio avaliativo, interfira negativamente na confiabilidade dos resultados do exame.

Ou seja, essa nova escala de avaliação pode ser proposta e testada em estudos futuros, de forma que fique mais fácil avaliar os construtos latentes que envolvem o que o exame adota como proficiência oral, diminuindo, assim, níveis de subjetividade que se instalam no processo avaliativo.

Reconhecemos que uma alteração significativa como esta na forma de avaliar requer dois cuidados básicos:

- **testagem**: a proposição de uma nova grade avaliativa deverá, obviamente, passar por testagens com dados reais e os resultados devem ser submetidos à análise estatística para que sejam estimadas a sua confiabilidade e validade;
- **capacitação dos sujeitos envolvidos**: uma vez que os avaliadores já se adaptaram ao modelo atual de mensuração do desempenho dos examinandos, a existência de uma nova grade demanda que esses sujeitos participem de processo(s) de capacitação eficiente.

Uma das limitações da pesquisa foi não ter dados suficientes para estabelecer comparações entre as três instâncias avaliativas de todas as edições analisadas, o que serviria para enriquecer discussões sobre a confiabilidade e traçar o perfil de avaliação que o Celpe-Bras vem adotando ao longo dos anos.

Os resultados apresentados neste trabalho permitem discussões a partir de outros vieses, a exemplos de: a maneira como os avaliadores (das três instâncias) são capacitados para atuar no exame; o estabelecimento de comparações de desempenho de examinandos a partir do perfil dos avaliadores (formação acadêmica, se docentes de PLE ou não, tempo de atuação na área de PLE, tempo de atuação no processo de aplicação e avaliação do exame, língua materna, localidade do posto aplicador etc); o estabelecimento de comparações de desempenho de examinandos a partir dos critérios da grade de avaliação analítica e da localidade do posto aplicador (se no Brasil ou no exterior); o estabelecimento de comparações de discrepâncias geradas ao longo das edições do exame, por localidade do posto aplicador e por estilo de avaliador; a identificação, a partir da grade de avaliação analítica, de aspectos que possam ser responsáveis pela geração de notas discrepantes; a tensão entre confiabilidade e validade; o impacto da necessidade de resultados confiáveis na dimensão *praticidade* de exames em larga escala.

Acreditamos que esses temas possam servir de base não só para a realização de pesquisas científicas, mas também para a tomada de decisões por parte dos administradores do exame, pois entendemos que um exame como o Celpe-Bras deva ser acompanhado permanentemente por estudos e evidências estatísticas que o sustentem.

REFERÊNCIAS

AGOSSA, Mahulikplimi Obed Brice. *O exame Celpe-Bras como instrumento de divulgação da cultura brasileira*: percepções de candidatos. 2017. Dissertação de mestrado. Centro Federal de Educação Tecnológica de Minas Gerias, Belo Horizonte, 2017.

ALMEIDA, Leandro da Silva; VIANA, Fernanda Leopoldina Parente. Testes centrados em critério (TCT). In: PASQUALI, Luiz *e col*. *Instrumentação psicológica*: fundamentos e práticas. Porto Alegre: Artmed, 2010, p. 242-261.

AMARAL, Deise. *A perspectiva dos examinadores sobre o uso da grade de avaliação oral do IELTS*. 2011. Dissertação de mestrado. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

BACHMAN, Lyle F. *Fundamental considerations in language testing*. New York: Oxford University Press, 1990, 408 p.

BACHMAN, Lyle F; BRIAN K. Lynch; MASON, Maureen. Investigating variability in tasks and rater judgements in a performance test of foreing language speaking. In: *Language testing*, 1995, vol. 12, n. 2, pp. 238-257. Disponível em: <https://doi.org/10.1177/026553229501200206>. Acesso em: 22 jan. 2018.

BACHMAN, Lyle F.; PALMER, Adrian S. *Language testing in practice*: designing and developing useful language tests. New York: Oxford University Press, 1996, 377 p.

BACHMAN, Lyle F. *Statistical analyses for language assessment*. New York, Cambridge University Press, 2004, 364p.

BACHMAN, Lyle F.; PALMER, Adrian S. *Language assessment in practice*. New York: Oxford University Press, 2010, 509p.

BAILEY, Kathleen M. *Learning about language assessment*: dilemmas, decisions, and directions. USA, Heinle, Cengale Learning, 1998, 258p.

BARNWELL, David. *Who is to judge how well others speak? An experiment with the ACTFL/ETS Oral Proficiency Scale*. Paper presented at the Eastern States Conference on Linguistics, Pittsburgh, PA. 1986. Disponível em: <https://eric.ed.gov/?id=ED296589>. Acesso em: 10 jun. 2018.

BAUER, Adriana; ALAVARSE, Ocimar Munhoz; OLIVEIRA, Romualdo Portela de. Avaliações em larga escala: uma sistematização do debate. *Educação e Pesquisa*, São Paulo, v. 41, p. 1367-1384, dec. 2015. ISSN 1678-4634. Disponível em: <https://www.revistas.usp.br/ep/article/view/109890/108390>. doi:http://dx.doi.org/10.1590/S1517-9702201508144607. Acesso em: 05 jan. 2018.

BOTTURA, Eleonara Bambozzi. *Exame Celpe-Bras*: uma investigação sobre o papel do entrevistador na interação face a face. 2014. 216 f. Dissertação de Mestrado. Universidade Federal de São Carlos, São Carlos, 2014.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Guia do participante*. Brasília-DF, 2013a. Disponível em: <http://download.inep.gov.br/outras_acoes/celpe_bras/estrutura_exame/2014/guia_participante_celpebras_caderno_provas_comentadas.pdf>. Acesso em: 03 nov. 2014.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Guia de Capacitação para Examinadores da Parte Oral do Celpe-Bras*. Brasília-DF, 2013b. Disponível em: <http://www.ufrgs.br/acervocelpebras/arquivos/guias/guia-de-capacitacao-para-examinadores-da-parte-oral>. Acesso em: 03 nov. 2014

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Portaria nº 334, de 2 de julho de 2013*. Dispõe sobre o credenciamento, recredenciamento e descredenciamento de Postos Aplicadores e define procedimentos para aplicação do Exame para obtenção do Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). Diário Oficial da União, Poder Executivo, Brasília, DF, 4 jul 2013c. Seção 1. p. 16-17. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=04/07/2013&jornal=1&pagina=16&totalArquivos=104>. Acesso em: 13 out. 2016.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Guia do examinando – versão simplificada*. Brasília-DF, 2013d.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 1, de 7 de fevereiro de 2013*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, publicado no D.O.U., em 28 de fevereiro de 2013, seção 3, p. 86 a 88, Brasília-DF, 2013e. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2013>. Acesso em: 16 mar. 2018.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 2, de 17 de julho de 2013*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, publicado no D.O.U., em 18 de julho de 2013, seção 3, p. 66 a 68, Brasília-DF, 2013f. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2013>. Acesso em: 16 mar. 2018.

_____. Agência Nacional de Aviação Civil. *Instruções para o candidato do Santos Dumont Englis Assessment*, ANAC, 2014a. Disponível em: <http://www.anac.gov.br/assuntos/setor-regulado/profissionais-da-aviacao-civil/paginas-complementares/santos-dumont-english-assessment-sdea>. Acesso em: 30 ago. 2016.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 2, de 7 de fevereiro de 2014*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, publicado no D.O.U., em 10 de fevereiro de 2014, seção 3, p. 62 a 64, Brasília-DF, 2014b. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2014>. Acesso em: 16 mar. 2018.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 17, de 15 de julho de 2014*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, publicado no D.O.U., em 16 de julho de 2014, seção 3, p. 55 a 58, Brasília-DF, 2014c. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2014>. Acesso em: 10 fev. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Manual do examinando* – edição novembro de 2015. Brasília-DF, 2015a. Disponível em http://download.inep.gov.br/outras_acoes/celpe_bras/manual/2012/manual_examinando_celpebras.pdf Acesso em: 21 out. 2016

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 2, de 10 de março de 2015*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, publicado no D.O.U., em 11 de março de 2015, seção 3, p. 72-75, Brasília-DF, 2015b. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2015-1>. Acesso em: 10 fev. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 13, de 30 de julho de 2015*, publicado no D.O.U., em 31 de julho de 2015, seção 3, p. 59-62, Brasília-DF, 2015c. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2015-1>. Acesso em: 10 fev. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Manual de orientação para os coordenadores de postos aplicadores do Celpe-Bras* – edição novembro de 2015. Brasília-DF, 2015d. Disponível em http://download.inep.gov.br/outras_acoes/celpe_bras/estrutura_exame/2015/manual_do_aplicador.pdf Acesso em: 12 jul. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 1, de 28 de janeiro de 2016*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros, publicado no D.O.U., em 29 de janeiro de 2016, seção 3, p. 72-77, Brasília-DF, 2016a. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2016>. Acesso em: 10 fev. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 20, de 26 de julho de 2016*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros, publicado no D.O.U., em 27 de julho de 2016, seção 3, p. 62-77, Brasília-DF, 2016b. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2016>. Acesso em: 16 mar. 2018.

_____. Instituto de Controle do Espaço Aéreo. *Manual do Candidato ao Exame de Proficiência em Inglês Aeronáutico do SISCEAB*. Versão 2017.1. São José dos Campos, 2017a. Disponível em: <http://eplis.icea.gov.br/public_html/EPLIS_wp/ManualCandidato/Manual_do_Candidato_versao_2017.1.pdf.>. Acesso em: 25 abr. 2017.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 1, de 2 de março de 2017*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros, publicado no D.O.U., em 3 de março de 2017, seção 3, p. 51-56, Brasília-DF, 2017b. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2017>. Acesso em: 16 mar. 2018.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital 45, de 4 de agosto de 2017*: Certificado de Proficiência em Língua Portuguesa para Estrangeiros, publicado no D.O.U., em 7 de agosto de 2017, seção 3, p. 46-51, Brasília-DF, 2017c. Disponível em: <http://www.ufrgs.br/acervocelpebras/acervo/2017>. Acesso em: 16 mar. 2018.

BROWN, Annie. *Interviewer variability in oral proficiency interviews*. Frankfurt: Peter Lang, 2005, 289p.

BROWN, Douglas H.; ABEYWICKRAMA, Priyanvada. *Language assessment*: principles and classroom practices. 2ª ed. White Plains, Pearson Education, 2010.

CAIRES, Martha da Rocha. *Percepções quanto à proficiência de PFOL*: uma análise comparativa com avaliadores iniciantes e experientes do exame oral do Celpe-Bras. 2014. 61 f. Trabalho de Conclusão de Curso. Universidade Tecnológica Federal do Paraná, Curitiba, 2014.

CAMPOLINA, Isabela Bertho. *Competência intercultural na prova oral do exame Celpe-Bras*: um estudo comparativo. 2017. Trabalho de Conclusão de Curso. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2017.

CÂNDIDO, Marcela Dezotti. *Avaliação da interação face a face no exame Celpe-Bras*: as características dos elementos provocadores e a atuação dos examinadores-interlocutores. 2015. 106 f. Dissertação de Mestrado. Universidade Estadual de Campinas, Campinas, 2015.

CARDOSO, Adnaldo Paulo. Adaptação transcultural e análise da confiabilidade da versão brasileira da Late Life Function and Disability Instrument (LLFDI) em uma amostra de idosos com alta escolaridade no município de Belo Horizonte. 2013. 78 f. Dissertação de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, 2013.

CASTRO, Maria Helena Guimarães. A consolidação da política de avaliação da Educação Básica no Brasil. *Meta*: avaliação. Rio de Janeiro, v. 1, n. 3, p. 271-296, set./dez. 2009. Disponível em http://revistas.cesgranrio.org.br/index.php/metaavaliacao/article/view/51/30 Acesso em: 20 ago. 2017.

CASTRO, Pedrina Barros de. *Produção escrita*: encontros e desencontros entre os livros didáticos de português do Brasil para estrangeiros e o exame Celpe-Bras. 2006. 131 f. Dissertação de Mestrado. Universidade Federal Fluminense, Niterói, 2006.

COELHO, Rafaela Pascoal. *Diferentes olhares sobre a formação de professores de Português como Língua Adicional no Estado de Minas Gerais*. 2015. 110f. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte,

2015.

COSTA, Francisco José da. *Mensuração e desenvolvimento de escalas*: aplicações em administração. Rio de Janeiro: Editora Ciência Moderna Ltda, 2011.

COSTA, Augusto da Silva. A composição das imagens dos elementos provocadores e a interação na parte oral do Celpe-Bras. In: DELL'ISOLA, Regina Lúcia Péret (Org.). *O exame de proficiência Celpe-Bras em foco*. Campinas: Pontes Editores, 2014, p. 87-95.

COSTA, Augusto da Silva. Avaliação de proficiência oral no Celpe-Bras: análise da condução das interações face a face. 2015. 193 f. Dissertação de Mestrado. Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

COURA-SOBRINHO, Jerônimo. O sistema de avaliação Celpe-Bras: o processo de correção e a certificação. In: HORA, Demerval da (Org.). *Língua(s) e Povos:* Unidade e Diversidade. João Pessoa: Idéia, 2006. p. 127-132.

COURA-SOBRINHO, Jerônimo; DELL´ISOLA, Regina Lúcia Péret. O contrato de comunicação na avaliação de proficiência em língua estrangeira. In: JUDICE, Norimar; DELL´ISOLA, Regina Lúcia Péret (Orgs.). *Português – língua estrangeira*: novos diálogos. Niterói: Intertexto, 2009. p. 89-102.

COURA-SOBRINHO, Jerônimo. The face to face interaction to evaluate the oral Portuguese language proficiency. *International Association for Educational Assessment*. Singapore, 40th Annual Conference (2014). Disponível em: <http://www.iaea.info/papers.aspx?id=82>. Acesso em: mar. 2015.

DAMAZO, Liliane Oliveira. *A modalização na produção de textos em português como língua estrangeira.* 2012. 220 f. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2012.

DAVIS, Larry. The influence of training and experience on rater performance in scoring spoken language. In: *Language Testing*, 2016, vol. 33, n. 1, pp. 117-135. Disponível em: <http://journals.sagepub.com/doi/abs/10.1177/0265532215582282>. Acesso em: dez. de 2017.

DINIZ, Leandro Alves Rodrigues; ZOPPI-FONTANA, Mónica Graciela. Política linguística no MERCOSUL: o caso do certificado de proficiência em língua portuguesa para estrangeiros (Celpe-Bras). In: HORA, Demerval da (Org.). *Língua(s) e Povos*: Unidade e Diversidade. João Pessoa: Ideia, 2006. p. 150-156.

DINIZ, Leandro Rodrigues Alves. *Mercado de línguas*: a instrumentalização brasileira do português como língua estrangeira. Campinas, Editora RG, 2010, 159p.

DÖRNYEI, Z. *Qualitative, quantitative and mixed methods research*. In: Research methods in Applied Linguistics: quantitative, qualitative and mixed methodologies. Oxford: OUP, 2007. p. 24-47

DUARTE, Ana Paula Andrade; OLIVEIRA, Regina Purri Brant Hemetério de; MIRANDA, Yara Carolina Campos de. Os gêneros textuais na interação face a face do Celpe-Bras. In: DELL´ISOLA, Regina Lúcia Péret (Org.). *O Exame de proficiência Celpe-Bras em foco*. Campinas: Pontes Editores, 2014. p. 97-110.

ECKES, Thomas. *Introduction to Many-Facet Rasch Measurement*: analyzing and evaluating rater-mediated assessments. Language Testing and Evaluation, vol. 22. 2ª ed. Frankfurt: Peter Lang, 2015.

FERREIRA, Laura Márcia Luiza. *Habilidades de leitura na proposta de interação face a face do exame Celpe-Bras*. 2012. 158 f. Dissertação de Mestrado, Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

_____. Avaliação da proficiência oral: atividades de pós-leitura de listas e gráficos no exame Celpe-Bras. In: DELL'ISOLA, Regina Lúcia Péret (Org.). *O exame de proficiência Celpe-Bras em foco*. Campinas: Pontes Editores, 2014, p. 111-130.

_____. *Avaliação da proficiência oral*: uma análise fatorial e de discriminação de itens do exame Celpe-Bras. 2018. Tese de Doutorado, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte (no prelo).

FONSECA, Ricardo Jorge Rodrigues Moita da; SILVA, Pedro José dos Santos Ponte da; SILVA, Rita Rocha da. Acordo inter-juízes: o caso do coeficiente kappa. *Laboratório de Psicologia*, Lisboa: Instituto Superior de Psicologia Aplicada, 2007, v. 5, n. 1, p. 81-90. Disponível em < http://hdl.handle.net/10400.12/1263>. Acesso em: 21 set. 2017.

FORTES, Melissa Santos. *Uma compreensão etnometodológica do trabalho de fazer ser membro na fala-em-interação de entrevista de proficiência oral em português como língua adicional*. 2009. 329 f. Tese de Doutorado, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

FRANCO, Creso. O SAEB - Sistema de Avaliação da Educação Básica: potencialidades, problemas e desafios. *Revista Brasileira de Educação* [online]. 2001, n.17, pp.127-133. <http://dx.doi.org/10.1590/S1413-24782001000200010>. Acesso em: 8 jan. de 2018.

FULCHER, Glenn. *Testing second language speaking*. London: Longman, 2003.

FULCHER, Glenn. *Practical language testing*. London: Hodder Educational, 2010.

FURTOSO, Viviane Aparecida Bagio. *Desempenho oral em português para falantes de outras línguas*: da avaliação à aprendizagem de línguas estrangeiras em contexto *online*. 2011, 285 f. Tese de Doutorado. Universidade Estadual Paulista, São José do Rio Preto, 2011a.

FURTOSO, Viviane Aparecida Bagio. Avaliação de proficiência em português para falantes de outras línguas: relação com ensino e aprendizagem. In: MENDES, Eleise (Org). *Diálogos interculturais*: ensino e formação em português língua estrangeira, Campinas, Pontes Editores, 2011b. p. 207-236.

GAYA, Karina Figueiredo. *Atividades de compreensão oral como insumo para a produção oral/escrita em português língua estrangeira: preparação para o exame Celpe-Bras*. 2010. 147 f. Dissertação de Mestrado, Universidade Federal do Pará, Belém, 2010.

HE, A. W. & YOUNG, R. Language proficiency interviews: a discourse approach. In: R. Young & A. W. He (Orgs.) *Talking and Testing*: discourse approaches to the assessment of oral proficiency. Philadelphia: John Benjamins, 1998, p. 1-26. Disponível em https://pdfs.semanticscholar.org/c6ce/dacd6ca8959d178ad2b361c77641b7c3804d.pdf. Acesso em: mar. 2017.

HAUCK FILHO, Nelson; ZANON, Cristian. Questões básicas sobre mensuração. In: HUTZ, Claudio Simon; BANDEIRA, Denise Ruschel; TRENTINI, Clarissa Marceli (Orgs.). *Psicometria*. Porto Alegre: Artmed, 2015, p 23-43.

HOGAN, Thomas P.; BENJAMIN, Amy; BREZINSKI, Kristen L. Reliability methods: a note on the frequency of use of various types. In:*Score reliability*: contemporary thinking on reliability issues. EUA: Bruce Thompson Editor, SAGE, 2003b, pp. 59-68.

HUGHES, Arthur. *Testing for language teachers.* 2ª ed. Cambridge: Cambridge University Press, 2003, 251p.

JHA, Naveen Kumar. *Materiais didáticos para o ensino de português língua estrangeira e sua relação com o exame Celpe-Bras*. 2016. 130 f. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

JÚDICE, Norimar. Avaliação: um instrumento de diálogo. In: JÚDICE, Norimar (Org.). *Português / língua estrangeira*: leitura, produção e avaliação de textos. Niterói: Intertexto, 2000, p. 55-64.

LEROY, Henrique Rodrigues. *Ensino de língua portuguesa para estrangeiros em contextos de imersão e não-imersão*: percepções interculturais dos aprendizes e do professor. 2011. 147 f. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2011.

LEVIN, Jack; FOX, James Alan; FORDE, David R. *Estatística para as ciências humanas*. 11ª ed. São Paulo: Pearson Education do Brasil, 2012, 458p.

LIMA, Ronaldo Amorim. *Representações do Brasil em textos do exame CELPE-BRAS.* 2008. 166 f. Tese de Doutorado. Universidade Federal Fluminense, Niterói, 2008.

LOEWEN, Shawn; REINDERS, Hayo. *Key concepts in second language acquisition*. Inglaterra: Macmillan, 2011, 187 p.

LUMLEY, Tom; McNAMARA, Tim. Rater characteristics and rater bias: implications for training. In: *Language Testing*, 1995, vol. 12, n. 1, pp. 54-71. Disponível em https://doi.org/10.1177/026553229501200104. Acesso em: 02 dez. de 2017.

MARÔCO, João; GARCIA-MARQUES, Tereza. *Qual a fiabilidade do Alfa de Cronbach? Questões antigas e soluções modernas?* Instituto Superior de Psicologia Aplicada,

Laboratório de Psicologia, 4, 2006, p. 65-90. Disponível em: <http://hdl.handle.net/10400.12/133> Acesso em: set. 2017.

MARÔCO, João. *Análise estatística*: com o SPSS *Statistics*. Pêro Pinheiro: Report Number, 6ª ed, 2014, 990p.

McNAMARA, Tim. *Language testing*. Oxford: Oxford University Press, 2000.

MEIRON, Beryl E.; SCHICK, Laurie S. Ratings, raters and test performance: an exploratory study. In: KUNNAN, Antony John. *Fairness and validation in language assessment*: selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida. New York: Cambridge University Press, 2000, p.153-176.

MOSKAL, Barbara M.; JON, A. Leydens (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, vol. 7, n. 10, 2000. Disponível em: <http://pareonline.net/getvn.asp?v=7&n=10>. Acesso em: jun. 2017.

MURPHY, Kevin R.; DAVIDSHOFER, Charles O. *Psychological testing*: principles and applications. 6ª ed. International Edition. New Jersey: Pearson Education, 2005.

NEVES, Liliane de Oliveira; AGOSSA, Maulikplimi Obed Brice; COURA-SOBRINHO, Jerônimo. Enquadramento temático na Parte Oral do Exame que Confere o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). In: COURA-SOBRINHO et al. (Orgs). *Estudos em português como língua estrangeira*: um panorama da área. Belo Horizonte: CEFET-MG, 2017, p. 233-247.

NIEDERAUER, Márcia. Competência interacional: critério para avaliação da produção oral em língua adicional. Trab. linguist. apl. [online]. 2014, vol.53, n.2, pp.403-424.

OLIVEIRA, Gilvan Müller de. Política linguística e internacionalização: a língua portuguesa no mundo globalizado do século XXI. *Trabalhos em Linguística Aplicada* [online]. V. 52, N.2, 2013, p. 409-433. ISSN 0103-1813. Disponível em: http://www.scielo.br/pdf/tla/v52n2/a10v52n2.pdf

PASQUALI, Luiz. Testes referentes a construto: teoria e modelo de construção. In: PASQUALI, Luiz *e col*. *Instrumentação psicológica*: fundamentos e práticas. Porto Alegre: Artmed, 2010, p. 165-198.

PINHEIRO, João Ismael D. *et al*. *Estatística básica*: a arte de trabalhar com dados. 2ª. ed. Rio de Janeiro: Elsevier, 2015, 360p.

PONCIANO, Leila; LONGORDO, Monique. Representações da cultura brasileira nos elementos provocadores do Celpe-Bras de 2013. In: DELL'ISOLA, Regina Lúcia Péret (Org.). *O exame de proficiência Celpe-Bras em foco*. Campinas: Pontes Editores, 2014, p. 69-86.

PRODANOV, C. C; FREITAS, E. C. *Metodologia do trabalho científico* [recurso eletrônico].: métodos e técnicas da pesquisa e do trabalho acadêmico. 2 ed. Novo Hamburgo: Feevale, 2013.

ROSA BECKER, Fernanda da. Avaliação educacional em larga escala: a experiência brasileira. *Revista Ibero-americana de Educação*, v. 53, n. 1, p. 1-10, 25 jun. 2010. Disponível em: <https://rieoei.org/RIE/article/view/1751>. Acesso em: 8 jan. de 2018.

SAKAMORI, Lieko. *A atuação do entrevistador na interação face a face do exame Celpe-Bras*. 2006. 190 f. Dissertação de Mestrado, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2006.

SAWILOWSKY, Shlomo S. Reliability as psychometrics versus datametrics. In:*Score reliability*: contemporary thinking on reliability issues. EUA: Bruce Thompson Editor, SAGE, 2003, pp. 103-121.

SANTOS JUNIOR, Elysio Soares. Descendo do salto: uma análise sobre mal-entendidos na interação face a face do Celpe-Bras. In: *International Congresso of Critical Applied Linguistics*. Anais... Brasília, 2015.

SCARAMUCCI, Matilde Virgínia Ricardi. Proficiência em LE: considerações terminológicas e conceituais. *Trabalhos em Linguística Aplicada*. Campinas: IEL/Unicamp, v. 36, p. 11-22, 2000.

_____. (2000/2001). Propostas curriculares e exames vestibulares: potencializando o efeito retroativo benéfico no ensino de LE (Inglês). *Contexturas 5*, p. 97-109, APLIESP, Ibilce, Unesp, São José do Rio Preto, SP.

_____. O Projeto Celpe-Bras no âmbito do Mercosul: contribuições para uma definição de proficiência comunicativa. In: ALMEIDA FILHO, J. C. P. de. (Org.). *Português para estrangeiros interface com o espanhol*. Campinas: Pontes Editores, 2001. p. 77-90.

_____. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. *Trabalhos em Linguística Aplicada*, Campinas, no 43, p. 203-226. 2004.

_____. O exame Celpe-Bras em contexto hispanofalante: percepções de professores e candidatos. In: WIEDEMANN, Lyris; SCARAMUCCI, Matilde V. R. (Orgs.) *Português para falantes de espanhol*: ensino e aquisição. Campinas: Pontes Editores, 2008. p. 175 – 190.

_____. A avaliação da leitura do inglês como língua estrangeira e a validade de construto. *Caleidoscópio*, Vol 7, No. 1, p. 30 – 48, jan/abr. 2009.

_____. Validade e consequências sociais das avaliações em contextos de ensino de línguas. *Lingvarvm Arena*. Vol. 2, Ano 2011 – 103-120.

_____. *Letramento em avaliação no contexto de línguas*. Brasília: UnB, 2014. (Comunicação oral). Disponível em: <https://www.youtube.com/watch?v=E3TnGJgc2wA>. Acesso em: 21 fev. 2017.

SCHLATTER, Margarete. O sistema de avaliação Celpe-Bras: características, implementação e perspectivas. In: HORA, Demerval da (Org.). *Língua(s) e Povos*: Unidade e Diversidade. João Pessoa: Idéia, 2006. p. 171-175.

SCHLATTER, Margarete. *et. al*. Celpe-Bras e CELU: impactos da construção de parâmetros comuns de avaliação de proficiência em português e em espanhol. In: ZOPPI FONTANA, Mónica Graciela (Org.) *O português do Brasil como língua transnacional*. Campinas: Editora RG, 2009.

SCHOFFEN, Juliana Roquele. *Avaliação de proficiência oral em língua estrangeira*: descrição dos níveis de candidatos falantes de espanhol no exame Celpe-Bras. 2003. 101 f. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

_____. *Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras*. 2009. 192 f. Tese de Doutorado, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

_____. Níveis de proficiência oral de examinandos falantes de espanhol no exame Celpe-Bras. In: SCHOFFEN, Juliana Roquele *et al* (Orgs.). *Português como língua adicional*: reflexões para a prática docente. Porto Alegre: Bem Brasil, 2012, p. 145-169.

SCHOLTES, Vanessa A.; TERWEE, Caroline B.; POOLMAN, Rudolf W. *What makes a measurement instrument valid and reliable?* Injury, Int. J. Care Injured 42, 2011, p. 236-240.

SILVA, Ricardo Moutinho Rodrigues da. *O efeito retroativo do Celpe-Bras na cultura de aprender de candidatos ao exame.* 2006. 142 f. Dissertação de mestrado, Universidade Federal de São Carlos, São Carlos, 2006.

SUDBRACK, Edite Maria; COCCO, Eliane Maria. Avaliação em larga escala no Brasil: potencial indutor de qualidade? *Roteiro*, Joaçaba, v. 39, n. 2, p. 347-370, jul./dez. 2014. Disponível em: http://editora.unoesc.edu.br/index.php/roteiro/article/view/4231. Acesso em: 15 set. de 2017.

THOMPSON, Bruce. Understanding reliability and coeficiente alpha, really. In:*Score reliability*: contemporary thinking on reliability issues. EUA: Bruce Thompson Editor, SAGE, 2003a, pp. 3-23.

THOMPSON, Bruce. Guidelines for authors reporting score reliability estimates. In:*Score reliability*: contemporary thinking on reliability issues. EUA: Bruce Thompson Editor, SAGE, 2003b, pp. 91-101.

TOFFOLI, Sônia Ferreira Lopes. *Avaliações em larga escala com itens de respostas construídas no contexto do Modelo Multifacetas de Rasch*. 2015. 315 f. Tese de Doutorado, Universidade Federal de Santa Catarina, Florianópolis, 2015.

TOFFOLI, Sônia Ferreira Lopes; ANDRADE, Dalton Francisco de; BORNIA, Antonio Cezar; QUEVEDO-CAMARGO, Gladys. *Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça*. *Educ. Pesqui.* [online]. 2016, vol.42, n.2, pp.343-358. ISSN 1517-9702. Disponível em: <http://dx.doi.org/10.1590/S1517-

9702201606135887>. Acesso em 15 set. 2017.

TROUCHE, Lygia Maria Gonçalves. Análise da interlocução em elementos provocadores do exame oral Celpe-Bras. In: XVIII Congresso Nacional de Linguística e Filologia, 2014, Rio de Janeiro. *Cadernos do CNLF, Vol. XVIII, n. 07 – Fonética, Fonologia, Ortografia e Política Linguística e de Ensino*, 2014. p. 52-62.

URBINA, Suzana. *Fundamentos da testagem psicológica*. Trad. Cláudia Dornelles. Porto Alegre: Artmed, 2007, 312 p.

VIANNA, Heraldo Marelim. Avaliações nacionais em larga escala: análises e propostas. *Estudos em avaliação educacional*, n. 27, jan-jun/2003. Disponível em: http://dx.doi.org/10.18222/eae02720032177. Acesso em: 17 ago. de 2017.

VIEIRA, Ana Luíza Gabatteli. *Curso online para a parte oral do Celpe-Bras*: contribuições da avaliação de proficiência para o ensino-aprendizagem de PLE. 2016. 199 f. Dissertação de Mestrado. Universidade de Brasília, Brasília, 2016.

WALSH, W. Bruce; BETZ, Nancy E. *Tests and assessment*. 3ª ed. New Jersey: Prentice-Hall, 1995.

ZANON, Cristian; HAUCK FILHO, Nelson. Fidedignidade. In: HUTZ, Claudio Simon; BANDEIRA, Denise Ruschel; TRENTINI, Clarissa Marceli (Orgs.). *Psicometria*. Porto Alegre: Artmed, 2015, p 85-95.

ANEXOS

Anexo A – Grade analítica utilizada pelo avaliador observador (AO)

Celpe-Bras 2011/2

Ficha de Avaliação da Interação Face a Face

Observador

INEP Ministério da Educação

Aspectos a serem avaliados

Com base nas descrições apresentadas na grade de avaliação, assinale o nível que melhor descreve cada aspecto do desempenho do candidato ao longo da interação face a face.

ATENÇÃO:

É obrigatória a avaliação de todos os aspectos.

Avaliação do Observador

Circule o número da descrição que melhor caracteriza o desempenho do candidato

Compreensão	5	4	3	2	1	0
Competência Interacional	5	4	3	2	1	0
Fluência	5	4	3	2	1	0
Adequação Lexical	5	4	3	2	1	0
Adequação Gramatical	5	4	3	2	1	0
Pronúncia	5	4	3	2	1	0

NOME DO OBSERVADOR

DATA: / /

Assinatura do Observador

Fonte: DAMAZO, 2012

Anexo B – Descritores da grade analítica utilizada pelo avaliador observador (AO)

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

Fonte: BRASIL, 2013b.

Anexo C – Grade holística utilizada pelo avaliador interlocutor (AI)

Celpe-Bras 2011/2

Ficha de Avaliação da Interação Face a Face

Entrevistador

INEP Ministério da Educação

Posto:

Inscrição:

Nome:

Sobrenome:

Elementos provocadores utilizados (circule os números correspondentes)

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

Avaliação do Entrevistador
Marque o número da descrição que melhor caracteriza o desempenho do candidato

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

NOME DO ENTREVISTADOR

DATA: / /

Assinatura do Entrevistador

Fonte: DAMAZO, 2012.

Anexo D – Exemplos de elementos provocadores

2017/2 CelpeBras

Interação Face a Face

Elemento Provocador 4

INEP

Supermercado *on-line*

SUPERMERCADO *ON-LINE*

O chamado e-commerce tem facilitado a vida de muita gente. Mas e quando se trata das compras de supermercado? Será que compensa largar a fila do caixa e correr para a tela do computador?

Com apenas alguns cliques, é possível abastecer a despensa com todos os itens necessários, de produtos de limpeza e higiene aos mais variados tipos de alimentos. São poucos, no entanto, os brasileiros que utilizam a internet para compras de supermercado. Um dos principais motivos disso é o receio de receber produtos de baixa qualidade.

Disponível em: https://www.portalvital.com/sua-vida/compras-pela-internet/supermercado-online (adaptado).

2017/2 CelpeBras

Interação Face a Face

Elemento Provocador 5

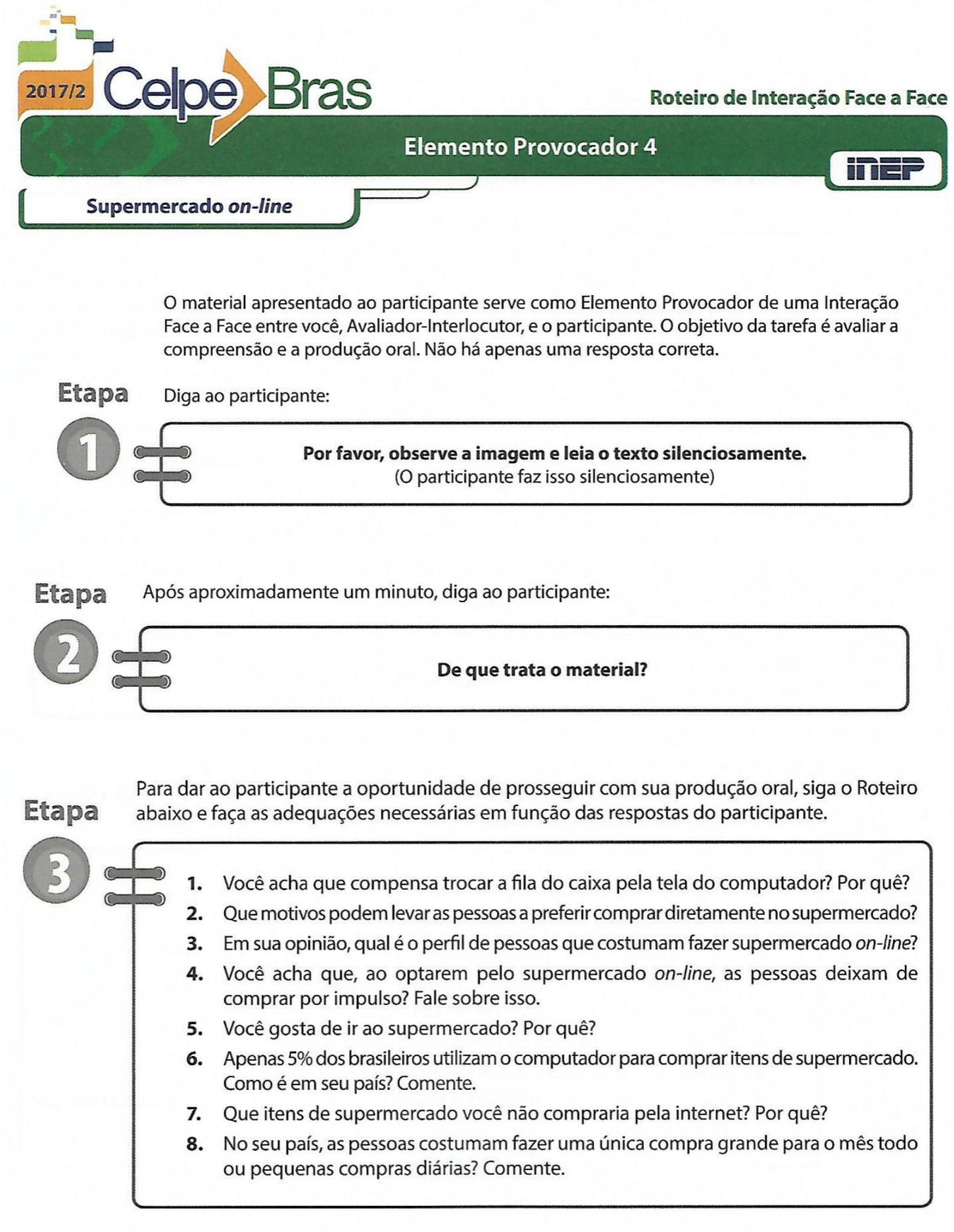
INEP

Machismo

Disponível em: https://s-media-cache-ak0.pinimg.com/736x/d9/11/d8/d911d8db05345928ac7914fce015eae6.jpg (adaptado).

Fonte: http://www.ufrgs.br/acervocelpebras/acervo/2017 - Acesso em 16 mar. 2018.

Anexo E – Exemplo de Roteiro de Interação Face a Face



2017/2 CelpeBras

Roteiro de Interação Face a Face

Elemento Provocador 4

INEP

Supermercado *on-line*

O material apresentado ao participante serve como Elemento Provocador de uma Interação Face a Face entre você, Avaliador-Interlocutor, e o participante. O objetivo da tarefa é avaliar a compreensão e a produção oral. Não há apenas uma resposta correta.

Etapa 1

Diga ao participante:

Por favor, observe a imagem e leia o texto silenciosamente.
(O participante faz isso silenciosamente)

Etapa 2

Após aproximadamente um minuto, diga ao participante:

De que trata o material?

Etapa 3

Para dar ao participante a oportunidade de prosseguir com sua produção oral, siga o Roteiro abaixo e faça as adequações necessárias em função das respostas do participante.

1. Você acha que compensa trocar a fila do caixa pela tela do computador? Por quê?
2. Que motivos podem levar as pessoas a preferir comprar diretamente no supermercado?
3. Em sua opinião, qual é o perfil de pessoas que costumam fazer supermercado *on-line*?
4. Você acha que, ao optarem pelo supermercado *on-line*, as pessoas deixam de comprar por impulso? Fale sobre isso.
5. Você gosta de ir ao supermercado? Por quê?
6. Apenas 5% dos brasileiros utilizam o computador para comprar itens de supermercado. Como é em seu país? Comente.
7. Que itens de supermercado você não compraria pela internet? Por quê?
8. No seu país, as pessoas costumam fazer uma única compra grande para o mês todo ou pequenas compras diárias? Comente.

Fonte: http://www.ufrgs.br/acervocelpebras/acervo/2017 - Acesso em 16 mar. 2018.

APÊNDICES DA PARTE I DO CAPÍTULO V

APÊNDICE 1.1 - Níveis de proficiência (nota final da prova oral) por edição

Edição			Nível Prova Oral					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
	1	Contagem	170	521	979	1001	785	3456
		% em No. Edição	4,9%	15,1%	28,3%	29,0%	22,7%	100,0%
	2	Contagem	174	594	1144	1208	1043	4163
		% em No. Edição	4,2%	14,3%	27,5%	29,0%	25,1%	100,0%
	3	Contagem	111	516	1214	1382	1290	4513
		% em No. Edição	2,5%	11,4%	26,9%	30,6%	28,6%	100,0%
	4	Contagem	152	619	1206	1321	1150	4448
		% em No. Edição	3,4%	13,9%	27,1%	29,7%	25,9%	100,0%
	5	Contagem	189	614	1211	1400	1171	4585
		% em No. Edição	4,1%	13,4%	26,4%	30,5%	25,5%	100,0%
	6	Contagem	269	646	1395	1364	1035	4709
		% em No. Edição	5,7%	13,7%	29,6%	29,0%	22,0%	100,0%
	7	Contagem	199	474	986	1051	1247	3957
		% em No. Edição	5,0%	12,0%	24,9%	26,6%	31,5%	100,0%
Total		Contagem	1264	3984	8135	8727	7721	**29831**
		% em No. Edição	4,2%	13,4%	27,3%	29,3%	25,9%	100,0%

(N=29.831)
Nota: esses dados estão apresentados no Gráfico 2.

APÊNDICE 1.2 – Tabulações cruzadas: níveis de proficiência na visão dos avaliadores da primeira instância, por edição.

1.2.1 – Edição 1

EDIÇÃO 1

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	181	55	5	1	1	243
		% Nível Observador	**74,5%**	22,6%	2,1%	0,4%	0,4%	100,0%
		% Nível Entrevistador	75,1%	9,5%	0,5%	0,1%	0,2%	7,0%
	Intermediário	Contagem	49	323	137	9	5	523
		% Nível Observador	9,4%	**61,8%**	26,2%	1,7%	1,0%	100,0%
		% Nível Entrevistador	20,3%	56,1%	12,8%	0,9%	0,8%	15,1%
	Intermediário Superior	Contagem	10	169	485	106	5	775
		% Nível Observador	1,3%	21,8%	**62,6%**	13,7%	0,6%	100,0%
		% Nível Entrevistador	4,1%	29,3%	45,4%	10,8%	0,8%	22,4%
	Avançado	Contagem	1	27	374	451	52	905
		% Nível Observador	0,1%	3,0%	41,3%	**49,8%**	5,7%	100,0%
		% Nível Entrevistador	0,4%	4,7%	35,0%	46,1%	8,8%	26,2%
	Avançado Superior	Contagem	0	2	68	411	529	1010
		% Nível Observador	0,0%	0,2%	6,7%	40,7%	**52,4%**	100,0%
		% Nível Entrevistador	0,0%	0,3%	6,4%	42,0%	89,4%	29,2%
Total		Contagem	241	576	1069	978	592	3456
		% Nível Observador	7,0%	16,7%	30,9%	28,3%	17,1%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=3.456)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.2 – Edição 2

EDIÇÃO 2

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	155	62	17	4	0	238
		% Nível Observador	**65,1%**	26,1%	7,1%	1,7%	0,0%	100,0%
		% Nível Entrevistador	69,8%	9,7%	1,3%	0,3%	0,0%	5,7%
	Intermediário	Contagem	46	320	141	13	2	522
		% Nível Observador	8,8%	**61,3%**	27,0%	2,5%	0,4%	100,0%
		% Nível Entrevistador	20,7%	49,8%	10,5%	1,0%	0,3%	12,5%
	Intermediário Superior	Contagem	14	193	590	112	11	920
		% Nível Observador	1,5%	21,0%	**64,1%**	12,2%	1,2%	100,0%
		% Nível Entrevistador	6,3%	30,1%	44,0%	8,9%	1,6%	22,1%
	Avançado	Contagem	6	56	484	555	59	1160
		% Nível Observador	0,5%	4,8%	41,7%	**47,8%**	5,1%	100,0%
		% Nível Entrevistador	2,7%	8,7%	36,1%	44,2%	8,4%	27,9%
	Avançado Superior	Contagem	1	11	108	571	632	1323
		% Nível Observador	0,1%	0,8%	8,2%	43,2%	**47,8%**	100,0%
		% Nível Entrevistador	0,5%	1,7%	8,1%	45,5%	89,8%	31,8%
Total		Contagem	222	642	1340	1255	704	4163
		% Nível Observador	5,3%	15,4%	32,2%	30,1%	16,9%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=4.163)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.3 – Edição 3

EDIÇÃO 3

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	130	63	6	1	0	200
		% Nível Observador	**65,0%**	31,5%	3,0%	0,5%	0,0%	100,0%
		% Nível Entrevistador	67,0%	9,5%	0,4%	0,1%	0,0%	4,4%
	Intermediário	Contagem	47	329	152	13	0	541
		% Nível Observador	8,7%	**60,8%**	28,1%	2,4%	0,0%	100,0%
		% Nível Entrevistador	24,2%	49,5%	10,9%	1,0%	0,0%	12,0%
	Intermediário Superior	Contagem	14	213	683	137	6	1053
		% Nível Observador	1,3%	20,2%	**64,9%**	13,0%	0,6%	100,0%
		% Nível Entrevistador	7,2%	32,1%	49,2%	10,2%	0,6%	23,3%
	Avançado	Contagem	2	51	431	648	72	1204
		% Nível Observador	0,2%	4,2%	35,8%	**53,8%**	6,0%	100,0%
		% Nível Entrevistador	1,0%	7,7%	31,0%	48,3%	7,8%	26,7%
	Avançado Superior	Contagem	1	8	117	542	847	1515
		% Nível Observador	0,1%	0,5%	7,7%	35,8%	**55,9%**	100,0%
		% Nível Entrevistador	0,5%	1,2%	8,4%	40,4%	91,6%	33,6%
Total		Contagem	194	664	1389	1341	925	4513
		% Nível Observador	4,3%	14,7%	30,8%	29,7%	20,5%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=4.513)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.4 – Edição 4

EDIÇÃO 4

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	71	55	6	2	1	135
		% Nível Observador	**52,6%**	40,7%	4,4%	1,5%	0,7%	100,0%
		% Nível Entrevistador	57,3%	8,6%	0,4%	0,1%	0,1%	3,0%
	Intermediário	Contagem	32	352	155	15	3	557
		% Nível Observador	5,7%	**63,2%**	27,8%	2,7%	0,5%	100,0%
		% Nível Entrevistador	25,8%	55,2%	11,2%	1,1%	0,3%	12,5%
	Intermediário Superior	Contagem	15	183	660	142	10	1010
		% Nível Observador	1,5%	18,1%	**65,3%**	14,1%	1,0%	100,0%
		% Nível Entrevistador	12,1%	28,7%	47,8%	10,0%	1,1%	22,7%
	Avançado	Contagem	2	35	466	626	71	1200
		% Nível Observador	0,2%	2,9%	38,8%	**52,2%**	5,9%	100,0%
		% Nível Entrevistador	1,6%	5,5%	33,7%	44,0%	8,0%	27,0%
	Avançado Superior	Contagem	4	13	95	637	797	1546
		% Nível Observador	0,3%	0,8%	6,1%	41,2%	**51,6%**	100,0%
		% Nível Entrevistador	3,2%	2,0%	6,9%	44,8%	90,4%	34,8%
Total		Contagem	124	638	1382	1422	882	4448
		% Nível Observador	2,8%	14,3%	31,1%	32,0%	19,8%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=4.448)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.5 – Edição 5

EDIÇÃO 5

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	166	64	17	6	1	254
		% Nível Observador	**65,4%**	25,2%	6,7%	2,4%	0,4%	100,0%
		% Nível Entrevistador	65,6%	8,7%	1,2%	0,4%	0,1%	5,5%
	Intermediário	Contagem	69	330	131	18	0	548
		% Nível Observador	12,6%	**60,2%**	23,9%	3,3%	0,0%	100,0%
		% Nível Entrevistador	27,3%	45,1%	9,4%	1,3%	0,0%	12,0%
	Intermediário Superior	Contagem	14	260	620	133	10	1037
		% Nível Observador	1,4%	25,1%	**59,8%**	12,8%	1,0%	100,0%
		% Nível Entrevistador	5,5%	35,5%	44,4%	9,9%	1,2%	22,6%
	Avançado	Contagem	3	73	519	632	71	1298
		% Nível Observador	0,2%	5,6%	40,0%	**48,7%**	5,5%	100,0%
		% Nível Entrevistador	1,2%	10,0%	37,2%	47,1%	8,3%	28,3%
	Avançado Superior	Contagem	1	5	110	554	778	1448
		% Nível Observador	0,1%	0,3%	7,6%	38,3%	**53,7%**	100,0%
		% Nível Entrevistador	0,4%	0,7%	7,9%	41,3%	90,5%	31,6%
Total		Contagem	253	732	1397	1343	860	4585
		% Nível Observador	5,5%	16,0%	30,5%	29,3%	18,8%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=4.585)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.6 – Edição 6

EDIÇÃO 6

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	144	62	2	1	0	209
		% Nível Observador	**68,9%**	29,7%	1,0%	0,5%	0,0%	100,0%
		% Nível Entrevistador	65,8%	9,2%	0,1%	0,1%	0,0%	4,4%
	Intermediário	Contagem	60	341	137	7	0	545
		% Nível Observador	11,0%	**62,6%**	25,1%	1,3%	0,0%	100,0%
		% Nível Entrevistador	27,4%	50,5%	8,9%	0,5%	0,0%	11,6%
	Intermediário Superior	Contagem	8	246	767	123	1	1145
		% Nível Observador	0,7%	21,5%	**67,0%**	10,7%	0,1%	100,0%
		% Nível Entrevistador	3,7%	36,4%	49,6%	8,4%	0,1%	24,3%
	Avançado	Contagem	5	22	595	764	69	1455
		% Nível Observador	0,3%	1,5%	40,9%	**52,5%**	4,7%	100,0%
		% Nível Entrevistador	2,3%	3,3%	38,5%	52,0%	8,6%	30,9%
	Avançado Superior	Contagem	2	4	45	575	729	1355
		% Nível Observador	0,1%	0,3%	3,3%	42,4%	**53,8%**	100,0%
		% Nível Entrevistador	0,9%	0,6%	2,9%	39,1%	91,2%	28,8%
Total		Contagem	219	675	1546	1470	799	4709
		% Nível Observador	4,7%	14,3%	32,8%	31,2%	17,0%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=4.709)
Nota: esses dados estão apresentados no Gráfico 3.

1.2.7 – Edição 7

EDIÇÃO 7

			Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	95	41	5	0	0	141
		% Nível Observador	**67,4%**	29,1%	3,5%	0,0%	0,0%	100,0%
		% Nível Entrevistador	57,2%	8,7%	0,4%	0,0%	0,0%	3,6%
	Intermediário	Contagem	58	229	107	6	1	401
		% Nível Observador	14,5%	**57,1%**	26,7%	1,5%	0,2%	100,0%
		% Nível Entrevistador	34,9%	48,5%	9,5%	0,5%	0,1%	10,1%
	Intermediário Superior	Contagem	8	186	503	108	2	807
		% Nível Observador	1,0%	23,0%	**62,3%**	13,4%	0,2%	100,0%
		% Nível Entrevistador	4,8%	39,4%	44,7%	8,7%	0,2%	20,4%
	Avançado	Contagem	5	10	459	546	105	1125
		% Nível Observador	0,4%	0,9%	40,8%	**48,5%**	9,3%	100,0%
		% Nível Entrevistador	3,0%	2,1%	40,8%	44,1%	11,0%	28,4%
	Avançado Superior	Contagem	0	6	51	579	847	1483
		% Nível Observador	0,0%	0,4%	3,4%	39,0%	**57,1%**	100,0%
		% Nível Entrevistador	0,0%	1,3%	4,5%	46,7%	88,7%	37,5%
Total		Contagem	166	472	1125	1239	955	3957
		% Nível Observador	4,2%	11,9%	28,4%	31,3%	24,1%	100,0%
		% Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

(N=3.957)
Nota: esses dados estão apresentados no Gráfico 3.

APÊNDICE 1.3 – Tabulação cruzada: percentual dos níveis de proficiência atribuídos pelos avaliadores da 1ª instância – Edição 5 Amostra B

	1ª instância de avaliação		Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	68	27	7	6	1	**109**
		% em Nível Observador	**62,4%**	24,8%	6,4%	5,5%	0,9%	100,0%
		% em Nível Entrevistador	52,3%	10,2%	2,6%	11,1%	9,1%	14,9%
		% do Total	9,3%	3,7%	1,0%	0,8%	0,1%	**14,9%**
	Intermediário	Contagem	44	83	22	11	0	**160**
		% em Nível Observador	27,5%	**51,9%**	13,8%	6,9%	0,0%	100,0%
		% em Nível Entrevistador	33,8%	31,2%	8,1%	20,4%	0,0%	21,8%
		% do Total	6,0%	11,3%	3,0%	1,5%	0,0%	**21,8%**
	Intermediário Superior	Contagem	14	79	87	5	10	**195**
		% em Nível Observador	7,2%	40,5%	**44,6%**	2,6%	5,1%	100,0%
		% em Nível Entrevistador	10,8%	29,7%	32,0%	9,3%	90,9%	26,6%
		% do Total	1,9%	10,8%	11,9%	0,7%	1,4%	**26,6%**
	Avançado	Contagem	3	72	85	21	0	**181**
		% em Nível Observador	1,7%	39,8%	47,0%	**11,6%**	0,0%	100,0%
		% em Nível Entrevistador	2,3%	27,1%	31,3%	38,9%	0,0%	24,7%
		% do Total	0,4%	9,8%	11,6%	2,9%	0,0%	**24,7%**
	Avançado Superior	Contagem	1	5	71	11	0	**88**
		% em Nível Observador	1,1%	5,7%	80,7%	12,5%	0,0%	100,0%
		% em Nível Entrevistador	0,8%	1,9%	26,1%	20,4%	**0,0%**	12,0%
		% do Total	0,1%	0,7%	9,7%	1,5%	0,0%	**12,0%**
Total		Contagem	**130**	**266**	**272**	**54**	**11**	733
		% em Nível Observador	17,7%	36,3%	37,1%	7,4%	1,5%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**17,7%**	**36,3%**	**37,1%**	**7,4%**	**1,5%**	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência atribuídos pelo observador (totais destacados nas linhas) e pelo entrevistador (totais destacados nas colunas) da 1ª instância de avaliação – Amostra B;

- os destaques em negrito na diagonal indicam o percentual de concordância entre os avaliadores, por nível de proficiência;
- esses dados estão apresentados nos Gráficos 4 e 5.

APÊNDICE 1.4 – Tabulação cruzada: percentual dos níveis de proficiência atribuídos pelos avaliadores da 2ª instância – Edição 5 Amostra B

	2ª instância de avaliação		Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	7	8	4	1	0	**20**
		% em Nível Observador	**35,0%**	40,0%	20,0%	5,0%	0,0%	100,0%
		% em Nível Entrevistador	15,6%	5,4%	1,5%	0,5%	0,0%	2,7%
		% do Total	1,0%	1,1%	0,5%	0,1%	0,0%	**2,7%**
	Intermediário	Contagem	11	24	31	7	0	**73**
		% em Nível Observador	15,1%	**32,9%**	42,5%	9,6%	0,0%	100,0%
		% em Nível Entrevistador	24,4%	16,2%	11,8%	3,2%	0,0%	10,0%
		% do Total	1,5%	3,3%	4,2%	1,0%	0,0%	**10,0%**
	Intermediário Superior	Contagem	15	60	94	36	4	**209**
		% em Nível Observador	7,2%	28,7%	**45,0%**	17,2%	1,9%	100,0%
		% em Nível Entrevistador	33,3%	40,5%	35,9%	16,4%	6,8%	28,5%
		% do Total	2,0%	8,2%	12,8%	4,9%	0,5%	**28,5%**
	Avançado	Contagem	10	44	86	94	20	**254**
		% em Nível Observador	3,9%	17,3%	33,9%	**37,0%**	7,9%	100,0%
		% em Nível Entrevistador	22,2%	29,7%	32,8%	42,9%	33,9%	34,7%
		% do Total	1,4%	6,0%	11,7%	12,8%	2,7%	**34,7%**
	Avançado Superior	Contagem	2	12	47	81	35	**177**
		% em Nível Observador	1,1%	6,8%	26,6%	45,8%	**19,8%**	100,0%
		% em Nível Entrevistador	4,4%	8,1%	17,9%	37,0%	59,3%	24,1%
		% do Total	0,3%	1,6%	6,4%	11,1%	4,8%	**24,1%**
Total		Contagem	**45**	**148**	**262**	**219**	**59**	733
		% em Nível Observador	6,1%	20,2%	35,7%	29,9%	8,0%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**6,1%**	**20,2%**	**35,7%**	**29,9%**	**8,0%**	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência atribuídos pelo observador (totais destacados nas linhas) e pelo entrevistador (totais destacados nas colunas) da 2ª instância de avaliação – Amostra B;

- os destaques em negrito na diagonal indicam o percentual de concordância entre os avaliadores, por nível de proficiência;
- esses dados estão apresentados nos Gráficos 4 e 5.

APÊNDICE 1.5 – Tabulação cruzada: percentual dos níveis de proficiência atribuídos pelos avaliadores da 1ª instância – Edição 5 Amostra C

	1ª instancia de avaliação		Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	16	10	0	1	0	27
		% em Nível Observador	**59,3%**	37,0%	0,0%	3,7%	0,0%	100,0%
		% em Nível Entrevistador	47,1%	18,9%	0,0%	25,0%	0,0%	20,8%
		% do Total	12,3%	7,7%	0,0%	0,8%	0,0%	**20,8%**
	Intermediário	Contagem	11	11	5	0	0	27
		% em Nível Observador	40,7%	**40,7%**	18,5%	0,0%	0,0%	100,0%
		% em Nível Entrevistador	32,4%	20,8%	13,5%	0,0%	0,0%	20,8%
		% do Total	8,5%	8,5%	3,8%	0,0%	0,0%	**20,8%**
	Intermediário Superior	Contagem	4	14	13	0	2	33
		% em Nível Observador	12,1%	42,4%	**39,4%**	0,0%	6,1%	100,0%
		% em Nível Entrevistador	11,8%	26,4%	35,1%	0,0%	100,0%	25,4%
		% do Total	3,1%	10,8%	10,0%	0,0%	1,5%	**25,4%**
	Avançado	Contagem	3	18	14	1	0	36
		% em Nível Observador	8,3%	50,0%	38,9%	**2,8%**	0,0%	100,0%
		% em Nível Entrevistador	8,8%	34,0%	37,8%	25,0%	0,0%	27,7%
		% do Total	2,3%	13,8%	10,8%	0,8%	0,0%	**27,7%**
	Avançado Superior	Contagem	0	0	5	2	0	7
		% em Nível Observador	0,0%	0,0%	71,4%	28,6%	**0,0%**	100,0%
		% em Nível Entrevistador	0,0%	0,0%	13,5%	50,0%	0,0%	5,4%
		% do Total	0,0%	0,0%	3,8%	1,5%	0,0%	**5,4%**
Total		Contagem	34	53	37	4	2	130
		% em Nível Observador	26,2%	40,8%	28,5%	3,1%	1,5%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**26,2%**	**40,8%**	**28,5%**	**3,1%**	**1,5%**	100,0%

(N=130)

Notas: - tabulação cruzada dos níveis de proficiência atribuídos pelo observador (totais destacados nas linhas) e pelo entrevistador (totais destacados nas colunas) da 1ª instância de avaliação – Amostra C;

- os destaques em negrito na diagonal indicam o percentual de concordância entre os avaliadores, por nível de proficiência;
- esses dados estão apresentados no Gráfico 6.

APÊNDICE 1.6 – Tabulação cruzada: percentual dos níveis de proficiência atribuídos pelos avaliadores da 2ª instância – Edição 5 Amostra C

	2ª instancia de avaliação		Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	2	0	1	1	0	4
		% em Nível Observador	**50,0%**	0,0%	25,0%	25,0%	0,0%	100,0%
		% em Nível Entrevistador	6,7%	0,0%	2,9%	33,3%	0,0%	3,1%
		% do Total	1,5%	0,0%	0,8%	0,8%	0,0%	**3,1%**
	Intermediário	Contagem	1	0	0	2	0	3
		% em Nível Observador	33,3%	**0,0%**	0,0%	66,7%	0,0%	100,0%
		% em Nível Entrevistador	3,3%	0,0%	0,0%	66,7%	0,0%	2,3%
		% do Total	0,8%	0,0%	0,0%	1,5%	0,0%	**2,3%**
	Intermediário Superior	Contagem	15	2	0	0	4	21
		% em Nível Observador	71,4%	9,5%	**0,0%**	0,0%	19,0%	100,0%
		% em Nível Entrevistador	50,0%	3,4%	0,0%	0,0%	100,0%	16,2%
		% do Total	11,5%	1,5%	0,0%	0,0%	3,1%	**16,2%**
	Avançado	Contagem	10	44	0	0	0	54
		% em Nível Observador	18,5%	81,5%	0,0%	**0,0%**	0,0%	100,0%
		% em Nível Entrevistador	33,3%	75,9%	0,0%	0,0%	0,0%	41,5%
		% do Total	7,7%	33,8%	0,0%	0,0%	0,0%	**41,5%**
	Avançado Superior	Contagem	2	12	34	0	0	48
		% em Nível Observador	4,2%	25,0%	70,8%	0,0%	**0,0%**	100,0%
		% em Nível Entrevistador	6,7%	20,7%	97,1%	0,0%	0,0%	36,9%
		% do Total	1,5%	9,2%	26,2%	0,0%	0,0%	**36,9%**
Total		Contagem	30	58	35	3	4	130
		% em Nível Observador	23,1%	44,6%	26,9%	2,3%	3,1%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**23,1%**	**44,6%**	**26,9%**	**2,3%**	**3,1%**	100,0%

(N=130)

Notas: - tabulação cruzada dos níveis de proficiência atribuídos pelo observador (totais destacados nas linhas) e pelo entrevistador (totais destacados nas colunas) da 2ª instância de avaliação – Amostra C;

- os destaques em negrito na diagonal indicam o percentual de concordância entre os avaliadores, por nível de proficiência;
- esses dados estão apresentados no Gráfico 6.

APÊNDICE 1.7 – Tabulação cruzada: percentual dos níveis de proficiência atribuídos pelos avaliadores da 1ª instância – Edição 5 Amostra D

	1ª instância de avaliação		Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador	Básico	Contagem	98	37	10	0	0	145
		% em Nível Observador	**67,6%**	25,5%	6,9%	0,0%	0,0%	100,0%
		% em Nível Entrevistador	79,7%	7,9%	0,9%	0,0%	0,0%	3,8%
		% do Total	2,5%	1,0%	0,3%	0,0%	0,0%	3,8%
	Intermediário	Contagem	25	247	109	7	0	388
		% em Nível Observador	6,4%	**63,7%**	28,1%	1,8%	0,0%	100,0%
		% em Nível Entrevistador	20,3%	53,0%	9,7%	0,5%	0,0%	10,1%
		% do Total	0,6%	6,4%	2,8%	0,2%	0,0%	10,1%
	Intermediário Superior	Contagem	0	181	533	128	0	842
		% em Nível Observador	0,0%	21,5%	**63,3%**	15,2%	0,0%	100,0%
		% em Nível Entrevistador	0,0%	38,8%	47,4%	9,9%	0,0%	21,9%
		% do Total	0,0%	4,7%	13,8%	3,3%	0,0%	21,9%
	Avançado	Contagem	0	1	434	611	71	1117
		% em Nível Observador	0,0%	0,1%	38,9%	**54,7%**	6,4%	100,0%
		% em Nível Entrevistador	0,0%	0,2%	38,6%	47,4%	8,4%	29,0%
		% do Total	0,0%	0,0%	11,3%	15,9%	1,8%	29,0%
	Avançado Superior	Contagem	0	0	39	543	778	1360
		% em Nível Observador	0,0%	0,0%	2,9%	39,9%	**57,2%**	100,0%
		% em Nível Entrevistador	0,0%	0,0%	3,5%	42,1%	91,6%	35,3%
		% do Total	0,0%	0,0%	1,0%	14,1%	20,2%	35,3%
Total		Contagem	123	466	1125	1289	849	3852
		% em Nível Observador	3,2%	12,1%	29,2%	33,5%	22,0%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	3,2%	12,1%	29,2%	33,5%	22,0%	100,0%

(N=3.582)

Notas: - tabulação cruzada dos níveis de proficiência atribuídos pelo observador (linhas) e pelo entrevistador (colunas) da 1ª instância de avaliação – Amostra D;
- os destaques em negrito na diagonal indicam o percentual de concordância entre os avaliadores, por nível de proficiência;
- esses dados estão apresentados no Gráfico 7.

APÊNDICE 1.8 – Tabulação cruzada: percentual dos níveis de proficiência das notas finais da 1ª e 2ª instâncias – Edição 5 Amostra B

	Níveis das notas finais – 1ª e 2ª instâncias		2ª Instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
1ª Instância	Básico	Contagem	26	52	59	13	1	151
		% em Nível Nota Final 1a Instância	**17,2%**	34,4%	39,1%	8,6%	0,7%	100,0%
		% em Nível Nota Final 2a Instância	86,7%	40,6%	24,4%	5,4%	1,1%	20,6%
		% do Total	3,5%	7,1%	8,0%	1,8%	0,1%	**20,6%**
	Intermediário	Contagem	4	53	76	51	10	194
		% em Nível Nota Final 1a Instância	2,1%	**27,3%**	39,2%	26,3%	5,2%	100,0%
		% em Nível Nota Final 2a Instância	13,3%	41,4%	31,4%	21,2%	10,9%	26,5%
		% do Total	0,5%	7,2%	10,4%	7,0%	1,4%	**26,5%**
	Intermediário Superior	Contagem	0	22	84	121	42	269
		% em Nível Nota Final 1a Instância	0,0%	8,2%	**31,2%**	45,0%	15,6%	100,0%
		% em Nível Nota Final 2a Instância	0,0%	17,2%	34,7%	50,2%	45,7%	36,7%
		% do Total	0,0%	3,0%	11,5%	16,5%	5,7%	**36,7%**
	Avançado	Contagem	0	1	23	56	39	119
		% em Nível Nota Final 1a Instância	0,0%	0,8%	19,3%	**47,1%**	32,8%	100,0%
		% em Nível Nota Final 2a Instância	0,0%	0,8%	9,5%	23,2%	42,4%	16,2%
		% do Total	0,0%	0,1%	3,1%	7,6%	5,3%	**16,2%**
Total		Contagem	30	128	242	241	92	733
		% em Nível Nota Final 1a Instância	4,1%	17,5%	33,0%	32,9%	12,6%	100,0%
		% em Nível Nota Final 2a Instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**4,1%**	**17,5%**	**33,0%**	**32,9%**	**12,6%**	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência da nota final da 1ª instância (totais destacados nas linhas) e da 2ª instancia (totais destacados nas colunas) – Amostra B;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados no Gráfico 8.

APÊNDICE 1.9 – Tabulação cruzada: percentual dos níveis de proficiência das notas finais da 1ª e 3ª instâncias – Edição 5 Amostra C

	Níveis das notas finais – 1ª e 3ª instâncias		3ª Instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
1ª Instância	Básico	Contagem	4	16	15	4	0	39
		% em Nível Nota Final 1a Instância	**10,3%**	41,0%	38,5%	10,3%	0,0%	100,0%
		% em Nível Nota Final 3a Instância	66,7%	66,7%	26,8%	10,5%	0,0%	30,0%
		% do Total	3,1%	12,3%	11,5%	3,1%	0,0%	**30,0%**
	Intermediário	Contagem	1	4	23	6	0	34
		% em Nível Nota Final 1a Instância	2,9%	**11,8%**	67,6%	17,6%	0,0%	100,0%
		% em Nível Nota Final 3a Instância	16,7%	16,7%	41,1%	15,8%	0,0%	26,2%
		% do Total	0,8%	3,1%	17,7%	4,6%	0,0%	**26,2%**
	Intermediário Superior	Contagem	1	4	17	21	4	47
		% em Nível Nota Final 1a Instância	2,1%	8,5%	**36,2%**	44,7%	8,5%	100,0%
		% em Nível Nota Final 3a Instância	16,7%	16,7%	30,4%	55,3%	66,7%	36,2%
		% do Total	0,8%	3,1%	13,1%	16,2%	3,1%	**36,2%**
	Avançado	Contagem	0	0	1	7	2	10
		% em Nível Nota Final 1a Instância	0,0%	0,0%	10,0%	**70,0%**	20,0%	100,0%
		% em Nível Nota Final 3a Instância	0,0%	0,0%	1,8%	18,4%	33,3%	7,7%
		% do Total	0,0%	0,0%	0,8%	5,4%	1,5%	**7,7%**
Total		Contagem	6	24	56	38	6	130
		% em Nível Nota Final 1a Instância	4,6%	18,5%	43,1%	29,2%	4,6%	100,0%
		% em Nível Nota Final 3a Instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**4,6%**	**18,5%**	**43,1%**	**29,2%**	**4,6%**	100,0%

(N=130)

Notas: - tabulação cruzada dos níveis de proficiência da nota final da 1ª instância (totais destacados nas linhas) e da 3ª instancia (totais destacados nas colunas) – Amostra C;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados nos Gráficos 9 e 10.

APÊNDICE 1.10 – Tabulação cruzada: percentual dos níveis de proficiência das notas finais da 2ª e 3ª instâncias – Edição 5 Amostra C

Níveis das notas finais – 2ª e 3ª instâncias			3ª Instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
2ª Instância	Básico	Contagem	2	3	2	0	0	7
		% em Nível Nota Final 2a Instância	**28,6%**	42,9%	28,6%	0,0%	0,0%	100,0%
		% em Nível Nota Final 3a Instância	33,3%	12,5%	3,6%	0,0%	0,0%	5,4%
		% do Total	1,5%	2,3%	1,5%	0,0%	0,0%	**5,4%**
	Intermediário	Contagem	3	11	9	2	0	25
		% em Nível Nota Final 2a Instância	12,0%	**44,0%**	36,0%	8,0%	0,0%	100,0%
		% em Nível Nota Final 3a Instância	50,0%	45,8%	16,1%	5,3%	0,0%	19,2%
		% do Total	2,3%	8,5%	6,9%	1,5%	0,0%	**19,2%**
	Intermediário Superior	Contagem	1	9	40	9	1	60
		% em Nível Nota Final 2a Instância	1,7%	15,0%	**66,7%**	15,0%	1,7%	100,0%
		% em Nível Nota Final 3a Instância	16,7%	37,5%	71,4%	23,7%	16,7%	46,2%
		% do Total	0,8%	6,9%	30,8%	6,9%	0,8%	**46,2%**
	Avançado	Contagem	0	1	5	27	5	38
		% em Nível Nota Final 2a Instância	0,0%	2,6%	13,2%	**71,1%**	13,2%	100,0%
		% em Nível Nota Final 3a Instância	0,0%	4,2%	8,9%	71,1%	83,3%	29,2%
		% do Total	0,0%	0,8%	3,8%	20,8%	3,8%	**29,2%**
Total		Contagem	6	24	56	38	6	130
		% em Nível Nota Final 2a Instância	4,6%	18,5%	43,1%	29,2%	4,6%	100,0%
		% em Nível Nota Final 3a Instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	**4,6%**	**18,5%**	**43,1%**	**29,2%**	**4,6%**	100,0%

(N=130)

Notas: - tabulação cruzada dos níveis de proficiência da nota final da 2ª instância (totais destacados nas linhas) e da 3ª instancia (totais destacados nas colunas) – Amostra C;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados nos Gráficos 9 e 10.

APÊNDICE 1.11 – Tabulação cruzada: percentual dos níveis de proficiência dos observadores da 1ª e 2ª instâncias – Edição 5 Amostra B

			Observador 2ª instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Observador 1ª instância	Básico	Contagem	7	19	55	23	5	109
		% em Nível Observador 1a instância	**6,4%**	17,4%	50,5%	21,1%	4,6%	100,0%
		% em Nível Observador 2a instância	35,0%	26,0%	26,3%	9,1%	2,8%	14,9%
		% do Total	1,0%	2,6%	7,5%	3,1%	0,7%	14,9%
	Intermediário	Contagem	9	34	39	54	24	160
		% em Nível Observador 1a instância	5,6%	**21,3%**	24,4%	33,8%	15,0%	100,0%
		% em Nível Observador 2a instância	45,0%	46,6%	18,7%	21,3%	13,6%	21,8%
		% do Total	1,2%	4,6%	5,3%	7,4%	3,3%	21,8%
	Intermediário Superior	Contagem	4	15	56	77	43	195
		% em Nível Observador 1a instância	2,1%	7,7%	**28,7%**	39,5%	22,1%	100,0%
		% em Nível Observador 2a instância	20,0%	20,5%	26,8%	30,3%	24,3%	26,6%
		% do Total	0,5%	2,0%	7,6%	10,5%	5,9%	26,6%
	Avançado	Contagem	0	4	42	67	68	181
		% em Nível Observador 1a instância	0,0%	2,2%	23,2%	**37,0%**	37,6%	100,0%
		% em Nível Observador 2a instância	0,0%	5,5%	20,1%	26,4%	38,4%	24,7%
		% do Total	0,0%	0,5%	5,7%	9,1%	9,3%	24,7%
	Avançado Superior	Contagem	0	1	17	33	37	88
		% em Nível Observador 1a instância	0,0%	1,1%	19,3%	37,5%	**42,0%**	100,0%
		% em Nível Observador 2a instância	0,0%	1,4%	8,1%	13,0%	20,9%	12,0%
		% do Total	0,0%	0,1%	2,3%	4,5%	5,0%	12,0%
Total		Contagem	20	73	209	254	177	733
		% em Nível Observador 1a instância	2,7%	10,0%	28,5%	34,7%	24,1%	100,0%
		% em Nível Observador 2a instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	2,7%	10,0%	28,5%	34,7%	24,1%	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência dos observadores da 1ª e 2ª instâncias – Amostra B;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados no Gráfico 11.

APÊNDICE 1.12 – Tabulação cruzada: percentual dos níveis de proficiência dos entrevistadores da 1ª e 2ª instâncias – Edição 5 Amostra B

			Entrevistador 2ª instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Entrevistador 1ª Instância	Básico	Contagem	30	44	40	16	0	130
		% em Nível Entrevistador 1a instância	**23,1%**	33,8%	30,8%	12,3%	0,0%	100,0%
		% em Nível Entrevistador 2a instância	66,7%	29,7%	15,3%	7,3%	0,0%	17,7%
		% do Total	4,1%	6,0%	5,5%	2,2%	0,0%	17,7%
	Intermediário	Contagem	13	70	109	63	11	266
		% em Nível Entrevistador 1a instância	4,9%	**26,3%**	41,0%	23,7%	4,1%	100,0%
		% em Nível Entrevistador 2a instância	28,9%	47,3%	41,6%	28,8%	18,6%	36,3%
		% do Total	1,8%	9,5%	14,9%	8,6%	1,5%	36,3%
	Intermediário Superior	Contagem	2	31	91	115	33	272
		% em Nível Entrevistador 1a instância	0,7%	11,4%	**33,5%**	42,3%	12,1%	100,0%
		% em Nível Entrevistador 2a instância	4,4%	20,9%	34,7%	52,5%	55,9%	37,1%
		% do Total	0,3%	4,2%	12,4%	15,7%	4,5%	37,1%
	Avançado	Contagem	0	2	18	20	14	54
		% em Nível Entrevistador 1a instância	0,0%	3,7%	33,3%	**37,0%**	25,9%	100,0%
		% em Nível Entrevistador 2a instância	0,0%	1,4%	6,9%	9,1%	23,7%	7,4%
		% do Total	0,0%	0,3%	2,5%	2,7%	1,9%	7,4%
	Avançado Superior	Contagem	0	1	4	5	1	11
		% em Nível Entrevistador 1a instância	0,0%	9,1%	36,4%	45,5%	**9,1%**	100,0%
		% em Nível Entrevistador 2a instância	0,0%	0,7%	1,5%	2,3%	1,7%	1,5%
		% do Total	0,0%	0,1%	0,5%	0,7%	0,1%	1,5%
Total		Contagem	45	148	262	219	59	733
		% em Nível Entrevistador 1a instância	6,1%	20,2%	35,7%	29,9%	8,0%	100,0%
		% em Nível Entrevistador 2a instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	6,1%	20,2%	35,7%	29,9%	8,0%	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência dos entrevistadores da 1ª e 2ª instâncias – Amostra B;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados no Gráfico 12.

APÊNDICE 1.13 – Tabulação cruzada: percentual dos níveis de proficiência das notas finais da 1ª e 2ª instâncias – Edição 5 Amostra B

			Nota Final 2ª instância					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nota Final 1ª Instância	Básico	Contagem	26	52	59	13	1	151
		% em Nível Nota Final 1a instância	**17,2%**	34,4%	39,1%	8,6%	0,7%	100,0%
		% em Nível Nota Final 2a instância	86,7%	40,6%	24,4%	5,4%	1,1%	20,6%
		% do Total	3,5%	7,1%	8,0%	1,8%	0,1%	20,6%
	Intermediário	Contagem	4	53	76	51	10	194
		% em Nível Nota Final 1a instância	2,1%	**27,3%**	39,2%	26,3%	5,2%	100,0%
		% em Nível Nota Final 2a instância	13,3%	41,4%	31,4%	21,2%	10,9%	26,5%
		% do Total	0,5%	7,2%	10,4%	7,0%	1,4%	26,5%
	Intermediário Superior	Contagem	0	22	84	121	42	269
		% em Nível Nota Final 1a instância	0,0%	8,2%	**31,2%**	45,0%	15,6%	100,0%
		% em Nível Nota Final 2a instância	0,0%	17,2%	34,7%	50,2%	45,7%	36,7%
		% do Total	0,0%	3,0%	11,5%	16,5%	5,7%	36,7%
	Avançado	Contagem	0	1	23	56	39	119
		% em Nível Nota Final 1a instância	0,0%	0,8%	19,3%	**47,1%**	32,8%	100,0%
		% em Nível Nota Final 2a instância	0,0%	0,8%	9,5%	23,2%	42,4%	16,2%
		% do Total	0,0%	0,1%	3,1%	7,6%	5,3%	16,2%
Total		Contagem	30	128	242	241	92	733
		% em Nível Nota Final 1a instância	4,1%	17,5%	33,0%	32,9%	12,6%	100,0%
		% em Nível Nota Final 2a instância	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	4,1%	17,5%	33,0%	32,9%	12,6%	100,0%

(N=733)

Notas: - tabulação cruzada dos níveis de proficiência dos das notas finais da 1ª e 2ª instâncias – Amostra B;
- os destaques em negrito na diagonal indicam o percentual de concordância entre as instâncias, por nível de proficiência;
- esses dados estão apresentados no Gráfico 13.

APÊNDICE 1.14 – Grau de similitude das avaliações: frequência tricotomizada de notas finais - Edição 5 Amostra B

Grau de similitude das **notas finais**				
Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	8	1,1	1,1	1,1
2a instância < 1a instância	133	18,1	18,1	19,2
2a instância > 1a instância	592	80,8	80,8	100
Total	733	100	100	

(N=733)
Notas: - os valores da frequência das notas finais foram tricotomizados;
- esses dados estão apresentados no Gráfico 14.

APÊNDICE 1.15 – Grau de similitude das avaliações: frequência tricotomizada de notas dos entrevistadores - Edição 5 Amostra B

Grau de similitude das avaliações dos **entrevistadores**				
Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	205	28	28	28
2a instância < 1a instância	80	10,9	10,9	38,9
2a instância > 1a instância	448	61,1	61,1	100
Total	733	100	100	

(N=733)
Notas: - os valores da frequência das notas dos entrevistadores foram tricotomizados;
- esses dados estão apresentados no Gráfico 15.

APÊNDICE 1.16 – Grau de similitude das avaliações: frequência tricotomizada de notas dos observadores - Edição 5 Amostra B

Grau de similitude das avaliações dos **observadores**				
Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	14	1,9%	1,9	1,9
2a instância < 1a instância	204	27,8%	27,8	29,7
2a instância > 1a instância	515	70,3%	70,3	100
Total	733	100	100	

(N=733)
Notas: - os valores da frequência das notas dos observadores foram tricotomizados;
- esses dados estão apresentados no Gráfico 16.

APÊNDICE 1.17 – Grau de similitude das avaliações: frequência tricotomizada dos critérios da grade analítica - Edição 5 Amostra B

Grau de similitude em Compreensão

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	371	50,6	50,6	50,6
2a instância < 1a instância	42	5,7	5,7	56,3
2a instância > 1a instância	320	43,7	43,7	100,0
Total	733	100,0	100,0	

Grau de similitude em Competência Interacional

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	207	28,2	28,2	28,2
2a instância < 1a instância	106	14,5	14,5	42,7
2a instância > 1a instância	420	57,3	57,3	100,0
Total	733	100,0	100,0	

Grau de similitude em Fluência

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	209	28,5	28,5	28,5
2a instância < 1a instância	170	23,2	23,2	51,7
2a instância > 1a instância	354	48,3	48,3	100,0
Total	733	100,0	100,0	

Grau de similitude em Adequação Lexical

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	220	30,0	30,0	30,0
2a instância < 1a instância	132	18,0	18,0	48,0
2a instância > 1a instância	381	52,0	52,0	100,0
Total	733	100,0	100,0	

Grau de similitude em Adequação Gramatical

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	215	29,3	29,3	29,3
2a instância < 1a instância	136	18,6	18,6	47,9
2a instância > 1a instância	382	52,1	52,1	100,0
Total	733	100,0	100,0	

Grau de similitude em Pronúncia

Tricotomização das notas	Frequência	Porcentagem	Porcentagem válida	Porcentagem cumulativa
2a instância = 1a instância	221	30,2	30,2	30,2
2a instância < 1a instância	172	23,5	23,5	53,6
2a instância > 1a instância	340	46,4	46,4	100,0
Total	733	100,0	100,0	

N=733, para cada. Nota: esses dados estão apresentados nos Gráficos 17 a 22.

APÊNDICES DA PARTE II DO CAPÍTULO V

APÊNDICE 2.1 – Estatísticas descritivas da população de estudo

2.1.1 – Estatística descritiva: nota final, do observador e do entrevistador

Edição 1 (N=3.456)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,442	0	5	1,050	0,305
Nota do Observador	3,568	0	5	1,012	0,284
Nota do Entrevistador	3,311	0	5	1,167	0,353

Edição 2 (N=4.163)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,519	0	5	0,994	0,282
Nota do Observador	3,663	0	5	0,964	0,263
Nota do Entrevistador	3,370	0	5	1,118	0,332

Edição 3 (N=4.513)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,598	0	5	0,990	0,275
Nota do Observador	3,721	0	5	0,949	0,255
Nota do Entrevistador	3,470	0	5	1,113	0,321

Edição 4 (N=4.448)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,636	0	5	0,932	0,256
Nota do Observador	3,751	0	5	0,902	0,241
Nota do Entrevistador	3,515	0	5	1,055	0,300

Edição 5 (N=4.585)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,537	0	5	1,004	0,284
Nota do Observador	3,676	0	5	0,968	0,263
Nota do Entrevistador	3,394	0	5	1,136	0,335

Edição 6 (N=4.709)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,552	0	5	0,951	0,268
Nota do Observador	3,685	0	5	0,894	0,243
Nota do Entrevistador	3,412	0	5	1,079	0,316

Edição 7 (N=3.957)					
Avaliações em 1a Instância	**Média**	**Mínimo**	**Máximo**	**Desvio padrão**	**C.V. Pearson**
Nota Final	3,714	0	5	0,972	0,262
Nota do Observador	3,832	0	5	0,911	0,238
Nota do Entrevistador	3,590	0	5	1,109	0,309

Legenda: C. V. Pearson = Coeficiente de Variação de Pearson – indica o quão volátil é a média: quanto menor o coeficiente, mais estável é a média. Sua fórmula é: o desvio padrão dividido pela média.

Nota: dados apresentados na Parte II do Capítulo V, item 5.3.

2.1.2 – Estatística descritiva: critérios da grade analítica – avaliação em 1ª instância

Edição 1 (N=3.456)						Edição 2 (N=4.163)				
Avaliações em 1a instância	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**
Compreensão	0	5	4,317	0,982	0,228	0	5	4,359	0,956	0,219
Competência Interacional	0	5	3,765	1,169	0,311	0	5	3,846	1,129	0,294
Fluência	0	5	3,601	1,169	0,325	0	5	3,732	1,124	0,301
Adequação Lexical	0	5	3,232	1,159	0,359	0	5	3,318	1,114	0,336
Adequação Gramatical	0	5	3,181	1,148	0,361	0	5	3,287	1,114	0,339
Pronúncia	0	5	3,429	1,136	0,331	0	5	3,587	1,093	0,305
Edição 3 (N=4.513)						**Edição 4 (N=4.448)**				
Avaliações em 1a instância	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**
Compreensão	0	5	4,373	0,929	0,212	0	5	4,464	0,834	0,187
Competência Interacional	0	5	3,899	1,090	0,280	0	5	3,956	1,037	0,262
Fluência	0	5	3,769	1,110	0,294	0	5	3,798	1,081	0,285
Adequação Lexical	0	5	3,410	1,119	0,328	0	5	3,399	1,092	0,321
Adequação Gramatical	0	5	3,365	1,109	0,330	0	5	3,380	1,081	0,320
Pronúncia	0	5	3,642	1,075	0,295	0	5	3,641	1,057	0,290

(Continua...)

(Conclusão...)

Edição 5 (N=4.585)						Edição 6 (N=4.709)				
Avaliações em 1a instância	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**
Compreensão	0	5	4,450	0,902	0,203	0	5	4,570	0,771	0,169
Competência Interacional	0	5	3,863	1,118	0,289	0	5	3,836	1,082	0,282
Fluência	0	5	3,748	1,141	0,304	0	5	3,737	1,098	0,294
Adequação Lexical	0	5	3,319	1,152	0,347	0	5	3,306	1,099	0,332
Adequação Gramatical	0	5	3,264	1,160	0,355	0	5	3,253	1,086	0,334
Pronúncia	0	5	3,541	1,124	0,317	0	5	3,556	1,095	0,308
Edição 7 (N=3.957)										
Avaliações em 1a instância	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**					
Compreensão	0	5	4,593	0,775	0,169					
Competência Interacional	0	5	3,965	1,056	0,266					
Fluência	0	5	3,859	1,087	0,282					
Adequação Lexical	0	5	3,501	1,129	0,322					
Adequação Gramatical	0	5	3,449	1,127	0,327					
Pronúncia	0	5	3,786	1,088	0,287					

Legenda: C. V. Pearson = Coeficiente de Variação de Pearson – indica o quão volátil é a média: quanto menor o coeficiente, mais estável é a média. Sua fórmula é: o desvio padrão dividido pela média.
Nota: dados apresentados na Parte II do Capítulo V, item 5.4.

APÊNDICE 2.2 – Estatísticas descritivas da Edição 5

2.2.1 Estatística descritiva da Amostra B – 1ª instância

	1a instância					2a instância				
Avaliações	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**	**C.V. Pearson**
Compreensão	0	5	4,124	1,102	0,267	2	5	4,794	0,498	0,104
Competência Interacional	0	5	3,267	1,239	0,379	0	5	4,060	1,042	0,257
Fluência	0	5	3,160	1,213	0,384	0	5	3,583	1,092	0,305
Adequação Lexical	0	5	2,618	1,088	0,416	0	5	3,142	1,020	0,325
Adequação Gramatical	0	5	2,600	1,100	0,423	0	5	3,161	0,986	0,312
Pronúncia	0	5	2,918	1,136	0,389	0	5	3,310	0,983	0,297
Nota Observador	0,17	5	3,088	0,990	0,321	0,67	5	3,661	0,767	0,210
Nota Entrevistador	0	5	2,378	0,928	0,390	0	5	3,130	1,039	0,332

Legenda: C. V. Pearson = Coeficiente de Variação de Pearson – indica o quão volátil é a média: quanto menor o coeficiente, mais estável é a média. Sua fórmula é: o desvio padrão dividido pela média.

Notas: - dados apresentados na Parte II do Capítulo V, item 5.5;
- N=733, para cada instância.

2.2.2 Estatística descritiva da Amostra C

Estatísticas Descritivas					
Instâncias	Mínimo	Máximo	Média	Desvio Padrão	C.V Pearson
1a	0,25	4,21	2,509	0,845	0,337
2a	1	4	2,990	0,840	0,281
3a	1	5	3,108	0,917	0,295
N válido (listwise) = 130					

Legenda: C. V. Pearson = Coeficiente de Variação de Pearson.

Notas: - dados apresentados na Parte II do Capítulo V, item 5.5;
- N=130, para cada instância.

APÊNDICE 2.3 – Testes de normalidade

2.3.1 - Edição 5: teste de normalidade: critérios da grade analítica – Amostra B (1ª e 2ª instâncias)

Teste de Kolmogorov-Smirnov de uma amostra

Avaliação em 1ª instância		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
N		733	733	733	733	733	733
Parâmetros normais[a,b]	Média	4,1241	3,2674	3,1596	2,6180	2,6003	2,9181
	Desvio Padrão	1,10247	1,23894	1,21260	1,08834	1,10010	1,13627
Diferenças Mais Extremas	Absoluto	,294	,187	,164	,194	,177	,175
	Positivo	,213	,129	,146	,158	,173	,145
	Negativo	-,294	-,187	-,164	-,194	-,177	-,175
Estatística do teste		,294	,187	,164	,194	,177	,175
Significância Assint. (Bilateral)		,000[c]	,000[c]	,000[c]	,000[c]	,000[c]	,000[c]

a. A distribuição do teste é Normal.
b. Calculado dos dados.
c. Correção de Significância de Lilliefors.

Teste de Kolmogorov-Smirnov de uma amostra

Avaliação em 2ª instância		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
N		733	733	733	733	733	733
Parâmetros normais[a,b]	Média	4,7940	4,0600	3,5825	3,1419	3,1610	3,3097
	Desvio Padrão	,49849	1,04238	1,09161	1,01978	,98555	,98257
Diferenças Mais Extremas	Absoluto	,494	,248	,193	,207	,199	,197
	Positivo	,340	,184	,159	,184	,199	,186
	Negativo	-,494	-,248	-,193	-,207	-,199	-,197
Estatística do teste		,494	,248	,193	,207	,199	,197
Significância Assint. (Bilateral)		,000[c]	,000[c]	,000[c]	,000[c]	,000[c]	,000[c]

a. A distribuição do teste é Normal.
b. Calculado dos dados.
c. Correção de Significância de Lilliefors.

2.3.2 - Edição 5: teste de normalidade: discrepâncias significativas – Amostra B

Resumo de processamento de casos

Tipo de discrepância	Casos					
	Válido		Omisso		Total	
	N	Porcentagem	N	Porcentagem	N	Porcentagem
≥ 1,50	230	100,0%	0	0,0%	230	100,0%
<1,50	503	100,0%	0	0,0%	503	100,0%

Descritivas

Tipo de discrepância			Estatística	Erro Padrão
≥ 1,50	Média		1,50	,033
	95% Intervalo de Confiança para Média	Limite inferior	1,44	
		Limite superior	1,57	
	5% da média aparada		1,50	
	Mediana		2,00	
	Variância		,251	
	Desvio Padrão		,501	
	Mínimo		1	
	Máximo		2	
	Amplitude		1	
	Amplitude interquartil		1	
	Assimetria		-,018	,160
	Curtose		-2,017	,320
<1,50	Média		1,72	,020
	95% Intervalo de Confiança para Média	Limite inferior	1,68	
		Limite superior	1,76	
	5% da média aparada		1,74	
	Mediana		2,00	
	Variância		,204	
	Desvio Padrão		,452	
	Mínimo		1	
	Máximo		2	
	Amplitude		1	
	Amplitude interquartil		1	
	Assimetria		-,959	,109
	Curtose		-1,084	,217

Testes de Normalidade

Tipo de discrepância	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
≥ 1,50	,343	230	,000	,636	230	,000
<1,50	,451	503	,000	,565	503	,000

a. Correlação de Significância de Lilliefors

APÊNDICE 2.4 – Tabulação cruzada: percentual de concordância em *Adequação Lexical* e *Adequação Gramatical* e teste de hipótese - Edição 5 Amostra B

Variáveis comparadas: Adequação Lexical e Adequação Gramatical (Amostra B)			Adequação Gramatical			Total
			2a instância = 1a instância	2a instância < 1a instância	2a instância > 1a instância	
Adequação Lexical	2a instância = 1a instância	Contagem	127	40	53	220
		% em Adequação Lexical	**57,7%**	18,2%	24,1%	100,0%
		% em Adequação Gramatical	59,1%	29,4%	13,9%	30,0%
		% do Total	17,3%	5,5%	7,2%	30,0%
	2a instância < 1a instância	Contagem	34	92	6	132
		% em Adequação Lexical	25,8%	**69,7%**	4,5%	100,0%
		% em Adequação Gramatical	15,8%	67,6%	1,6%	18,0%
		% do Total	4,6%	12,6%	0,8%	18,0%
	2a instância > 1a instância	Contagem	54	4	323	381
		% em Adequação Lexical	14,2%	1,0%	**84,8%**	100,0%
		% em Adequação Gramatical	25,1%	2,9%	84,6%	52,0%
		% do Total	7,4%	0,5%	44,1%	52,0%
	Total	Contagem	215	136	382	733
		% em Adequação Lexical	29,3%	18,6%	52,1%	100,0%
		% em Adequação Gramatical	100,0%	100,0%	100,0%	100,0%
		% do Total	29,3%	18,6%	52,1%	100,0%

(N=733)

Notas: - tabulação cruzada de Adequação Lexical e Adequação Gramatica da 1ª e 2ª instâncias – Amostra B;
- os destaques em negrito na diagonal indicam o percentual de concordância nos dois critérios;
- esses dados estão apresentados no Gráfico 23.

Testes qui-quadrado

	Valor	gl	Significância Assintótica (Bilateral)
Qui-quadrado de Pearson	508,378[a]	4	,000
Razão de verossimilhança	506,780	4	,000
Associação Linear por Linear	217,244	1	,000
Nº de Casos Válidos	733		

a. 0 células (0,0%) esperavam uma contagem menor que 5. A contagem mínima esperada é 24,49.

APÊNDICE 2.5 – Teste de hipótese: medianas dos critérios analíticos - Edição 5 Amostra B

2ª instância x 1ª instância		N	Posto Médio	Soma de Postos
Compreensão	Postos Negativos	42[a]	127,57	5358,00
	Postos Positivos	320[b]	188,58	60345,00
	Empates	371[c]		
	Total	733		
Competência Interacional	Postos Negativos	106[d]	208,13	22061,50
	Postos Positivos	420[e]	277,48	116539,50
	Empates	207[f]		
	Total	733		
Fluência	Postos Negativos	170[g]	238,63	40566,50
	Postos Positivos	354[h]	273,96	96983,50
	Empates	209[i]		
	Total	733		
Adequação Lexical	Postos Negativos	132[j]	223,89	29553,00
	Postos Positivos	381[k]	268,47	102288,00
	Empates	220[l]		
	Total	733		
Adequação Gramatical	Postos Negativos	136[m]	211,74	28797,00
	Postos Positivos	382[n]	276,50	105624,00
	Empates	215[o]		
	Total	733		
Pronúncia	Postos Negativos	172[p]	225,93	38860,00
	Postos Positivos	340[q]	271,96	92468,00
	Empates	221[r]		
	Total	733		

a. Compreensão 2a instância < Compreensão 1a instância
b. Compreensão 2a instância > Compreensão 1a instância
c. Compreensão 2a instância = Compreensão 1a instância
d. Competência Interacional 2a instância < Competência Interacional 1a instância
e. Competência Interacional 2a instância > Competência Interacional 1a instância
f. Competência Interacional 2a instância = Competência Interacional 1a instância
g. Fluência 2a instância < Fluência 1a instância
h. Fluência 2a instância > Fluência 1a instância
i. Fluência 2a instância = Fluência 1a instância
j. Adequação Lexical 2a instância < Adequação Lexical 1a instância
k. Adequação Lexical 2a instância > Adequação Lexical 1a instância
l. Adequação Lexical 2a instância = Adequação Lexical 1a instância
m. Adequação Gramatical 2a instância < Adequação Gramatical 1a instância
n. Adequação Gramatical 2a instância > Adequação Gramatical 1a instância
o. Adequação Gramatical 2a instância = Adequação Gramatical 1a instância
p. Pronúncia 2a instância < Pronúncia 1a instância
q. Pronúncia 2a instância > Pronúncia 1a instância
r. Pronúncia 2a instância = Pronúncia 1a instância

Estatísticas de teste[a]

Critérios para a par:	Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Z	-14,186[b]	-13,865[b]	-8,430[b]	-11,293[b]	-11,688[b]	-8,359[b]
Significância Assint. (Bilateral)	,000	,000	,000	,000	,000	,000

a. Teste de Postos Assinados por Wilcoxon
b. Com base em postos negativos.

APÊNDICE 2.6 – Correlação entre as notas do observador e do entrevistador: população de estudo

2.6.1 – Edição 1

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,572**
		Sig. (bilateral)	,000
		N	3456
	Competência Interacional	Coeficiente de Correlação	,719**
		Sig. (bilateral)	,000
		N	3456
	Fluência	Coeficiente de Correlação	,771**
		Sig. (bilateral)	,000
		N	3456
	Adequação Lexical	Coeficiente de Correlação	,819**
		Sig. (bilateral)	,000
		N	3456
	Adequação Gramatical	Coeficiente de Correlação	,812**
		Sig. (bilateral)	,000
		N	3456
	Pronúncia	Coeficiente de Correlação	,758**
		Sig. (bilateral)	,000
		N	3456

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.2 – Edição 2

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,551**
		Sig. (bilateral)	,000
		N	4163
	Competência Interacional	Coeficiente de Correlação	,685**
		Sig. (bilateral)	,000
		N	4163
	Fluência	Coeficiente de Correlação	,713**
		Sig. (bilateral)	,000
		N	4163
	Adequação Lexical	Coeficiente de Correlação	,781**
		Sig. (bilateral)	,000
		N	4163
	Adequação Gramatical	Coeficiente de Correlação	,773**
		Sig. (bilateral)	,000
		N	4163
	Pronúncia	Coeficiente de Correlação	,675**
		Sig. (bilateral)	,000
		N	4163

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.3 – Edição 3

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,539**
		Sig. (bilateral)	,000
		N	4513
	Competência Interacional	Coeficiente de Correlação	,699**
		Sig. (bilateral)	,000
		N	4513
	Fluência	Coeficiente de Correlação	,743**
		Sig. (bilateral)	,000
		N	4513
	Adequação Lexical	Coeficiente de Correlação	,799**
		Sig. (bilateral)	,000
		N	4513
	Adequação Gramatical	Coeficiente de Correlação	,791**
		Sig. (bilateral)	,000
		N	4513
	Pronúncia	Coeficiente de Correlação	,722**
		Sig. (bilateral)	,000
		N	4513

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.4 – Edição 4

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,496**
		Sig. (bilateral)	,000
		N	4448
	Competência Interacional	Coeficiente de Correlação	,654**
		Sig. (bilateral)	,000
		N	4448
	Fluência	Coeficiente de Correlação	,722**
		Sig. (bilateral)	,000
		N	4448
	Adequação Lexical	Coeficiente de Correlação	,779**
		Sig. (bilateral)	,000
		N	4448
	Adequação Gramatical	Coeficiente de Correlação	,768**
		Sig. (bilateral)	,000
		N	4448
	Pronúncia	Coeficiente de Correlação	,705**
		Sig. (bilateral)	,000
		N	4448

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.5 – Edição 5

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,477**
		Sig. (bilateral)	,000
		N	4585
	Competência Interacional	Coeficiente de Correlação	,673**
		Sig. (bilateral)	,000
		N	4585
	Fluência	Coeficiente de Correlação	,712**
		Sig. (bilateral)	,000
		N	4585
	Adequação Lexical	Coeficiente de Correlação	,794**
		Sig. (bilateral)	,000
		N	4585
	Adequação Gramatical	Coeficiente de Correlação	,776**
		Sig. (bilateral)	,000
		N	4585
	Pronúncia	Coeficiente de Correlação	,715**
		Sig. (bilateral)	,000
		N	4585

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.6 – Edição 6

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,471**
		Sig. (bilateral)	,000
		N	4709
	Competência Interacional	Coeficiente de Correlação	,679**
		Sig. (bilateral)	,000
		N	4709
	Fluência	Coeficiente de Correlação	,744**
		Sig. (bilateral)	,000
		N	4709
	Adequação Lexical	Coeficiente de Correlação	,814**
		Sig. (bilateral)	,000
		N	4709
	Adequação Gramatical	Coeficiente de Correlação	,801**
		Sig. (bilateral)	,000
		N	4709
	Pronúncia	Coeficiente de Correlação	,711**
		Sig. (bilateral)	,000
		N	4709

**. A correlação é significativa no nível 0,01 (bilateral).

2.6.7 – Edição 7

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,467**
		Sig. (bilateral)	,000
		N	3957
	Competência Interacional	Coeficiente de Correlação	,679**
		Sig. (bilateral)	,000
		N	3957
	Fluência	Coeficiente de Correlação	,739**
		Sig. (bilateral)	,000
		N	3957
	Adequação Lexical	Coeficiente de Correlação	,814**
		Sig. (bilateral)	,000
		N	3957
	Adequação Gramatical	Coeficiente de Correlação	,802**
		Sig. (bilateral)	,000
		N	3957
	Pronúncia	Coeficiente de Correlação	,710**
		Sig. (bilateral)	,000
		N	3957

**. A correlação é significativa no nível 0,01 (bilateral).

Nota: dados apresentados no Gráfico 24.

APÊNDICE 2.7 – Correlação entre as notas do observador e do entrevistador: Edição 5 – Amostra B

2.7.1 – Primeira instância

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,300**
		Sig. (bilateral)	,000
		N	733
	Competência Interacional	Coeficiente de Correlação	,445**
		Sig. (bilateral)	,000
		N	733
	Fluência	Coeficiente de Correlação	,471**
		Sig. (bilateral)	,000
		N	733
	Adequação Lexical	Coeficiente de Correlação	,531**
		Sig. (bilateral)	,000
		N	733
	Adequação Gramatical	Coeficiente de Correlação	,535**
		Sig. (bilateral)	,000
		N	733
	Pronúncia	Coeficiente de Correlação	,468**
		Sig. (bilateral)	,000
		N	733

**. A correlação é significativa no nível 0,01 (bilateral).

2.7.2 – Segunda instância

			Nota Entrevistador
rô de Spearman	Compreensão	Coeficiente de Correlação	,199**
		Sig. (bilateral)	,000
		N	733
	Competência Interacional	Coeficiente de Correlação	,313**
		Sig. (bilateral)	,000
		N	733
	Fluência	Coeficiente de Correlação	,413**
		Sig. (bilateral)	,000
		N	733
	Adequação Lexical	Coeficiente de Correlação	,456**
		Sig. (bilateral)	,000
		N	733
	Adequação Gramatical	Coeficiente de Correlação	,439**
		Sig. (bilateral)	,000
		N	733
	Pronúncia	Coeficiente de Correlação	,355**
		Sig. (bilateral)	,000
		N	733

**. A correlação é significativa no nível 0,01 (bilateral).

Nota: dados apresentados no Gráfico 25.

APÊNDICE 2.8 – Teste de hipótese das discrepâncias: Edição 5

2.8.1 Discrepâncias ≥ 1,50 ponto

			Discrepância ≥ 1,50		Total
			Não discrepância	Discrepância	
Localidade do posto	BRASIL	Contagem	1647	114	1761
		Contagem Esperada	1672,7	88,3	1761,0
		% em Localidade do posto	93,5%	6,5%	100,0%
		% em Discrepância ≥ 1,50	37,8%	49,6%	38,4%
		% do Total	35,9%	2,5%	38,4%
	EXTERIOR	Contagem	2708	116	2824
		Contagem Esperada	2682,3	141,7	2824,0
		% em Localidade do posto	95,9%	4,1%	100,0%
		% em Discrepância ≥ 1,50	62,2%	50,4%	61,6%
		% do Total	59,1%	2,5%	61,6%
Total		Contagem	4355	230	4585
		Contagem Esperada	4355,0	230,0	4585,0
		% em Localidade do posto	95,0%	5,0%	100,0%
		% em Discrepância ≥ 1,50	100,0%	100,0%	100,0%
		% do Total	95,0%	5,0%	100,0%

	Valor	gl	Significância Assintótica (Bilateral)	Sig exata (2 lados)	Sig exata (1 lado)
Qui-quadrado de Pearson	12,743[a]	1	,000		
Correção de continuidade[b]	12,251	1	,000		
Razão de verossimilhança	12,421	1	,000		
Teste Exato de Fisher				,000	,000
Nº de Casos Válidos	4585				

a. 0 células (0,0%) esperavam uma contagem menor que 5. A contagem mínima esperada é 88,34.
b. Computado apenas para uma tabela 2x2

2.8.2 Discrepâncias < 1,50 ponto

			Discrepância < 1,50		Total
			Não discrepância	Discrepância	
Localidade do posto	BRASIL	Contagem	1618	143	1761
		Contagem Esperada	1567,8	193,2	1761,0
		% em Localidade do posto	91,9%	8,1%	100,0%
		% em Discrepância < 1,50	39,6%	28,4%	38,4%
		% do Total	35,3%	3,1%	38,4%
	EXTERIOR	Contagem	2464	360	2824
		Contagem Esperada	2514,2	309,8	2824,0
		% em Localidade do posto	87,3%	12,7%	100,0%
		% em Discrepância < 1,50	60,4%	71,6%	61,6%
		% do Total	53,7%	7,9%	61,6%
Total		Contagem	4082	503	4585
		Contagem Esperada	4082,0	503,0	4585,0
		% em Localidade do posto	89,0%	11,0%	100,0%
		% em Discrepância < 1,50	100,0%	100,0%	100,0%
		% do Total	89,0%	11,0%	100,0%

	Valor	gl	Significância Assintótica (Bilateral)	Sig exata (2 lados)	Sig exata (1 lado)
Qui-quadrado de Pearson	23,780[a]	1	,000		
Correção de continuidade[b]	23,309	1	,000		
Razão de verossimilhança	24,667	1	,000		
Teste Exato de Fisher				,000	,000
Nº de Casos Válidos	4585				

a. 0 células (0,0%) esperavam uma contagem menor que 5. A contagem mínima esperada é 193,19.
b. Computado apenas para uma tabela 2x2

APÊNDICES DA PARTE III DO CAPÍTULO V

APÊNDICE 3.1 – Análise dos Componentes Principais (ACP) da população de estudo

3.1.1 ACP Edição 1

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,725	,697	,652	,630	,611
	Competência Interacional	,725	1,000	,846	,755	,733	,690
	Fluência	,697	,846	1,000	,814	,796	,770
	Adequação Lexical	,652	,755	,814	1,000	,896	,822
	Adequação Gramatical	,630	,733	,796	,896	1,000	,821
	Pronúncia	,611	,690	,770	,822	,821	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,901
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	21455,341
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,639
Competência Interacional	1,000	,789
Fluência	1,000	,853
Adequação Lexical	1,000	,861
Adequação Gramatical	1,000	,839
Pronúncia	1,000	,782

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,763	79,383	79,383	4,763	79,383	79,383
2	,503	8,384	87,767			
3	,292	4,864	92,631			
4	,201	3,348	95,979			
5	,139	2,308	98,288			
6	,103	1,712	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,800
Competência Interacional	,888
Fluência	,923
Adequação Lexical	,928
Adequação Gramatical	,916
Pronúncia	,884

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.2 ACP Edição 2

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,716	,691	,616	,601	,576
	Competência Interacional	,716	1,000	,830	,731	,714	,644
	Fluência	,691	,830	1,000	,781	,767	,712
	Adequação Lexical	,616	,731	,781	1,000	,879	,776
	Adequação Gramatical	,601	,714	,767	,879	1,000	,782
	Pronúncia	,576	,644	,712	,776	,782	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,897
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	23307,686
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,625
Competência Interacional	1,000	,777
Fluência	1,000	,830
Adequação Lexical	1,000	,833
Adequação Gramatical	1,000	,820
Pronúncia	1,000	,730

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,615	76,916	76,916	4,615	76,916	76,916
2	,552	9,194	86,109			
3	,315	5,244	91,354			
4	,238	3,964	95,318			
5	,160	2,674	97,991			
6	,121	2,009	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,791
Competência Interacional	,881
Fluência	,911
Adequação Lexical	,913
Adequação Gramatical	,905
Pronúncia	,854

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.3 ACP Edição 3

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,693	,675	,583	,581	,549
	Competência Interacional	,693	1,000	,830	,720	,701	,638
	Fluência	,675	,830	1,000	,796	,777	,721
	Adequação Lexical	,583	,720	,796	1,000	,877	,795
	Adequação Gramatical	,581	,701	,777	,877	1,000	,796
	Pronúncia	,549	,638	,721	,795	,796	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,896
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	25370,764
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,589
Competência Interacional	1,000	,763
Fluência	1,000	,842
Adequação Lexical	1,000	,837
Adequação Gramatical	1,000	,822
Pronúncia	1,000	,740

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,592	76,534	76,534	4,592	76,534	76,534
2	,590	9,829	86,363			
3	,323	5,383	91,746			
4	,219	3,649	95,395			
5	,155	2,581	97,976			
6	,121	2,024	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,768
Competência Interacional	,873
Fluência	,917
Adequação Lexical	,915
Adequação Gramatical	,907
Pronúncia	,860

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.4 ACP Edição 4

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,669	,655	,557	,538	,507
	Competência Interacional	,669	1,000	,813	,687	,678	,601
	Fluência	,655	,813	1,000	,773	,756	,709
	Adequação Lexical	,557	,687	,773	1,000	,879	,786
	Adequação Gramatical	,538	,678	,756	,879	1,000	,788
	Pronúncia	,507	,601	,709	,786	,788	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,887
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	23801,387
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,554
Competência Interacional	1,000	,735
Fluência	1,000	,830
Adequação Lexical	1,000	,827
Adequação Gramatical	1,000	,813
Pronúncia	1,000	,724

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,484	74,727	74,727	4,484	74,727	74,727
2	,652	10,874	85,601			
3	,342	5,696	91,297			
4	,234	3,903	95,200			
5	,168	2,799	97,998			
6	,120	2,002	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,745
Competência Interacional	,857
Fluência	,911
Adequação Lexical	,910
Adequação Gramatical	,902
Pronúncia	,851

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.5 ACP Edição 5

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,671	,658	,561	,550	,517
	Competência Interacional	,671	1,000	,813	,711	,700	,614
	Fluência	,658	,813	1,000	,770	,765	,703
	Adequação Lexical	,561	,711	,770	1,000	,893	,811
	Adequação Gramatical	,550	,700	,765	,893	1,000	,800
	Pronúncia	,517	,614	,703	,811	,800	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,889
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	25485,037
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,557
Competência Interacional	1,000	,748
Fluência	1,000	,821
Adequação Lexical	1,000	,842
Adequação Gramatical	1,000	,828
Pronúncia	1,000	,734

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,531	75,509	75,509	4,531	75,509	75,509
2	,639	10,654	86,164			
3	,339	5,652	91,815			
4	,210	3,504	95,320			
5	,174	2,901	98,221			
6	,107	1,779	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,746
Competência Interacional	,865
Fluência	,906
Adequação Lexical	,917
Adequação Gramatical	,910
Pronúncia	,857

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.6 ACP Edição 6

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,599	,592	,497	,487	,441
	Competência Interacional	,599	1,000	,782	,669	,657	,553
	Fluência	,592	,782	1,000	,764	,741	,664
	Adequação Lexical	,497	,669	,764	1,000	,877	,768
	Adequação Gramatical	,487	,657	,741	,877	1,000	,755
	Pronúncia	,441	,553	,664	,768	,755	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,882
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	22956,782
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,478
Competência Interacional	1,000	,702
Fluência	1,000	,808
Adequação Lexical	1,000	,829
Adequação Gramatical	1,000	,808
Pronúncia	1,000	,687

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,312	71,864	71,864	4,312	71,864	71,864
2	,708	11,793	83,656			
3	,405	6,753	90,409			
4	,256	4,272	94,681			
5	,198	3,296	97,978			
6	,121	2,022	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,691
Competência Interacional	,838
Fluência	,899
Adequação Lexical	,911
Adequação Gramatical	,899
Pronúncia	,829

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.1.7 ACP Edição 7

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,618	,589	,511	,498	,489
	Competência Interacional	,618	1,000	,795	,694	,683	,592
	Fluência	,589	,795	1,000	,777	,763	,708
	Adequação Lexical	,511	,694	,777	1,000	,885	,789
	Adequação Gramatical	,498	,683	,763	,885	1,000	,766
	Pronúncia	,489	,592	,708	,789	,766	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,884
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	20474,169
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,492
Competência Interacional	1,000	,726
Fluência	1,000	,821
Adequação Lexical	1,000	,837
Adequação Gramatical	1,000	,816
Pronúncia	1,000	,723

Método de Extração: Análise de Componente Principal.

Variância total explicada

Componente	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,414	73,560	73,560	4,414	73,560	73,560
2	,664	11,067	84,627			
3	,385	6,420	91,047			
4	,242	4,041	95,088			
5	,181	3,017	98,106			
6	,114	1,894	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,701
Competência Interacional	,852
Fluência	,906
Adequação Lexical	,915
Adequação Gramatical	,903
Pronúncia	,850

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

Nota: dados apresentados na Parte III do Capítulo V.

APÊNDICE 3.2 – Análise dos Componentes Principais (ACP) da Edição 5 – Amostra B

3.2.1 ACP Primeira instância

N=733

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,671	,671	,550	,534	,509
	Competência Interacional	,671	1,000	,813	,697	,687	,563
	Fluência	,671	,813	1,000	,736	,737	,652
	Adequação Lexical	,550	,697	,736	1,000	,880	,803
	Adequação Gramatical	,534	,687	,737	,880	1,000	,787
	Pronúncia	,509	,563	,652	,803	,787	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,878
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	3895,029
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,563
Competência Interacional	1,000	,736
Fluência	1,000	,800
Adequação Lexical	1,000	,828
Adequação Gramatical	1,000	,815
Pronúncia	1,000	,703

Método de Extração: Análise de Componente Principal.

3.2.1 ACP Primeira instância (continuidade)
N=733

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	4,446	74,094	74,094	4,446	74,094	74,094
2	,688	11,469	85,563			
3	,366	6,101	91,664			
4	,209	3,476	95,140			
5	,173	2,890	98,030			
6	,118	1,970	100,000			

Método de Extração: Análise de Componente Principal.

Matriz de componente[a]

	Componente
	1
Compreensão	,750
Competência Interacional	,858
Fluência	,895
Adequação Lexical	,910
Adequação Gramatical	,903
Pronúncia	,839

Método de Extração: Análise de Componente Principal.
a. 1 componentes extraídos.

Matriz de componente rotativa[a]

a. Apenas um componente foi extraído. A solução não pode ser girada.

3.2.2 ACP Segunda instância

N=733

Matriz de correlações

		Compreensão	Competência Interacional	Fluência	Adequação Lexical	Adequação Gramatical	Pronúncia
Correlação	Compreensão	1,000	,479	,419	,332	,262	,195
	Competência Interacional	,479	1,000	,699	,552	,485	,299
	Fluência	,419	,699	1,000	,723	,696	,527
	Adequação Lexical	,332	,552	,723	1,000	,855	,694
	Adequação Gramatical	,262	,485	,696	,855	1,000	,690
	Pronúncia	,195	,299	,527	,694	,690	1,000

Teste de KMO e Bartlett

Medida Kaiser-Meyer-Olkin de adequação de amostragem.		,831
Teste de esfericidade de Bartlett	Aprox. Qui-quadrado	2798,737
	gl	15
	Sig.	,000

Comunalidades

	Inicial	Extração
Compreensão	1,000	,745
Competência Interacional	1,000	,746
Fluência	1,000	,784
Adequação Lexical	1,000	,871
Adequação Gramatical	1,000	,860
Pronúncia	1,000	,760

Método de Extração: Análise de Componente Principal.

Variância total explicada

	Autovalores iniciais			Somas de extração de carregamentos ao quadrado			Somas de rotação de carregamentos ao quadrado		
Componente	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa	Total	% de variância	% cumulativa
1	3,725	62,076	62,076	3,725	62,076	62,076	2,885	48,075	48,075
2	1,041	17,345	79,421	1,041	17,345	79,421	1,881	31,345	79,421
3	,556	9,262	88,682						
4	,301	5,024	93,707						
5	,239	3,978	97,685						
6	,139	2,315	100,000						

Método de Extração: Análise de Componente Principal.

3.2.2 ACP Segunda instância (continuidade)

Matriz de componente[a]

	Componente	
	1	2
Compreensão	,513	,694
Competência Interacional	,737	,449
Fluência	,877	,119
Adequação Lexical	,909	-,209
Adequação Gramatical	,878	-,298
Pronúncia	,741	-,458

Método de Extração: Análise de Componente Principal.
a. 2 componentes extraídos.

Matriz de componente rotativa[a]

	Componente	
	1	2
Compreensão	,037	,862
Competência Interacional	,360	,785
Fluência	,661	,590
Adequação Lexical	,871	,336
Adequação Gramatical	,895	,244
Pronúncia	,871	,035

Método de Extração: Análise de Componente Principal.
Método de Rotação: Varimax com Normalização de Kaiser.
a. Rotação convergida em 3 iterações.

Matriz de transformação de componente

Componente	1	2
1	,829	,559
2	-,559	,829

Método de Extração: Análise de Componente Principal.
Método de Rotação: Varimax com Normalização de Kaiser.

APÊNDICE 3.3 – Cálculo do Coeficiente *Alfa de Cronbach*: população de estudo

3.3.1 Alfa Edição 1

N=3.456

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,948	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	Coeficiente *Alfa de Cronbach* se o item for excluído
Compreensão	17,2080	27,927	,726	,951
Competência Interacional	17,7595	25,234	,836	,939
Fluência	17,9236	24,757	,886	,933
Adequação Lexical	18,2925	24,793	,891	,932
Adequação Gramatical	18,3438	25,056	,874	,934
Pronúncia	18,0955	25,586	,831	,939

3.3.2 Alfa Edição 2

N=4.163

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,940	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	17,7694	25,148	,712	,941
Competência Interacional	18,2820	22,733	,825	,928
Fluência	18,3963	22,406	,866	,923
Adequação Lexical	18,8107	22,472	,869	,922
Adequação Gramatical	18,8415	22,566	,858	,924
Pronúncia	18,5414	23,355	,789	,932

3.3.3 Alfa Edição 3

N=4.513

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,939	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	18,0842	24,604	,684	,942
Competência Interacional	18,5582	22,221	,813	,928
Fluência	18,6878	21,529	,875	,920
Adequação Lexical	19,0474	21,484	,871	,920
Adequação Gramatical	19,0924	21,655	,860	,922
Pronúncia	18,8156	22,490	,797	,930

3.3.4 Alfa Edição 4

N=4.448

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,932	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	18,1742	22,836	,655	,937
Competência Interacional	18,6823	20,304	,789	,921
Fluência	18,8402	19,363	,864	,911
Adequação Lexical	19,2394	19,274	,864	,911
Adequação Gramatical	19,2590	19,449	,853	,913
Pronúncia	18,9973	20,188	,784	,922

3.3.5 Alfa Edição 5

N=4.585

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,935	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	17,7352	26,168	,658	,940
Competência Interacional	18,3226	23,188	,800	,924
Fluência	18,4371	22,475	,858	,916
Adequação Lexical	18,8661	22,222	,875	,914
Adequação Gramatical	18,9215	22,248	,865	,915
Pronúncia	18,6445	23,215	,792	,925

3.3.6 Alfa Edição 6

N=4.709

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,921	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	17,6874	23,210	,593	,929
Competência Interacional	18,4207	19,736	,760	,909
Fluência	18,5205	18,937	,844	,897
Adequação Lexical	18,9514	18,784	,863	,894
Adequação Gramatical	19,0038	19,011	,846	,896
Pronúncia	18,7010	19,714	,751	,910

3.3.7 Alfa Edição 7

N=3.957

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,928	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	18,5605	24,009	,605	,936
Competência Interacional	19,1880	20,607	,780	,916
Fluência	19,2944	19,748	,856	,905
Adequação Lexical	19,6520	19,283	,872	,903
Adequação Gramatical	19,7048	19,432	,855	,906
Pronúncia	19,3672	20,343	,782	,916

APÊNDICE 3.4 – Cálculo do Coeficiente *Alfa de Cronbach*: edição 5 – Amostra B

3.4.1 Alfa Primeira Instância

N=733

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,929	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	14,5634	26,235	,665	,931
Competência Interacional	15,4202	23,820	,795	,916
Fluência	15,5280	23,556	,845	,908
Adequação Lexical	16,0696	24,548	,856	,908
Adequação Gramatical	16,0873	24,547	,845	,909
Pronúncia	15,7694	25,066	,757	,920

3.4.2 Alfa Segunda Instância

N=733

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,875	6

Estatísticas de item-total

	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	17,2551	18,316	,408	,891
Competência Interacional	17,9891	14,344	,617	,865
Fluência	18,4666	12,826	,804	,830
Adequação Lexical	18,9072	13,032	,846	,822
Adequação Gramatical	18,8881	13,523	,801	,831
Pronúncia	18,7394	14,677	,619	,864

APÊNDICE 3.5 – Cálculo do Coeficiente *Alfa de Cronbach*: edição 5 – Amostra B, 2ª instância (componentes extraídos na ACP)

3.5.1 Componente 1 (Fluência, Adequação Lexical, Adequação Gramatical e Pronúncia)

N=733

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,901	4

Estatísticas de item-total

COMPONENTE 1	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Fluência	9,6126	7,423	,713	,899
Adequação Lexical	10,0532	7,094	,869	,839
Adequação Gramatical	10,0341	7,328	,854	,846
Pronúncia	9,8854	8,031	,693	,902

3.5.2 Componente 2 (Compreensão e Competência Interacional)

N=733

Estatísticas de confiabilidade

Coeficiente *Alfa de Cronbach*	N de itens
,543	2

Estatísticas de item-total

COMPONENTE 2	Média de escala se o item for excluído	Variância de escala se o item for excluído	Correlação de item total corrigida	*Alfa de Cronbach* se o item for excluído
Compreensão	4,0600	1,087	,479	.
Competência Interacional	4,7940	,248	,479	.

APÊNDICE 3.6 – Coeficiente *Kappa*: população de estudo

3.6.1 Kappa Edição 1

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	181	55	5	1	1	243
		Contagem Esperada	16,9	40,5	75,2	68,8	41,6	243,0
		% em Nível Observador	74,5%	22,6%	2,1%	0,4%	0,4%	100,0%
		% em Nível Entrevistador	75,1%	9,5%	0,5%	0,1%	0,2%	7,0%
		% do Total	5,2%	1,6%	0,1%	0,0%	0,0%	7,0%
	Intermediário	Contagem	49	323	137	9	5	523
		Contagem Esperada	36,5	87,2	161,8	148,0	89,6	523,0
		% em Nível Observador	9,4%	61,8%	26,2%	1,7%	1,0%	100,0%
		% em Nível Entrevistador	20,3%	56,1%	12,8%	0,9%	0,8%	15,1%
		% do Total	1,4%	9,3%	4,0%	0,3%	0,1%	15,1%
	Intermediário Superior	Contagem	10	169	485	106	5	775
		Contagem Esperada	54,0	129,2	239,7	219,3	132,8	775,0
		% em Nível Observador	1,3%	21,8%	62,6%	13,7%	0,6%	100,0%
		% em Nível Entrevistador	4,1%	29,3%	45,4%	10,8%	0,8%	22,4%
		% do Total	0,3%	4,9%	14,0%	3,1%	0,1%	22,4%
	Avançado	Contagem	1	27	374	451	52	905
		Contagem Esperada	63,1	150,8	279,9	256,1	155,0	905,0
		% em Nível Observador	0,1%	3,0%	41,3%	49,8%	5,7%	100,0%
		% em Nível Entrevistador	0,4%	4,7%	35,0%	46,1%	8,8%	26,2%
		% do Total	0,0%	0,8%	10,8%	13,0%	1,5%	26,2%
	Avançado Superior	Contagem	0	2	68	411	529	1010
		Contagem Esperada	70,4	168,3	312,4	285,8	173,0	1010,0
		% em Nível Observador	0,0%	0,2%	6,7%	40,7%	52,4%	100,0%
		% em Nível Entrevistador	0,0%	0,3%	6,4%	42,0%	89,4%	29,2%
		% do Total	0,0%	0,1%	2,0%	11,9%	15,3%	29,2%
Total		Contagem	241	576	1069	978	592	3456
		Contagem Esperada	241,0	576,0	1069,0	978,0	592,0	3456,0
		% em Nível Observador	7,0%	16,7%	30,9%	28,3%	17,1%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	7,0%	16,7%	30,9%	28,3%	17,1%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,446	,011	50,182	,000
Nº de Casos Válidos		3456			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.2 Kappa Edição 2

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	155	62	17	4	0	238
		Contagem Esperada	12,7	36,7	76,6	71,7	40,2	238,0
		% em Nível Observador	65,1%	26,1%	7,1%	1,7%	0,0%	100,0%
		% em Nível Entrevistador	69,8%	9,7%	1,3%	0,3%	0,0%	5,7%
		% do Total	3,7%	1,5%	0,4%	0,1%	0,0%	5,7%
	Intermediário	Contagem	46	320	141	13	2	522
		Contagem Esperada	27,8	80,5	168,0	157,4	88,3	522,0
		% em Nível Observador	8,8%	61,3%	27,0%	2,5%	0,4%	100,0%
		% em Nível Entrevistador	20,7%	49,8%	10,5%	1,0%	0,3%	12,5%
		% do Total	1,1%	7,7%	3,4%	0,3%	0,0%	12,5%
	Intermediário Superior	Contagem	14	193	590	112	11	920
		Contagem Esperada	49,1	141,9	296,1	277,3	155,6	920,0
		% em Nível Observador	1,5%	21,0%	64,1%	12,2%	1,2%	100,0%
		% em Nível Entrevistador	6,3%	30,1%	44,0%	8,9%	1,6%	22,1%
		% do Total	0,3%	4,6%	14,2%	2,7%	0,3%	22,1%
	Avançado	Contagem	6	56	484	555	59	1160
		Contagem Esperada	61,9	178,9	373,4	349,7	196,2	1160,0
		% em Nível Observador	0,5%	4,8%	41,7%	47,8%	5,1%	100,0%
		% em Nível Entrevistador	2,7%	8,7%	36,1%	44,2%	8,4%	27,9%
		% do Total	0,1%	1,3%	11,6%	13,3%	1,4%	27,9%
	Avançado Superior	Contagem	1	11	108	571	632	1323
		Contagem Esperada	70,6	204,0	425,9	398,8	223,7	1323,0
		% em Nível Observador	0,1%	0,8%	8,2%	43,2%	47,8%	100,0%
		% em Nível Entrevistador	0,5%	1,7%	8,1%	45,5%	89,8%	31,8%
		% do Total	0,0%	0,3%	2,6%	13,7%	15,2%	31,8%
Total		Contagem	222	642	1340	1255	704	4163
		Contagem Esperada	222,0	642,0	1340,0	1255,0	704,0	4163,0
		% em Nível Observador	5,3%	15,4%	32,2%	30,1%	16,9%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	5,3%	15,4%	32,2%	30,1%	16,9%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,403	,010	49,119	,000
Nº de Casos Válidos		4163			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.3 Kappa Edição 3

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	130	63	6	1	0	200
		Contagem Esperada	8,6	29,4	61,6	59,4	41,0	200,0
		% em Nível Observador	65,0%	31,5%	3,0%	0,5%	0,0%	100,0%
		% em Nível Entrevistador	67,0%	9,5%	0,4%	0,1%	0,0%	4,4%
		% do Total	2,9%	1,4%	0,1%	0,0%	0,0%	4,4%
	Intermediário	Contagem	47	329	152	13	0	541
		Contagem Esperada	23,3	79,6	166,5	160,8	110,9	541,0
		% em Nível Observador	8,7%	60,8%	28,1%	2,4%	0,0%	100,0%
		% em Nível Entrevistador	24,2%	49,5%	10,9%	1,0%	0,0%	12,0%
		% do Total	1,0%	7,3%	3,4%	0,3%	0,0%	12,0%
	Intermediário Superior	Contagem	14	213	683	137	6	1053
		Contagem Esperada	45,3	154,9	324,1	312,9	215,8	1053,0
		% em Nível Observador	1,3%	20,2%	64,9%	13,0%	0,6%	100,0%
		% em Nível Entrevistador	7,2%	32,1%	49,2%	10,2%	0,6%	23,3%
		% do Total	0,3%	4,7%	15,1%	3,0%	0,1%	23,3%
	Avançado	Contagem	2	51	431	648	72	1204
		Contagem Esperada	51,8	177,1	370,6	357,8	246,8	1204,0
		% em Nível Observador	0,2%	4,2%	35,8%	53,8%	6,0%	100,0%
		% em Nível Entrevistador	1,0%	7,7%	31,0%	48,3%	7,8%	26,7%
		% do Total	0,0%	1,1%	9,6%	14,4%	1,6%	26,7%
	Avançado Superior	Contagem	1	8	117	542	847	1515
		Contagem Esperada	65,1	222,9	466,3	450,2	310,5	1515,0
		% em Nível Observador	0,1%	0,5%	7,7%	35,8%	55,9%	100,0%
		% em Nível Entrevistador	0,5%	1,2%	8,4%	40,4%	91,6%	33,6%
		% do Total	0,0%	0,2%	2,6%	12,0%	18,8%	33,6%
Total		Contagem	194	664	1389	1341	925	4513
		Contagem Esperada	194,0	664,0	1389,0	1341,0	925,0	4513,0
		% em Nível Observador	4,3%	14,7%	30,8%	29,7%	20,5%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	4,3%	14,7%	30,8%	29,7%	20,5%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,453	,010	56,007	,000
Nº de Casos Válidos		4513			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.4 Kappa Edição 4

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	71	55	6	2	1	135
		Contagem Esperada	3,8	19,4	41,9	43,2	26,8	135,0
		% em Nível Observador	52,6%	40,7%	4,4%	1,5%	0,7%	100,0%
		% em Nível Entrevistador	57,3%	8,6%	0,4%	0,1%	0,1%	3,0%
		% do Total	1,6%	1,2%	0,1%	0,0%	0,0%	3,0%
	Intermediário	Contagem	32	352	155	15	3	557
		Contagem Esperada	15,5	79,9	173,1	178,1	110,4	557,0
		% em Nível Observador	5,7%	63,2%	27,8%	2,7%	0,5%	100,0%
		% em Nível Entrevistador	25,8%	55,2%	11,2%	1,1%	0,3%	12,5%
		% do Total	0,7%	7,9%	3,5%	0,3%	0,1%	12,5%
	Intermediário Superior	Contagem	15	183	660	142	10	1010
		Contagem Esperada	28,2	144,9	313,8	322,9	200,3	1010,0
		% em Nível Observador	1,5%	18,1%	65,3%	14,1%	1,0%	100,0%
		% em Nível Entrevistador	12,1%	28,7%	47,8%	10,0%	1,1%	22,7%
		% do Total	0,3%	4,1%	14,8%	3,2%	0,2%	22,7%
	Avançado	Contagem	2	35	466	626	71	1200
		Contagem Esperada	33,5	172,1	372,8	383,6	237,9	1200,0
		% em Nível Observador	0,2%	2,9%	38,8%	52,2%	5,9%	100,0%
		% em Nível Entrevistador	1,6%	5,5%	33,7%	44,0%	8,0%	27,0%
		% do Total	0,0%	0,8%	10,5%	14,1%	1,6%	27,0%
	Avançado Superior	Contagem	4	13	95	637	797	1546
		Contagem Esperada	43,1	221,8	480,3	494,2	306,6	1546,0
		% em Nível Observador	0,3%	0,8%	6,1%	41,2%	51,6%	100,0%
		% em Nível Entrevistador	3,2%	2,0%	6,9%	44,8%	90,4%	34,8%
		% do Total	0,1%	0,3%	2,1%	14,3%	17,9%	34,8%
Total		Contagem	124	638	1382	1422	882	4448
		Contagem Esperada	124,0	638,0	1382,0	1422,0	882,0	4448,0
		% em Nível Observador	2,8%	14,3%	31,1%	32,0%	19,8%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	2,8%	14,3%	31,1%	32,0%	19,8%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,422	,010	51,124	,000
Nº de Casos Válidos		4448			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.5 Kappa Edição 5

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	166	64	17	6	1	254
		Contagem Esperada	14,0	40,6	77,4	74,4	47,6	254,0
		% em Nível Observador	65,4%	25,2%	6,7%	2,4%	0,4%	100,0%
		% em Nível Entrevistador	65,6%	8,7%	1,2%	0,4%	0,1%	5,5%
		% do Total	3,6%	1,4%	0,4%	0,1%	0,0%	5,5%
	Intermediário	Contagem	69	330	131	18	0	548
		Contagem Esperada	30,2	87,5	167,0	160,5	102,8	548,0
		% em Nível Observador	12,6%	60,2%	23,9%	3,3%	0,0%	100,0%
		% em Nível Entrevistador	27,3%	45,1%	9,4%	1,3%	0,0%	12,0%
		% do Total	1,5%	7,2%	2,9%	0,4%	0,0%	12,0%
	Intermediário Superior	Contagem	14	260	620	133	10	1037
		Contagem Esperada	57,2	165,6	316,0	303,7	194,5	1037,0
		% em Nível Observador	1,4%	25,1%	59,8%	12,8%	1,0%	100,0%
		% em Nível Entrevistador	5,5%	35,5%	44,4%	9,9%	1,2%	22,6%
		% do Total	0,3%	5,7%	13,5%	2,9%	0,2%	22,6%
	Avançado	Contagem	3	73	519	632	71	1298
		Contagem Esperada	71,6	207,2	395,5	380,2	243,5	1298,0
		% em Nível Observador	0,2%	5,6%	40,0%	48,7%	5,5%	100,0%
		% em Nível Entrevistador	1,2%	10,0%	37,2%	47,1%	8,3%	28,3%
		% do Total	0,1%	1,6%	11,3%	13,8%	1,5%	28,3%
	Avançado Superior	Contagem	1	5	110	554	778	1448
		Contagem Esperada	79,9	231,2	441,2	424,1	271,6	1448,0
		% em Nível Observador	0,1%	0,3%	7,6%	38,3%	53,7%	100,0%
		% em Nível Entrevistador	0,4%	0,7%	7,9%	41,3%	90,5%	31,6%
		% do Total	0,0%	0,1%	2,4%	12,1%	17,0%	31,6%
Total		Contagem	253	732	1397	1343	860	4585
		Contagem Esperada	253,0	732,0	1397,0	1343,0	860,0	4585,0
		% em Nível Observador	5,5%	16,0%	30,5%	29,3%	18,8%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	5,5%	16,0%	30,5%	29,3%	18,8%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,414	,010	52,524	,000
Nº de Casos Válidos		4585			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.6 Kappa Edição 6

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	144	62	2	1	0	209
		Contagem Esperada	9,7	30,0	68,6	65,2	35,5	209,0
		% em Nível Observador	68,9%	29,7%	1,0%	0,5%	0,0%	100,0%
		% em Nível Entrevistador	65,8%	9,2%	0,1%	0,1%	0,0%	4,4%
		% do Total	3,1%	1,3%	0,0%	0,0%	0,0%	4,4%
	Intermediário	Contagem	60	341	137	7	0	545
		Contagem Esperada	25,3	78,1	178,9	170,1	92,5	545,0
		% em Nível Observador	11,0%	62,6%	25,1%	1,3%	0,0%	100,0%
		% em Nível Entrevistador	27,4%	50,5%	8,9%	0,5%	0,0%	11,6%
		% do Total	1,3%	7,2%	2,9%	0,1%	0,0%	11,6%
	Intermediário Superior	Contagem	8	246	767	123	1	1145
		Contagem Esperada	53,3	164,1	375,9	357,4	194,3	1145,0
		% em Nível Observador	0,7%	21,5%	67,0%	10,7%	0,1%	100,0%
		% em Nível Entrevistador	3,7%	36,4%	49,6%	8,4%	0,1%	24,3%
		% do Total	0,2%	5,2%	16,3%	2,6%	0,0%	24,3%
	Avançado	Contagem	5	22	595	764	69	1455
		Contagem Esperada	67,7	208,6	477,7	454,2	246,9	1455,0
		% em Nível Observador	0,3%	1,5%	40,9%	52,5%	4,7%	100,0%
		% em Nível Entrevistador	2,3%	3,3%	38,5%	52,0%	8,6%	30,9%
		% do Total	0,1%	0,5%	12,6%	16,2%	1,5%	30,9%
	Avançado Superior	Contagem	2	4	45	575	729	1355
		Contagem Esperada	63,0	194,2	444,9	423,0	229,9	1355,0
		% em Nível Observador	0,1%	0,3%	3,3%	42,4%	53,8%	100,0%
		% em Nível Entrevistador	0,9%	0,6%	2,9%	39,1%	91,2%	28,8%
		% do Total	0,0%	0,1%	1,0%	12,2%	15,5%	28,8%
Total		Contagem	219	675	1546	1470	799	4709
		Contagem Esperada	219,0	675,0	1546,0	1470,0	799,0	4709,0
		% em Nível Observador	4,7%	14,3%	32,8%	31,2%	17,0%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	4,7%	14,3%	32,8%	31,2%	17,0%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,448	,010	56,311	,000
Nº de Casos Válidos		4709			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.6.7 Kappa Edição 7

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	95	41	5	0	0	141
		Contagem Esperada	5,9	16,8	40,1	44,1	34,0	141,0
		% em Nível Observador	67,4%	29,1%	3,5%	0,0%	0,0%	100,0%
		% em Nível Entrevistador	57,2%	8,7%	0,4%	0,0%	0,0%	3,6%
		% do Total	2,4%	1,0%	0,1%	0,0%	0,0%	3,6%
	Intermediário	Contagem	58	229	107	6	1	401
		Contagem Esperada	16,8	47,8	114,0	125,6	96,8	401,0
		% em Nível Observador	14,5%	57,1%	26,7%	1,5%	0,2%	100,0%
		% em Nível Entrevistador	34,9%	48,5%	9,5%	0,5%	0,1%	10,1%
		% do Total	1,5%	5,8%	2,7%	0,2%	0,0%	10,1%
	Intermediário Superior	Contagem	8	186	503	108	2	807
		Contagem Esperada	33,9	96,3	229,4	252,7	194,8	807,0
		% em Nível Observador	1,0%	23,0%	62,3%	13,4%	0,2%	100,0%
		% em Nível Entrevistador	4,8%	39,4%	44,7%	8,7%	0,2%	20,4%
		% do Total	0,2%	4,7%	12,7%	2,7%	0,1%	20,4%
	Avançado	Contagem	5	10	459	546	105	1125
		Contagem Esperada	47,2	134,2	319,8	352,3	271,5	1125,0
		% em Nível Observador	0,4%	0,9%	40,8%	48,5%	9,3%	100,0%
		% em Nível Entrevistador	3,0%	2,1%	40,8%	44,1%	11,0%	28,4%
		% do Total	0,1%	0,3%	11,6%	13,8%	2,7%	28,4%
	Avançado Superior	Contagem	0	6	51	579	847	1483
		Contagem Esperada	62,2	176,9	421,6	464,4	357,9	1483,0
		% em Nível Observador	0,0%	0,4%	3,4%	39,0%	57,1%	100,0%
		% em Nível Entrevistador	0,0%	1,3%	4,5%	46,7%	88,7%	37,5%
		% do Total	0,0%	0,2%	1,3%	14,6%	21,4%	37,5%
Total		Contagem	166	472	1125	1239	955	3957
		Contagem Esperada	166,0	472,0	1125,0	1239,0	955,0	3957,0
		% em Nível Observador	4,2%	11,9%	28,4%	31,3%	24,1%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	4,2%	11,9%	28,4%	31,3%	24,1%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,414	,010	46,741	,000
Nº de Casos Válidos		3957			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

APÊNDICE 3.7 – Coeficiente *Kappa*: Edição 5

3.7.1 Kappa Amostra B primeira instância

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	68	27	7	6	1	109
		Contagem Esperada	19,3	39,6	40,4	8,0	1,6	109,0
		% em Nível Observador	62,4%	24,8%	6,4%	5,5%	0,9%	100,0%
		% em Nível Entrevistador	52,3%	10,2%	2,6%	11,1%	9,1%	14,9%
		% do Total	9,3%	3,7%	1,0%	0,8%	0,1%	14,9%
	Intermediário	Contagem	44	83	22	11	0	160
		Contagem Esperada	28,4	58,1	59,4	11,8	2,4	160,0
		% em Nível Observador	27,5%	51,9%	13,8%	6,9%	0,0%	100,0%
		% em Nível Entrevistador	33,8%	31,2%	8,1%	20,4%	0,0%	21,8%
		% do Total	6,0%	11,3%	3,0%	1,5%	0,0%	21,8%
	Intermediário Superior	Contagem	14	79	87	5	10	195
		Contagem Esperada	34,6	70,8	72,4	14,4	2,9	195,0
		% em Nível Observador	7,2%	40,5%	44,6%	2,6%	5,1%	100,0%
		% em Nível Entrevistador	10,8%	29,7%	32,0%	9,3%	90,9%	26,6%
		% do Total	1,9%	10,8%	11,9%	0,7%	1,4%	26,6%
	Avançado	Contagem	3	72	85	21	0	181
		Contagem Esperada	32,1	65,7	67,2	13,3	2,7	181,0
		% em Nível Observador	1,7%	39,8%	47,0%	11,6%	0,0%	100,0%
		% em Nível Entrevistador	2,3%	27,1%	31,3%	38,9%	0,0%	24,7%
		% do Total	0,4%	9,8%	11,6%	2,9%	0,0%	24,7%
	Avançado Superior	Contagem	1	5	71	11	0	88
		Contagem Esperada	15,6	31,9	32,7	6,5	1,3	88,0
		% em Nível Observador	1,1%	5,7%	80,7%	12,5%	0,0%	100,0%
		% em Nível Entrevistador	0,8%	1,9%	26,1%	20,4%	0,0%	12,0%
		% do Total	0,1%	0,7%	9,7%	1,5%	0,0%	12,0%
Total		Contagem	130	266	272	54	11	733
		Contagem Esperada	130,0	266,0	272,0	54,0	11,0	733,0
		% em Nível Observador	17,7%	36,3%	37,1%	7,4%	1,5%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	17,7%	36,3%	37,1%	7,4%	1,5%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,166	,022	8,991	,000
Nº de Casos Válidos		733			

a. Não assumindo a hipótese nula.

b. Uso de erro padrão assintótico considerando a hipótese nula.

3.7.2 Kappa Amostra B segunda instância

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	7	8	4	1	0	20
		Contagem Esperada	1,2	4,0	7,1	6,0	1,6	20,0
		% em Nível Observador	35,0%	40,0%	20,0%	5,0%	0,0%	100,0%
		% em Nível Entrevistador	15,6%	5,4%	1,5%	0,5%	0,0%	2,7%
		% do Total	1,0%	1,1%	0,5%	0,1%	0,0%	2,7%
	Intermediário	Contagem	11	24	31	7	0	73
		Contagem Esperada	4,5	14,7	26,1	21,8	5,9	73,0
		% em Nível Observador	15,1%	32,9%	42,5%	9,6%	0,0%	100,0%
		% em Nível Entrevistador	24,4%	16,2%	11,8%	3,2%	0,0%	10,0%
		% do Total	1,5%	3,3%	4,2%	1,0%	0,0%	10,0%
	Intermediário Superior	Contagem	15	60	94	36	4	209
		Contagem Esperada	12,8	42,2	74,7	62,4	16,8	209,0
		% em Nível Observador	7,2%	28,7%	45,0%	17,2%	1,9%	100,0%
		% em Nível Entrevistador	33,3%	40,5%	35,9%	16,4%	6,8%	28,5%
		% do Total	2,0%	8,2%	12,8%	4,9%	0,5%	28,5%
	Avançado	Contagem	10	44	86	94	20	254
		Contagem Esperada	15,6	51,3	90,8	75,9	20,4	254,0
		% em Nível Observador	3,9%	17,3%	33,9%	37,0%	7,9%	100,0%
		% em Nível Entrevistador	22,2%	29,7%	32,8%	42,9%	33,9%	34,7%
		% do Total	1,4%	6,0%	11,7%	12,8%	2,7%	34,7%
	Avançado Superior	Contagem	2	12	47	81	35	177
		Contagem Esperada	10,9	35,7	63,3	52,9	14,2	177,0
		% em Nível Observador	1,1%	6,8%	26,6%	45,8%	19,8%	100,0%
		% em Nível Entrevistador	4,4%	8,1%	17,9%	37,0%	59,3%	24,1%
		% do Total	0,3%	1,6%	6,4%	11,1%	4,8%	24,1%
Total		Contagem	45	148	262	219	59	733
		Contagem Esperada	45,0	148,0	262,0	219,0	59,0	733,0
		% em Nível Observador	6,1%	20,2%	35,7%	29,9%	8,0%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	6,1%	20,2%	35,7%	29,9%	8,0%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,133	,022	6,701	,000
Nº de Casos Válidos		733			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.

3.7.3 Kappa Amostra D

			Nível Entrevistador					Total
			Básico	Intermediário	Intermediário Superior	Avançado	Avançado Superior	
Nível Observador	Básico	Contagem	98	37	10	0	0	145
		Contagem Esperada	4,6	17,5	42,3	48,5	32,0	145,0
		% em Nível Observador	67,6%	25,5%	6,9%	0,0%	0,0%	100,0%
		% em Nível Entrevistador	79,7%	7,9%	0,9%	0,0%	0,0%	3,8%
		% do Total	2,5%	1,0%	0,3%	0,0%	0,0%	3,8%
	Intermediário	Contagem	25	247	109	7	0	388
		Contagem Esperada	12,4	46,9	113,3	129,8	85,5	388,0
		% em Nível Observador	6,4%	63,7%	28,1%	1,8%	0,0%	100,0%
		% em Nível Entrevistador	20,3%	53,0%	9,7%	0,5%	0,0%	10,1%
		% do Total	0,6%	6,4%	2,8%	0,2%	0,0%	10,1%
	Intermediário Superior	Contagem	0	181	533	128	0	842
		Contagem Esperada	26,9	101,9	245,9	281,8	185,6	842,0
		% em Nível Observador	0,0%	21,5%	63,3%	15,2%	0,0%	100,0%
		% em Nível Entrevistador	0,0%	38,8%	47,4%	9,9%	0,0%	21,9%
		% do Total	0,0%	4,7%	13,8%	3,3%	0,0%	21,9%
	Avançado	Contagem	0	1	434	611	71	1117
		Contagem Esperada	35,7	135,1	326,2	373,8	246,2	1117,0
		% em Nível Observador	0,0%	0,1%	38,9%	54,7%	6,4%	100,0%
		% em Nível Entrevistador	0,0%	0,2%	38,6%	47,4%	8,4%	29,0%
		% do Total	0,0%	0,0%	11,3%	15,9%	1,8%	29,0%
	Avançado Superior	Contagem	0	0	39	543	778	1360
		Contagem Esperada	43,4	164,5	397,2	455,1	299,8	1360,0
		% em Nível Observador	0,0%	0,0%	2,9%	39,9%	57,2%	100,0%
		% em Nível Entrevistador	0,0%	0,0%	3,5%	42,1%	91,6%	35,3%
		% do Total	0,0%	0,0%	1,0%	14,1%	20,2%	35,3%
Total		Contagem	123	466	1125	1289	849	3852
		Contagem Esperada	123,0	466,0	1125,0	1289,0	849,0	3852,0
		% em Nível Observador	3,2%	12,1%	29,2%	33,5%	22,0%	100,0%
		% em Nível Entrevistador	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
		% do Total	3,2%	12,1%	29,2%	33,5%	22,0%	100,0%

Medidas Simétricas

		Valor	Erro Padronizado Assintótico[a]	T Aproximado[b]	Significância Aproximada
Medida de concordância	Kappa	,450	,011	49,915	,000
Nº de Casos Válidos		3852			

a. Não assumindo a hipótese nula.
b. Uso de erro padrão assintótico considerando a hipótese nula.