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

Laura Márcia Luiza Ferreira

Avaliação da proficiência oral: uma análise fatorial e de discriminação dos itens do exame Celpe-Bras

Belo Horizonte

Abril de 2018

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

Laura Márcia Luiza Ferreira

Avaliação da proficiência oral: uma análise fatorial e de discriminação dos itens do exame Celpe-Bras

Tese apresentada ao Programa de Pós-Graduação em Estudos de Linguagens do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) como requisito parcial para a obtenção do título de Doutora em Estudos de Linguagens.

Orientador: Prof. Dr. Jerônimo Coura Sobrinho

Belo Horizonte

Abril de 2018

Ferreira, Laura Márcia Luiza.
F368a Avaliação da proficiência oral: uma análise fatorial e de discriminação de itens do exame Celpe-Bras / Laura Márcia Luiza Ferreira. - 2018.
242 f. : il., grafs., tabs.
Orientador: Jerônimo Coura Sobrinho

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

1. Língua portuguesa - Testes de aptidão. 2. Proficiência oral - Avaliação. 3. Análise fatorial. 4. Teoria de resposta ao item. I. Sobrinho, Jerônimo Coura. II. Título.

CDD: 469.824

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

CEFET-MG

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

LAURA MÁRCIA LUIZA FERREIRA

Avaliação da proficiência oral: uma análise fatorial e de discriminação dos itens do exame Celpe-Bras

Tese apresentada ao Programa de Pós-Graduação em Estudos de Linguagens do Centro Federal de Educação Tecnológica de Minas Gerais em 02 de abril de 2018, como requisito parcial para obtenção do título de Doutor em Estudos de Linguagens, aprovada pela Banca Examinadora constituída pelos professores:

Prof. Jerônimo Coura Sobrinho, Dr. - Orientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Cristiano Mauro Assis Gomes, Dr.
Universidade Federal de Minas Gerais

Prof. Frederico Neves Condé, Dr.
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Prof. Luiz Antônio dos Prazeres, Dr.
Universidade Federal de Ouro Preto

Prof. Renato Caixeta da Silva, Dr.
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Vicente Aguimar Parreiras, Dr. - Suplente
Centro Federal de Educação Tecnológica de Minas Gerais

Profª. Ana Maria Nápoles Villela, Drª. - Suplente
Centro Federal de Educação Tecnológica de Minas Gerais

Agradecimentos

Aos servidores do CEFET-MG, especialmente aos docentes e aos servidores envolvidos no curso de Pós-Graduação em Estudos de Linguagens, por fornecerem as condições de realização deste trabalho;

À UNILA, agradeço aos colegas professores e técnicos, que apoiaram minha licença para capacitação para que pudesse cumprir com alguns dos requisitos do curso de doutorado;

Ao INEP, por fornecer os dados de análise deste trabalho;

Ao meu orientador, Prof. Dr. Jerônimo Coura-Sobrinho, pelo incentivo e confiança;

Ao Leonard Assis, pela paciência e ajuda com os cálculos que apresento neste trabalho;

Aos professores Dr. Renato Caixeta da Silva, Dr. Cristiano Mauro Assis Gomes, Dr. Frederico Neves Condé, Dr. Luiz Antônio dos Prazeres e Dra. Marcia Niederauer, pela leitura cuidadosa do trabalho;

Aos colegas do grupo INFORTEC, pela troca e incentivo;

À Reisila e Cleuza, por cuidarem da minha casa, enquanto me dedicava ao trabalho da tese;

À minha mãe, Marli Luiza Ferreira, por prover as condições para que eu pudesse chegar até aqui;

Ao Henrique, pela paciência e companhia;

À Elis e Eva, pela esperança renovada.

RESUMO: Como as avaliações podem admitir erros sistemáticos, os escores podem ser afetados, por isso é preciso validá-las por meio da análise de evidências variadas. Na perspectiva de Messick (1987), a validade é um conceito único, que consiste em avaliar evidências empíricas e teóricas sobre a pertinência das inferências feitas a partir da nota de um teste. O objetivo do presente trabalho é analisar as escalas de avaliação da prova oral do exame Celpe-Bras forma a coletar evidências da validade interna dos instrumentos. Inicialmente, argumento que a prova oral é uma tarefa cuja situação é a entrevista de proficiência oral. Em seguida, apresento a Teoria de Validade proposta por Messick (1987) e a relaciono com a maneira como o conceito tem sido definido por especialistas em avaliação de línguas. Para analisar os sete itens que compõem as duas escalas de avaliação da prova oral, quanto à sua dimensionalidade, apresento uma análise fatorial exploratória. Quanto ao ajuste e à quantidade de informação de cada item, apresento uma análise de discriminação de itens, por meio do modelo Rasch básico na extensão *Partial Credit Model.* A escala do avaliador-intelocutor é composta por um item e a escala do avaliador-observador contém seis itens, a saber: c*ompreensão, competência interacional, fluência, adequação lexical, adequação gramatical* e *pronúncia.* O conjunto de dados analisados nesta pesquisa é composto por notas atribuídas para 1.000 participantes que se submeteram ao exame na primeira edição de 2016. O resultado da análise fatorial sugere que a nota da prova oral seja uma medida unidimensional. Os valores de peso dos itens da escala, do maior para o menor, são de 0.36 para *nota do avaliador-interlocutor,* 0.19 para *adequação lexical*, 0.18 para *fluência,* 0.13 para *adequação gramatical,* seguidas dos itens *competência interacional, pronúncia* e *compreensão* com os valores de 0.09, 0.06 e 0.04, respectivamente. A partir da análise de discriminação de itens, os valores da média quadrada de *infit* variou de 0.431 a 1.386, sugerindo bom ajuste dos itens. Porém, quanto ao valor da média quadrada de outfit, o item *compreensão* apresentou um valor acima de 2.0, apontando necessidade de revisão do item na escala. Ao considerar intervalos entre os valores de *threshold* dos sete itens das escalas, posso afirmar que uma mesma faixa de nota em cada um dos itens pode não discriminar da mesma forma o mesmo perfil de examinandos. Com base na análise fatorial e na análise de discriminação de itens, discuto uma proposta de mudança de peso dos itens na composição da nota oral, especialmente quanto ao item compreensão, bem como a necessidade de investimento na revisão dos descritores das escalas, da tarefa e da situação de entrevista de proficiência oral.

Palavras-chave: avaliação em línguas, validade, Celpe-Bras, análise fatorial, teoria de resposta ao item, Rasch

ABSTRACT: Since sistematic errors may affect scores, assessments are imperfect samples of constructs to be validated. According to Messick's Validation Theory (1987), assessment validation is a matter of gathering multiple sources of empirical evidence and theoretical rationales to support the inferences that will be made from the tests results. In order to gather evidences to discuss dimensionality, item fit and information on the oral Celpe-Bras exam scales, I perform a exploratory factor analysis and an item discrimination analyses, using Rasch's Partial Credit Model. First, I describe the oral exam task as an oral proficiency interview. Then I discuss the Messick's Validation Theory and it's influence on applied linguistics and on second language assessment theoretical works. It was analysed 1.000 examinee's scores from the first exam edition of 2016. The oral Celpe-Bras's score model is organized in seven itens in two scales: one rated by the interviewer and six itens by the rater itself. The interviewer scores one single note for the performance while the rater scores one note for the following itens *comprehension, interactional competence, fluency, lexical adequacy, grammatical adequacy* and *pronunciation*. After running the exploratory factor analysis, we concluded that the test score is an one-dimensional measure. According to the weight values, the itens that weighed the most on the fator oral proficiency were the *interviewer's score* (0.36), followed by *lexical adequacy* (0.19), *fluency (*0.18) and *grammatical adequacy* (0.13). *Interactional competence, pronunciation and comprehension* weighted 0.09, 0.06 and 0.04, respectively. According to Rasch's analyses, the infit msq values for the itens was reasonable (0.431 to 1.386). However, outfit msq value for the item *comprehension* was >2.0, sugesting poor fit. Since the threshold's value categories were not homogeneous through the scales, the itens may not distinguish in the same way examinees from the same profile ability. Finally, I discuss a score's composite proposal, the relevance of iten *comprehension* in this oral situation as well as the investment on descriptor's and task's review.

Keywords: second language assessment, Celpe-Bras, factorial analyses, Rasch

LISTA DE ELEMENTOS GRÁFICOS

SUMÁRIO

1. INTRODUÇÃO

A avaliação[1] é um processo que tem como objetivo a coleta de informações para gerar subsídios para que os avaliadores interpretem a capacidade potencial sobre o uso ou conhecimento de umalgum determinado domínio por parte dos examinandos. O contexto de coleta das informações em uma situação de avaliação é controlado ou padronizado de forma a permitir que os examinandos tenham mais ou menos as mesmas chances de demonstrarem o que sabem. A partir das respostas ao teste, os avaliadores inferem ou julgam a proficiência, a habilidade, o conhecimento ou a capacidade dos examinandos.

As teorias relacionadas à avaliação de línguas ou à análise de algum instrumento específico podem mobilizar tanto conhecimentos da área da Linguística Aplicada ao ensino de línguas, quanto de outras áreas, como a da Psicometria, campo de estudos da Psicologia que analisa o significado das medidas e, para tanto, se apoia na Teoria da Medida e em metodologias estatísticas. Neste sentido, o estudo de avaliações de línguas é potencialmente interdisciplinar, uma vez que outros campos do conhecimento contribuem para o debate sobre a qualidade dos instrumentos. McNamara (2000) compara o complexo processo de elaboração de testes de línguas à criação de carros; em ambos os casos, os elaboradores devem submeter a invenção a diversas análises para avaliar possíveis fontes de inconsistências.

[1] Neste trabalho os termos *avaliação, teste* e *exame* serão usados como sinônimos.

Com relação à institucionalização do campo de estudos de avaliação de línguas, os pesquisadores que publicam sobre esse tema em língua inglesa têm espaços de debate sobre seus estudos em periódicos e eventos específicos, e podem se filiar a associações como a *International Language Testing Association* (ILTA). No Brasil, em eventos mais gerais sobre Linguística Aplicada ou sobre Avaliação Educacional, trabalhos sobre avaliação de línguas são bem-vindos. O evento mais específico para pesquisas sobre o tópico é o Simpósio Internacional sobre Celpe-Bras (Sincelpe), destinado ao debate de estudos que tenham como objeto de pesquisa o exame de Certificação de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). Embora haja indícios de que o campo de avaliação de línguas esteja se institucionalizando, McNamara (2004) afirma que tais estudos ocupam um lugar marginal na área da Linguística Aplicada ao ser considerado como um aspecto do ensino de línguas. Os estudos sobre avaliação de línguas estiveram na linha de frente do desenvolvimento de conceitos fundamentais no âmbito da perspectiva comunicativa do ensino de línguas. Como os testes deveriam operacionalizar domínios e conceitos de linguagem que se quer medir, os instrumentos de avaliação funcionariam também para retroalimentar discussões teóricas sobre ensino de língua, e conceitos como o de proficiência linguística, dentre outros (BYGATE, 2009; MCNAMARA, 2004). McNamara (2004) apresenta ainda mais um argumento sobre as potencialidades dos estudos em avaliação: eles poderiam contribuir para o debate sobre as noções de validade e confiabilidade e sobre o processo de validação desses instrumentos, uma vez que a avaliação é usada em muitas pesquisas como instrumento de coleta de dados.

McNamara (2004), ao discorrer sobre o lugar da avaliação de línguas na área da Linguística Aplicada ao longo do tempo, aponta seu desenvolvimento a partir de 1950, em que os testes apresentaram fundamentação científica vindas principalmente da área da Linguística e de Psicologia. A partir do final da década de 1970, os métodos comunicativos impactaram no desenho dos testes de forma que a ênfase passou a ser dada ao desempenho. A emergência do debate sobre o ensino para fins específicos deu ainda mais impulso para que as avaliações fossem elaboradas de forma a refletir os desempenhos relacionados a algum uso específico da língua. Atualmente, segundo o autor, a área vive tempos de sofisticação técnica com adoção de metodologias e ferramentas estatísticas que permitem investigar variados aspectos da validade e da confiabilidade a partir das notas. O autor cita os modelos estatísticos desenvolvidos a partir da Teoria de Resposta ao Item (TRI) como um dos exemplos de ferramenta que permitem investigar e identificar prováveis componentes responsáveis por variações na nota.

Uma outra forma de entender o lugar dos estudos sobre avaliação de línguas nas discussões recentes é por meio do debate sobre o letramento em avaliação de línguas. O letramento em avaliação de línguas, segundo Taylor (2013), é um dentre os diversos domínios de uso da linguagem no âmbito acadêmico. Esta discussão surge em um momento de crescente trabalho na área da avaliação em línguas, o que pressupõe um contingente de pessoas envolvidas nos processos de elaboração, aplicação e pesquisa relacionadas aos testes. Luckesi (2011), no contexto do debate sobre a formação dos professores brasileiros - não só os de línguas mas também de outras áreas –, faz uma defesa da importância da aprendizagem sobre avaliação. Scarino (2013) sugere que a formação de professores deveria trabalhar a avaliação simultaneamente ao conteúdo objeto estudado, tanto como prática para transformar a avaliação em um benefício para o processo de ensino aprendizagem, como para desenvolver nos professores uma autocompreensão e conscientização da natureza do próprio fenômeno da avaliação, seu papel e suas práticas de professores avaliadores. A autora também sugere a inclusão, durante o processo de desenvolvimento do letramento em avaliação, a exploração de pesquisas sobre avaliação e dos conceitos de validade e confiabilidade. Tais conhecimentos, Scarino (2003) defende, devem fazer parte do repertório do fazer docente e não apenas dos especialistas em avaliação de línguas. Os autores que discutem o desenvolvimento do letramento de línguas tendem a concordar que a validade e a confiabilidade são conceitos centrais não só para elaboração e análise de exames de larga escala, como também para os testes voltados para o contexto de sala de aula.

Esta tese se encontra no campo de estudos de avaliação de línguas adicionais, e tem como objetivo a análise do modelo de atribuição de notas da prova oral do examinando que se submete ao processo de avaliação da proficiência por meio do exame de Certificação de Proficiência em Língua Portuguesa para Estrangeiros, doravante, Celpe-Bras. O trabalho está organizado de forma que, neste primeiro capítulo, apresentarei o contexto da certificação da proficiência em Língua Portuguesa a falantes de outras línguas emitido por meio do Celpe-Bras. Em seguida, descreverei as faixas de certificação do exame e detalharei a etapa de avaliação oral. Por fim, argumentarei que a avaliação oral no contexto do Celpe-Bras é uma situação de entrevista de proficiência oral na qual tarefas são propostas pelos elaboradores do exame para guiar a interação face a face entre *avaliador-interlocutor* e examinando. No segundo capítulo, o tema da entrevista de proficiência oral será discutido. Apresentarei um breve histórico das entrevistas de proficiência oral, bem como um debate sobre os fatores que podem interferir no julgamento do desempenho nesta situação específica. Finalizarei o capítulo expondo a complexidade de definir e

organizar parâmetros para a avaliação da proficiência oral. No terceiro capítulo, fundamentado na Teoria de Validade proposta por Messick (1987), apresentarei os conceitos de validade e confiabilidade, e buscarei inseri-los no debate proposto por Messick (1987) com algumas definições de validade elaboradas por especialistas da área de avaliação em línguas. Em seguida, no quarto capítulo, retomarei os percursos de pesquisa que culminaram nas análises que apresentarei neste trabalho tal como está e fundamentarei e discutirei com mais detalhes as escolhas metodológicas. No quinto capítulo, apresentarei as análises empíricas bem como as discussões que emergiram de seus resultados para, ao final do texto, apontar algumas considerações finais.

Tendo introduzido e apresentado de maneira geral o tema do estudo e como o texto está organizado, discutirei, a seguir, o contexto da pesquisa: o exame Celpe-Bras.

1.1. O CELPE-BRAS

O Celpe-Bras é o exame oficial do Governo do Brasil para avaliação e, se for o caso, comprovação de proficiência de estrangeiros em Língua Portuguesa. A certificação de proficiência atestada por meio do Celpe-Bras pode ser exigida dos estrangeiros que não tenham a Língua Portuguesa como língua materna. O certificado atesta a proficiência em Português de estudantes que queiram se candidatar a programas de cooperação educacional financiados pelo Governo do Brasil, tais como os Programas de Estudante Convênio – Graduação ou Programas de Estudante Convênio – Pós-Graduação e, ainda, de profissionais que queiram ter seus diplomas revalidados. O exame pode ser consideravelmente relevante (*high stake test*), uma vez que decisões para a vida do examinando podem ser tomadas a partir de seu resultado.

Segundo informações disponíveis no *site* do Inep (2017), o perfil dos inscritos no exame é formado majoritariamente por falantes de espanhol como língua materna. Corroborando com este perfil, Gonçalves e Christófolo (2012) apontam que o Brasil é um dos países mais procurados pelos estudantes colombianos, por exemplo. Segundo os autores, houve um aumento de 123% no número de vistos de estudantes concedidos pela Embaixada do Brasil em Bogotá no período de 2007 a 2010. De acordo com Schoffen (2013), 56% dos inscritos no processo de certificação aplicado em outubro de 2011 foram provenientes da América Latina. Isso é resultado do investimento feito pelo Estado Brasileiro na promoção do ensino de português na

América Latina e é neste contexto político que o exame Celpe-Bras se fortaleceu e expandiu seus postos de aplicação pela região. Os postos aplicadores do exame são normalmente instituições de ensino credenciadas a fazer a aplicação do exame. De acordo com Diniz (2012), o Celpe-Bras é um instrumento de política linguística promovido pelo Estado Brasileiro que historicamente já mantinha a Rede Brasileira de Ensino no Exterior da qual faz parte a manutenção e abertura de centros e institutos culturais para promoção da língua portuguesa falada no Brasil e da cultura nacional pelo mundo, especialmente América Latina. Em 2016 havia 93 postos aplicadores do exame Celpe-Bras, sendo 29 no Brasil, 31 na América Latina e 34 espalhados por outras regiões como África, Europa, Ásia e Oriente Médio. Aproximadamente 6 mil examinandos por edição se inscrevem no Celpe-Bras nos últimos cinco anos.

O exame foi instituído em 1994. A primeira aplicação ocorreu em 1998 e contou, segundo Coura-Sobrinho (2006), com 127 candidatos. O exame é aplicado duas vezes por ano, no primeiro e no segundo semestre. As inscrições são feitas pelo *site* do Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) mediante um cadastro *online* e o pagamento de uma taxa de inscrição para manutenção dos custos de aplicação, que varia de posto para posto de aplicação.

O Celpe-Bras é um exame que certifica a proficiência nos níveis ou faixas de certificação *intermediário*, *intermediário superior*, *avançado* e *avançado superior* e cada nível está relacionado a uma nota final ou escore que varia de 0 a 5 pontos. Os examinandos que obtêm uma nota igual ou inferior a 1.99 pontos não são certificados (BRASIL, 2016a). De maneira geral, a nota final de certificação é composta pela menor nota entre as etapas oral e escrita do exame. A parte oral é composta por sete itens e a prova escrita por quatro itens. Na prova oral, há dois examinadores, o avaliador-observador avalia seis itens, a saber: compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia; e o avaliador-interlocutor avalia de maneira holística o desenvolvimento do examinando atribuindo-lhe uma nota única. Trata-se de uma prova oral de sete itens ou parâmetros porque a nota final da prova oral é composta a partir de sete itens, sendo seis notas atribuídas pelo avaliador-observador e uma nota atribuída pelo avaliador-interlocuor. Na prova escrita, como uma nota é atribuída a cada um dos textos escritos pelos examinandos, trata-se de quatro itens que irão compor a nota da prova final da parte escrita. Na quadro 1, coloco a descrição da proficiência para cada faixa certificada pelo exame, a partir de informações apresentadas no Manual do Examinando e no site do Inep.

QUADRO 1

Descrição da proficiência para cada faixa certificada pelo exame Celpe-Bras

faixa	nota	descrição
intermediário	2 a 2,75	Conferido a examinandos/as que evidenciem domínio operacional parcial da Língua Portuguesa, e demonstrem ser capazes de compreender e produzir textos orais e escritos sobre assuntos limitados, em contextos conhecidos e situações do cotidiano, sendo admitidas, neste nível, inadequações e interferências da língua materna e/ou de outra(s) língua(s) estrangeira(s) mais frequentes em situações desconhecidas, não suficientes, entretanto, para comprometer a comunicação.
intermediário superior	2,76 a 3,50	Conferido a examinandos/as que preenchem as características descritas no nível Intermediário, com a diferença de que, neste nível, as inadequações e interferências de língua materna e/ou de outra(s) língua(s) estrangeira(s) na pronúncia e na escrita devem ser menos frequentes que naquele nível.
avançado	3,51 a 4,25	Conferido a examinandos/as que evidenciem domínio operacional amplo da Língua Portuguesa, e demonstrem ser capazes de compreender e produzir textos orais e escritos sobre assuntos variados em contextos conhecidos e desconhecidos, sendo admitidas, neste nível, inadequações ocasionais na comunicação, principalmente em contextos desconhecidos, não suficiente, entretanto, para comprometer a comunicação.
avançado superior	4,26 a 5	Conferido a examinandos/as que preencham todos os requisitos do nível Avançado, mas com inadequações menos frequentes do que naquele nível.

Fonte: Informações organizadas a partir do Manual do examinando (2015) e *site* do Inep.

De acordo com o Manual do examinando (2015), são características ou domínios do exame a ênfase no uso da língua, o uso de textos autênticos e a avaliação integrada da compreensão e da produção na modalidade oral e/ou escrita da Língua Portuguesa. Sobre o desenho do exame, o Manual do examinando afirma que se trata de um teste de desempenho, no qual a resposta às tarefas podem ser avaliadas e classificadas em distintas faixas de proficiência. Neste sentido, não fazem parte da avaliação do Celpe-Bras tarefas isoladamente desenhadas que acessam conhecimentos e habilidades para usar estruturas linguísticas da língua portuguesa de forma fragmentada e descontextualizada de uma situação comunicativa. Além disso, diferentemente do que acontece com alguns exames de línguas, as tarefas não são desenhadas para certificar uma faixa específica de proficiência. O exame é composto de tarefas por meio das quais os examinandos devem produzir textos orais e escritos que estejam adequados ao propósito, ao interlocutor e ao registro de linguagem especificados pelas situações de interação delimitadas nas tarefas (FURTOSO, 2011; SCARAMUCCI, 2008; SCHOFFEN, 2003). O objetivo das tarefas é o de estimular o engajamento do examinando em uma interação, seja por meio da escrita de um texto

cujos interlocutores e propósitos são definidos no comando da tarefa, seja por meio da oralidade em situação de interação face a face, na qual o examinando tenha que comentar e conversar sobre algum assunto publicado na mídia impressa brasileira, por exemplo. As tarefas são objeto de pesquisa e discussão de diversos autores da área da educação e da linguística aplicada ao ensino de línguas.

Nas próximas seções, argumentarei que a proposta de interação face a face da prova oral é uma tarefa cuja situação é a entrevista de proficiência oral.

A seguir, trato das tarefas do exame com foco na interação face a face.

1.2. A INTERAÇÃO FACE A FACE DO EXAME CELPE-BRAS

Tratarei a seguir da avaliação oral, objeto de estudo deste trabalho, porém antes de descrevê-la, convém apresentar brevemente a prova escrita.

A parte escrita do exame é composta por quatro tarefas de produção, cada tarefa gera uma nota que é atribuída a partir de uma grade holística. A prova escrita é compostar por 4 itens, sendo cada tarefa um item. Na primeira tarefa, o examinando é convidado a assistir um vídeo; na segunda, a escutar um áudio; e na terceira e na quarta, a ler textos. A partir da compreensão do vídeo, do áudio e dos textos impressos, o examinando deve interagir, por meio da escrita, conforme o comando de cada uma das quatro tarefas. A partir das respostas dessas tarefas, a avaliação é feita com foco na capacidade do examinando mobilizar conhecimentos para organizar um texto adequado textual e discursivamente. De acordo com o Manual do examinando (2015), os parâmetros de avaliação das adequações textuais estão relacionados com a coesão, a coerência, o uso adequado de recursos linguísticos e a clareza; por sua vez, os elementos discursivos estão associados à adequação do propósito especificado no comando da tarefa como o gênero discursivo e a interlocução. É importante ressaltar que tais parâmetros compõem a descrição de uma nota única; ou seja, a avaliação é feita do texto como um todo, de forma holística.[2] Para cada um desses quatro textos é atribuída uma nota em uma escala de 0 a 5 por dois avaliadores independentes durante os eventos de correção do exame. Caso as notas atribuídas a uma mesma tarefa por dois avaliadores independentes estejam discrepantes por mais de um ponto, um terceiro avaliador reavaliará o desempenho do examinando na tarefa e decidirá por uma nota (BRASIL,2016a). A partir da média aritmética da nota nas quatro tarefas escritas, calcula-se a nota final da parte escrita

[2] Para uma discussão mais aprofundada sobre a maneira de atribuição de notas da prova escrita do exame Celpe-Bras, consulte Schoffen (2009).

do exame, ou seja, a nota de cada uma das tarefas têm peso igual na composição da nota final da parte escrita (BRASIL,2016a). Na composição final da nota do exame, tanto a nota final da prova escrita quanto a nota final da prova oral são levadas em consideração, de maneira que a menor nota dentre as duas modalidades de prova é a que prevalecerá para fins de certificação da proficiência (BRASIL,2016a).

Quanto à avaliação oral, o Manual do examinando a define como uma conversa de 20 minutos entre avaliador-interlocutor (AI) e examinando, a partir das informações pessoais que constam na ficha de inscrição do examinando e sobre tópicos do cotidiano e de interesse geral (BRASIL, 2015). Na ficha de inscrição, o examinando pode informar sua nacionalidade, profissão, como aprendeu o português, o que gosta de fazer, etc. A parte oral é uma conversa dividida em duas etapas: a primeira, se dá a partir das informações que constam no questionário de inscrição do examinando, com duração de 5 minutos; e a segunda, motivada a partir da leitura de três elementos provocadores previamente selecionados pelo avaliador-interlocutor por aproximadamente 15 minutos (BRASIL, 2015). Após os primeiros cinco minutos de avaliação, os elementos provocadores são o ponto de partida para a interação, uma vez que têm o objetivo de provocar e pautar o assunto da conversa (BRASIL, 2015). Vale ressaltar que a grande maioria dos textos utilizados para compor os elementos provocadores são recortes de reportagem retirados de revistas e jornais impressos brasileiros (FERREIRA, 2012). No Manual do examinando (2015) afirma-se que são avaliadas a capacidade de produção e compreensão oral. Além do *avaliador-interlocutor*, a interação é avaliada também por um segundo avaliador, chamado de *avaliador-observador*. As duas avaliações de uma mesma interação são feitas de forma independente entre o *avaliador-observador* e *avaliador-interlocutor* que não devem trocar impressões sobre o julgamento das notas (2016a). A avaliação é feita em fichas separadas e são enviadas ao Inep para que seja calculada uma nota única final da prova oral, com base no julgamento independente feito pelo *avaliador-observador* e pelo *avaliador-interlocutor*. Segundo o edital de inscrições (2016a), a nota final da prova oral é calculada a partir de uma média entre as duas notas, ou seja, cada nota tem um peso de 50% na composição da nota final oral da prova. Caso as notas atribuídas entre os avaliadores sejam divergente por mais de um ponto e meio (1.5), a interação é reavaliada em eventos de correção do exame. Além disso, a prova oral pode ser reavaliada por um terceiro avaliador, quando a nota da prova oral é discrepante em até dois pontos em relação à nota da prova escrita, ou quando a diferença de notas entre as duas modalidade de prova implicar mudança do nível de certificação, e ou quando a nota final na prova escrita for superior à oral.

O Inep publicou a portaria n.334 (BRASIL, 2013) com o objetivo de estabelecer parâmetros de credenciamento, recredenciamento e descredenciamento de postos aplicadores. O documento é um importante instrumento legal para garantir a padronização do exame, principalmente no que diz respeito à prova oral. Caso os avaliadores não sejam falantes de português como língua materna, exige-se que os examinadores devam possuir o nível de certificação a*vançado superior* do exame Celpe-Bras. O documento explicita também que os avaliadores devem fazer capacitações, que deverão ser oferecidas presencialmente ou à distância pelo Inep. Além disso, quanto às habilidades do examinador, ressalta-se que:

> 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 exame Celpe-Bras e agir com cordialidade, lembrando-se que estão em situação formal de interação. (BRASIL, 2013)

É importante ressaltar que, para além da condução da interação, que exige capacidade de gerenciamento da conversa, os avaliadores farão o julgamento do desempenho do examinando; por este motivo, o conhecimento do construto[3] do exame, bem como da grade de avaliação e da delimitação dos níveis do exame deve ser de conhecimento de todos que aplicam o exame oral.

O Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe) tem feito a gestão do exame e uma de suas ações envolve a oferta de cursos online de capacitação para aplicadores do exame. No que diz respeito à prova oral, o curso envolve a leitura dos manuais do exame e atividades de atribuição de notas a partir da escuta das interações. No curso de capacitação que precedeu a aplicação do exame no primeiro semestre de 2017, um fórum para dirimir dúvidas sobre o exame como um todo foi aberto e eram bastante frequentes as dúvidas sobre a atribuição de notas da parte oral como o comentário abaixo que sintetiza o debate:

Avaliador: Em casos como o do áudio 3, no qual o examinando está claramente no nível avançado, acho difícil definir um ponto de corte no qual eu decido se ele é apenas avançado ou se é avançado superior. E isso me acontece tanto na grade

[3] Construto é o que pode ser observado e medido. O construto é o grau de habilidade, conhecimento ou capacidade que o examinando possui que é inferido por meio do seu desempenho em um determinado teste (MESSICK, 1987).

holística como na analítica, visto que as nuances de "bastante", "raras", "algumas", etc, nem sempre são tão claras para mim. (INEP/CEBRASPE, 2017)

No comentário, o avaliador relata dificuldade na interpretação dos descritores da grade e como consequência, a dificuldade de classificar o examinando em uma faixa ou outra de nota. Em exames de desempenho como é o caso de Celpe-Bras, a atribuição da nota é feita pelos avaliadores que utilizam as grades de parâmetros para guiar o julgamento da performance do examinando. Eckes (2015) afirma que, em exames de desempenho, não só a habilidade do examinando e a dificuldade da tarefa definem a nota do examinando, mas o comportamento de avaliação também interfere no julgamento da nota. Segundo o autor, o avaliador pode tender a fazer julgamentos mais lenientes ou severos, ou pode, até mesmo, demonstrar preferência pelas faixas de certificação médias da escala, por exemplo. Eckes (2015) explica que, na situação de interação face a face, há pelo menos cinco facetas[4] e variadas maneiras de interação entre estas que podem impactar no resultado da nota final, a saber: o examinando, a tarefa, o entrevistador, o parâmetro de avaliação da grade e o avaliador.

Quanto aos parâmetros de atribuição de notas, o Manual do examinando (2015) define que a nota é atribuída a partir de evidências de que o examinando compreende a fala do *avaliador-interlocutor*, demonstra competência para interagir em Língua Portuguesa, tem domínio de estruturas linguísticas ao falar sobre diferentes temas, e possui pronúncia adequada. Estes parâmetros de avaliação estão organizados e descritos em duas fichas de avaliação. Na ficha de avaliação do *avaliador-interlocutor* (anexo 1), há uma descrição holística que engloba vários destes parâmetros com o objetivo de caracterizar o desempenho do examinando. A descrição holística é graduada em uma escala de 0 a 5 pontos. A nota zero ou *sem certificação* representa o desempenho de um examinando que pouco contribui para a interação, com frequentes pausas e hesitações, interrupção do fluxo da conversa, fluxo de conversa em outra língua, uso de recursos linguísticos limitado e/ou inadequado e problemas de compreensão da fala do *avaliador-interlocutor*. Um desempenho nota 5 ou *avançado superior* é atribuído ao examinando que demonstra autonomia, fluência, domínio amplo de estruturas linguísticas, pronúncia adequada e compreensão. Na ficha de avaliação do *avaliador-interlocutor*, os parâmetros de avaliação compõem uma descrição única de desempenho para cada nota de 0 a 5.

Na grade do *avaliador-observador* (anexo 2), uma nota é atribuída a cada um dos parâmetros ou itens que compõem a grade: *compreensão, competência interacional,*

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

fluência, adequação lexical, adequação gramatical e *pronúncia*. Cada um destes parâmetros ou itens de avaliação tem um peso diferente na composição da nota final do *avaliador-observador,* sendo 16.6% para *compreensão; 16.6%, competência interacional; 16.6%, fluência; 21%, adequação lexical; 21%, adequação gramatical e 8%, pronúncia.* Para cada nota de 0 a 5, há uma descrição dos seis parâmetros. Uma nota 0 em *adequação lexical* é caracterizada por evidências de que o examinando demonstra vocabulário muito inadequado e limitado, com interferências de outras línguas que comprometem a interação, ao passo que uma nota 5 se refere a um desempenho oral que demonstra vocabulário amplo e adequado com raras inadequações. Entre as notas 0 e 5, há uma graduação na maneira como o desempenho é descrito no limite desses extremos. Assim como a nota de *adequação lexical* é graduada, os outros cinco parâmetros de avaliação são também descritos, e cabe ao *avaliador-observador* julgar o desempenho a partir de cada um desses seis parâmetros e atribuir-lhe uma nota. Ao final, o *avaliador-observador* atribui seis notas que representam diferentes dimensões do desempenho oral do examinando, sendo que, o *avaliador-interlocutor* atribui uma nota única holística ao mesmo desempenho, ou seja, trata-se de duas formas de organizar os parâmetros da avaliação de um mesmo desempenho. Embora as grades organizem os parâmetros de avaliação de maneira distinta, no processo de avaliação, as notas atribuídas por meio da grade analítica e holística são consideradas equivalentes e entram inclusive como parâmetro para detectar situações de discrepância no julgamento do desempenho do examinando, conforme já explicitado.

O modelo de atribuição de notas se refere tanto à maneira como a nota é composta – ou seja, como as notas por cada parâmetro de avaliação são combinadas para formar uma nota final – quanto à maneira como as notas são atribuídas – ou seja, aos sistemas de controle das notas tais como avaliação em pares, controle de discrepância, etc. No caso da prova do Celpe-Bras, há uma composição numérica que corresponde à cada um dos sete parâmetros de avaliação: *nota do avaliador-interlocutor,* nota de *compreensão*, nota de *competência interacional*, nota de *fluência*, nota de *adequação lexical*, nota de *adequação gramatical* e nota de *pronúncia* que gera uma nota final única para a avaliação oral. Esta nota final, por sua vez, está relacionada com as faixas de certificação geral do exame (tabela 1). Messick (1987) denomina como modelo composto quando se tem uma nota em número que corresponde a uma faixa de proficiência. Ele alerta que, neste caso, o modelo é desafiador para o processo de validação de construto, por ser necessário investigar não só o significado da nota final em si, mas também a validade da nota de corte de cada uma das faixas de certificação. Além disso, por se tratar de duas grades distintas,

uma holística e outra analítica, para medir um mesmo desempenho, cabe questionar empiricamente até que ponto estas duas medidas teriam pesos equivalentes, de 50% cada uma na nota final, como vigora hoje. E ainda, cabe questionar até que ponto cada um dos parâmetros analíticos que compõem a nota final do *avaliador-observador* apresentam empiricamente o peso praticado atualmente. Estudar a interpretação da nota envolve entender o modelo estrutural de sua composição.

O objetivo desta pesquisa é discutir alguns dos possíveis argumentos empíricos sobre a validade das medidas por meio da análise do significado das notas atribuídas ao desempenho oral do examinando do exame Celpe-Bras, de forma a buscar respostas às seguintes perguntas:

- *O quanto cada uma das diferentes notas que representam diferentes parâmetros do desempenho oral do examinando do Celpe-Bras contribui para a composição da nota final da prova oral?*

- *Como cada uma das notas atribuídas pelos avaliadores contribui para a discriminação de cada um dos níveis de proficiência certificados pelo Celpe-Bras?*

Por se tratar de uma pesquisa com foco na análise empírica do resultado da avaliação, ou seja, no significado das notas atribuídas na prova oral, a discussão aqui proposta está diretamente relacionada à validade e confiabilidade em testes de línguas. O trabalho busca discutir alguns argumentos empíricos sobre a validade do significado das medidas que representam o desempenho oral do examinando. De maneira geral, a validade está relacionada à eficiência do instrumento de gerar informações pertinentes para o cumprimento do propósito do exame. Isto quer dizer que o significado das notas deve ser coerente com o propósito do exame, definido por meio das especificações que podem estar nos manuais, na página dos organizadores do exame, nas fichas de avaliação, etc.

Há vários aspectos da avaliação que podem ser analisados quanto à sua validade. Nesta tese, a validade é entendida como investigação científica de inferência, ou seja, a investigação da pertinência das interpretações feitas a partir dos resultados do teste. Fundamento-me em Messick (1987), para o qual a inferência está relacionada com a hipótese e, por isso, validar uma inferência é verificar uma hipótese. Se a nota é uma inferência sobre a proficiência do examinando, ou seja, uma hipótese sobre a capacidade do examinando se comunicar oralmente em português, discutirei neste trabalho alguns aspectos da correspondência entre as notas atribuídas na parte oral e as inferências que são feitas a partir dos resultados. Para tanto, é preciso avaliar

empiricamente os níveis de descrição de cada uma das notas e sua relação com a faixa de certificação definida nos documentos do exame pelo estudo da nota.

A validação é um processo em construção constante que carece de evidências teóricas e empíricas (MESSICK, 1987; FULCHER, 2003, BACHMAN, 1990; MCNAMARA, 2004). Quanto ao desenho do teste, por exemplo, é provável que uma prova oral válida ofereça ao examinando a oportunidade de interagir oralmente, mas apenas o desenho da prova não garante que os resultados do teste sejam válidos para nortear as ações tomadas a partir das notas. O que precisa ser validado não é o teste como ele é, mas as inferências que são feitas a partir dos resultados dos testes. Por isso, é preciso analisar não só o desenho do teste, mas o significado da nota gerada a partir do teste (MESSICK, 1987).

A seguir, a contextualização da análise da avaliação oral será detalhada.

1.3. A TAREFA NO ÂMBITO DO ENSINO E DA AVALIAÇÃO

Nesta seção, argumento que a prova oral é uma tarefa cuja situação é a entrevista de proficiência oral. Alguns autores como Bachman e Palmer (1996) e Fulcher (2003) consideram que o desenho da tarefa também está relacionado com as grades de avaliação. Nesse sentido, Fulcher (2003) aponta a necessidade de definição da tarefa de avaliação para que os envolvidos no processo de avaliação possam prever o tipo de performance que se espera do examinando. Como o objetivo do presente estudo é analisar o significado das notas, vale ressaltar que a medida está também relacionada com o contexto específico da avaliação. Construirei uma argumentação de forma a primeiro trazer algumas discussões sobre desenvolvimento da proficiência oral no âmbito dos estudos sobre ensino de línguas. Em seguida, tratarei do conceito de tarefa no contexto do ensino para, finalmente, abordar a tarefa no contexto da avaliação.

Long (2011) afirma que estudiosos da história dos métodos para ensino de línguas adicionais[5] identificam um movimento pendular quanto à escolha de abordagens ora intervencionistas, ora *laissez-faire*. Entendem-se por intervencionistas as metodologias estruturalistas como as de tradução, audiolingual, etc., ao passo que as *laissez-faire* estão relacionadas aos processos que envolvem competências e habilidades linguísticas, discursivas, pragmáticas, interculturais, etc. As primeiras são sintéticas, ao proporem o estudo das partes da língua de forma isolada. As últimas são holísticas e

[5] Neste trabalho, não faremos distinção entre língua adicional, segunda língua ou língua estrangeira.Para uma discussão sobre o uso dos termos, consulte Jordão (2014).

levam em conta o *syllabus* interno do aprendiz, ao propor uma análise contextualizada das regras, significados e funções da língua. Segundo Long (2011), os métodos holísticos para ensino e aprendizagem são eficazes no desenvolvimento da interlíngua[6] embora o foco na forma dos métodos intervencionistas seja também relevante quando se pretende desenvolver a proficiência avançada em língua adicional (LA). Tendo em vista os variados contextos e objetivos de ensino de LA, o autor afirma que seria despropositada a oferta de um método universal. Cabe salientar que, ainda segundo Long (2011), as pesquisas não advogam a favor de nenhum método e que a área de investigação sobre ensino e aprendizagem de línguas está em contínuo processo de desenvolvimento e que, por isso, é de responsabilidade dos professores selecionar abordagens e procedimentos teórico-metodológicos e justificar sua escolha em detrimento de outras, apoiando-se em pesquisas, e no que o docente acredita ser mais eficiente de acordo com o resultado de sua prática. Ainda que sejam diversos os contextos de ensino, aprender uma língua adicional é também um processo cognitivo relativamente similar entre aprendizes adultos, por isso é possível elencar um conjunto mínimo de princípios metodológicos a serem utilizados de maneira mais ou menos consensual entre professores, que podem variar com as circunstâncias de sala de aula. Dentre suas descobertas, Long (2011) destaca como relevantes as seguintes conclusões de resultados de pesquisa (1) os aprendizes, e não o professor, têm mais controle sobre o processo de aprendizagem; (2) a interlíngua é uma versão individualizada do aprendiz da língua alvo, cujo desenvolvimento pode estar relacionado à instrução, à proximidade da primeira língua com a língua adicional, dentre outros fatores; (3) além da proximidade entre as línguas, outros processos criativos desempenham um papel importante na aquisição de itens lexicais; (4) o desenvolvimento da interlíngua não é necessariamente linear; (5) a instrução tem efeito mínimo no desenvolvimento da interlíngua, porém o ensino pode acelerar o processo de aquisição e melhorar a previsão de aquisição de formas e funções raras, pouco salientes e específicas a um contexto comunicativo; (6) tanto o foco na forma quanto a instrução implícita são úteis para cumprir objetivos simples e complexos de aprendizagem.

A partir das conclusões de pesquisa, o autor postula que é incompatível o uso de uma ementa baseada em itens isolados da língua no contexto do ensino de línguas porque o controle da aprendizagem de um item isolado não garante a aquisição da forma linguística. Além do mais, a instrução tem um papel limitado no percurso individual do desenvolvimento da interlíngua. Também, o conjunto de itens pouco ou nada se relacionam aos objetivos comunicativos de estudo da língua pelos aprendizes.

[6] No contexto da discussão proposta por Long (2011), a interlíngua é individualizada e representa estágios de desenvolvimento da língua alvo.

Enquanto a proposta de itens isolados é questionável do ponto de vista da aquisição, propostas holísticas parecem ser mais coerentes com os resultados das pesquisas. As abordagens holísticas oferecem um insumo rico e realista, e, se forem elaboradas a partir de uma análise de necessidades comunicativas do aprendiz, serão ainda mais eficazes, uma vez que se considerará o uso futuro da língua. Deste modo, planejar o ensino de acordo com as necessidades dos aprendizes e baseado em tarefas significa aproximar a sala de aula aos contextos autênticos de uso da língua. As tarefas têm objetivos comunicativos e são elaboradas a partir de atividades, gêneros e ou pré-gêneros textuais orais e escritos de forma sequenciada (SWALLES, 1990).

De acordo com Norris (2011), o ensino de línguas baseado em tarefas é uma abordagem que se apoia em evidências científicas sobre ensino e cujas bases teóricas têm origem no campo do ensino de línguas adicionais e da filosofia da educação. O autor, ao discorrer sobre a origem da abordagem, afirma que o ensino baseado em tarefas surgiu como superação dos métodos pós-segunda guerra, que privilegiavam o ensino da forma desarticulada ao contexto de comunicação. O autor afirma que os princípios de abordagens como a da tarefa começaram a surgir a partir de 1980, quando os métodos sintéticos, ou seja, as metodologias que focam o estudo das partes da língua de forma isolada, sofreram diversas críticas, ao mesmo tempo em que os resultados de pesquisas apontavam que a aprendizagem se dá por meio não só de instrução, mas também em contexto de imersão, isto é, contextos em que o aprendiz utiliza a língua adicional no seu dia a dia e ou em atividades laborais, sem necessariamente ter uma reflexão sistemática ou metalinguística do uso da língua. As tarefas integram pressupostos pedagógicos teóricos e empíricos a fim de atingir objetivos realistas, ou seja, considera o que o aprendiz deverá ser capaz de fazer na língua alvo.

Quanto às características da abordagem, Norris (2011) afirma que a estrutura das atividades de uma abordagem baseada em tarefas é holística e procura oferecer aos aprendizes uma série limitada de experiências e informações relevantes que podem ser compreendidas e que, por meio das quais, os aprendizes podem se engajar nas oportunidades de desenvolvimento de aprendizagem. O autor faz uma resenha de uma década de trabalhos que discutiram a abordagem por meio de tarefas para elencar alguns de seus princípios, a saber: (1) a análise de objetivos de aprendizagem, que busca prever o contexto de uso da língua alvo no futuro; (2) a seleção e o sequenciamento de tarefas; (3) o desenvolvimento de materiais e da instrução que buscam didatizar as tarefas; (4) o ensino que potencializa a performance dos aprendizes ao desempenharem as tarefas; (5) a avaliação como mecanismo de

prover *feedback* relevante quanto à performance dos aprendizes ao realizarem as tarefas alvo.

As tarefas são recomendadas pelo Quadro europeu comum de referência para as línguas (EUROPA, 2001). O documento é resultado de um demanda política por uma definição de parâmetros comuns para o ensino e avaliação de línguas para facilitar a mobilidade entre os cidadãos dos países membros da União Europeia. As diretrizes foram elaboradas a partir da expectativa de diversos professores de línguas adicionais na Europa quanto à progressão de estudo na língua. O quadro reúne o que é consenso entre os professores sobre o que o estudante é capaz de fazer desde os níveis iniciais aos mais avançados. Fulcher e Davidson (2007) problematizaram o lugar da teoria na formulação da progressão da aprendizagem ou ensino de línguas proposto no Quadro, ao afirmarem que o desenvolvimento das habilidades descritas ao longo dos níveis não estariam baseadas em um arcabouço teórico sobre aquisição, ensino e aprendizagem de línguas e que as informações sobre os níveis seriam baseadas nas expectativas dos professores europeus que participaram da construção do Quadro, quanto ao desempenho na língua adicional. Embora a organização dos parâmetros do Quadro possa carecer de fundamentação teórica, como apontam Fulcher e Davidson (2007), o documento influenciou e continua influenciando diversos contextos de ensino de línguas. Cabe ressaltar que o conceito de tarefa configura entre as diretrizes. Segundo o documento, as tarefas são sinônimos de ações relevantes para os aprendizes, como escrever um livro ou deslocar um armário. As tarefas nem sempre são sinônimos de escrever textos. No documento europeu, por exemplo, as tarefas podem até não necessariamente envolver o uso da língua oral ou escrita, mas outras formas de comunicação que são ou deveriam ser igualmente objeto de ensino da língua alvo, na perspectiva do que seria ensino de línguas apresentada no documento. Long (2011) também atenta para o fato de que a didática baseada em tarefas nem sempre tenha que utilizar um texto como objetivo do ensino. Assim como Norris (2011), Long (2011) faz um resumo de estudos sobre os princípios e características metodológicas do ensino baseado em tarefas e ressalta: (1) o uso de tarefas e não de textos para nortear a didática; (2) a promoção da aprendizagem por meio da prática; (3) a elaboração do *input* para além dos textos autênticos; (4) o fomento à indução por parte do aprendiz; (5) a promoção do *input* rico e diferenciado; (6) o foco na forma contextualizado; (7) o *feedback*; (8) o respeito ao desenvolvimento individual do aluno; (9) a promoção da aprendizagem colaborativa e cooperativa; (10) a coerência entre ensino e uso futuro da língua.

Quanto aos testes de língua, segundo Norris (2011), a tarefa pode ser um método coerente quando o objetivo dos elaboradores é o de interpretar a habilidade dos

examinandos em usar a língua para comunicar, como no caso do exame Celpe-Bras. O Manual do Examinando explica resumidamente o conceito de tarefa como “um convite para interagir no mundo usando a linguagem com um propósito social” (BRASIL, 2010, p.4). Scaramucci (2001) considera que a tarefa é elaborada a partir de textos retirados de revistas, jornais, livros, etc. e que tem um propósito comunicativo que busca simular a linguagem usada no cotidiano. Outro aspecto importante da tarefa são os objetivos dos participantes. Scaramucci *et al* (2004) ressaltaram que as tarefas não dependem somente das instruções do comando, mas também dos objetivos de seus participantes; neste sentido, os autores consideram que as tarefas são sinônimo de interações situadas.

No contexto do Celpe-Bras, as interações são escritas e orais. Os textos escritos estão relacionados com a parte escrita do exame e envolvem a compreensão de um texto oral ou escrito para que o examinando possa escrever um outro em atendimento aos propósitos estabelecidos no comando da tarefa. Assumo que a interação oral face a face do exame Celpe-Bras é uma tarefa cuja situação é uma entrevista de proficiência oral nos quais avaliadores e examinandos têm objetivos e propósitos definidos.

Na edição de 2015 do Manual do Examinando, o termo *conversa* passou a caracterizar o exame oral, embora os manuais anteriores a definissem como uma *conversa* e *entrevista*, sendo os primeiros cinco minutos a parte da *conversa* e os quinze minutos seguintes a *entrevista* (BRASIL, 2010). Nos manuais anteriores e recentes, o termo *tarefa* é usado para definir as atividades da prova escrita apenas. Fulcher (2003) argumenta que, embora a definição de tarefa enfatize a presença de um propósito marcado, como nos comandos das tarefas de produção escrita do Celpe-Bras, há teóricos que incluem na definição de tarefa questões relacionadas ao processo de interação. Corroboro, assim como Fulcher (2003), Norris (2011) e Long (2011), uma compreensão do conceito de tarefa em um sentido amplo, que pode envolver tanto produção escrita quanto oral. Defino a tarefa como uma atividade não só de ensino como de avaliação, que pode ser oral e/ou escrita na qual os participantes interagem para atingir objetivos previamente definidos.

Fulcher (2003) propõe diretrizes para analisar tarefas orais de avaliação. A partir de algumas dessas diretrizes, analiso e problematizo a seguir a tarefa oral do Celpe-Bras:

1. Orientação da tarefa: guiada e flexível, composta por um roteiro de perguntas a serem feitas a partir de um dos 20 Elementos Provocadores. As perguntas podem ser feitas e editadas pelo avaliador-interlocutor, a depender da interação com o

examinando. Em situações de teste, recomenda-se que a tarefa seja o mais padronizada possível, por isso o fato da tarefa ser flexível pode comprometer a confiabilidade das medidas.

2. Relação entre os interagentes: o avaliador-interlocutor e examinando interagem de maneira que é desejado e esperado que o examinando fale por mais tempo, ou seja, o examinando tende a concentrar turnos maiores de fala.

3. *Status* do *avaliador-interlocutor* com relação ao examinando: assimétrico, o avaliador é provavelmente um professor ou uma autoridade na instituição em que o exame é aplicado; além disso, é responsabilidade do avaliador-interlocutor julgar o desempenho do examinando.

4. Tópicos: variados. Na primeira parte da prova os assuntos dizem respeito à vida do examinando e, na segunda parte, os temas são pautados pelo que circula na mídia impressa brasileira por meio dos Elementos Provocadores. Para cada edição, são enviados 20 Elementos Provocadores e seus respectivos roteiros de pergunta que contém normalmente de sete a oito perguntas cada.

5. Situação: entrevista de proficiência oral.

Quanto à análise da tarefa oral de avaliação no que diz respeito ao seu propósito, para Fulcher (2003), se a tarefa envolve interação entre interlocutores, como é o caso da tarefa oral do exame Celpe-bras, haverá comunicação e, por isso, há um propósito.

Em trabalho anterior (FERREIRA, 2012), os Elementos Provocadores (EP) e as suas respectivas perguntas que compõem os Roteiros de interação face a face de três edições do exame aplicados no primeiro semestre de 2004, 2007 e 2010 foram analisados, a fim de verificar as atividades de compreensão escrita implicadas no encaminhado da tarefa oral de avaliação. Observei que a maioria dos elementos provocadores analisados foram formados por recortes de reportagem que envolvia título, e um pequeno texto que poderia vir acompanhado ou não de uma imagem. Outros tipos de textos como campanhas publicitárias institucionais também foram encontrados, porém em menor quantidade. Normalmente a primeira pergunta dos roteiros analisados envolvia convidar o examinando a expor uma opinião sobre o assunto tratado no EP, definir um termo chave, explicitar alguma ideia do texto ou comentar o assunto do EP, dentre outras. De maneira geral, houve um movimento das perguntas no sentido de primeiro explorar ou verificar a compreensão das ideias do texto para, então, explorar o tema do EP, sem necessariamente contemplar as informações apresentadas no EP. Em todos os anos, houve uma tendência a relacionar o tema discutido ao contexto cultural do examinando, principalmente nas

últimas perguntas do roteiro. Quanto à diversidade de atividades de compreensão, as perguntas eram variadas: pediam que o candidato solucionasse problemas, propusesse explicações, desse dicas, inferisse informações,etc., colaborando para que o examinando produzisse sentidos a partir das leituras e releituras do EP ao longo da interação.

A edição de 2015 do Manual do examinando caracteriza o conjunto de perguntas de maneira que a primeira pergunta preveja a exploração da compreensão do EP, seguido de perguntas que podem contemplar as opiniões e experiências do examinando relacionadas ao tema do EP e a questões culturais do examinando e dos brasileiros. Exemplificamos abaixo com o décimo EP e seu respectivo roteiro de perguntas, que fizeram parte da aplicação do primeiro semestre de 2016:

FIGURA 1 - Elementos provocadores e roteiro de interação face a face
Fonte: BRASIL, 2016b; BRASIL, 2016c

Conforme aponta a pesquisa de Ferreira (2012) e o Manual do examinando (2015), a primeira pergunta *comente o material,* que corresponde ao EP *Lugares para aprender,* contempla a exploração da compreensão global do texto pelo examinando. Normalmente, nesta etapa verifica-se se está claro para o examinando o assunto que será tratado nos minutos seguintes da interação. Na sequência, as perguntas são diversificadas, sendo que a primeira implica o posicionamento do examinando com relação à ideia de que as experiências fora da escola transformam os conteúdos. Em seguida, o examinando pode localizar no EP exemplos de lugares para aprender ao formular uma resposta à segunda pergunta. Nas perguntas subsequentes, o

examinando pode ser convidado a falar de suas experiências de aprendizado fora da escola, ou seja, relaciona-se o assunto do EP com sua vivência. Nas duas últimas perguntas do roteiro, explora-se a relação entre o assunto do EP e o país de origem do examinando.

Embora a tarefa tenha o objetivo de oferecer oportunidade ao examinando de demonstrar sua proficiência oral a partir de uma situação de entrevista oral com foco nos assuntos do EP, a leitura está implícita no desenho da tarefa da prova oral, assim como a compreensão oral de áudio e vídeo está implícita no desenho de algumas das tarefas que compõem a prova escrita[7]. É importante ressaltar que, embora esteja argumentando que a tarefa oral pressupõe a leitura de textos escrito e imagens do EP, não afirmo que a compreensão escrita deva configurar entre os parâmetros de avaliação da prova oral das grades de avaliação.

Tendo definido os principais aspectos da estrutura da tarefa da prova oral, apresentarei a seguir um debate sobre a questão da validade de entrevistas de proficiência oral.

[7] Para uma discussão sobre desenho de tarefas integradas para prova escrita, veja Pileggi (2015).

2. ENTREVISTA DE PROFICIÊNCIA ORAL

Hughes (1989) parte do pressuposto de que o objetivo do ensino de línguas, no que diz respeito à produção oral, é de preparar os aprendizes para interagir com desenvoltura na língua alvo. O autor propõe que as avaliações devam representar tarefas a serem desempenhadas pelos examinandos: “as tarefas avaliativas devem prever comportamentos que representem de fato a habilidade do examinando e que, por meio das quais, as notas atribuídas sejam válidas e confiáveis”[8] (HUGHES, 1989, 101p.). Cabe ressaltar que, neste trabalho, a validade está relacionada com as inferências que são feitas a partir do resultado do teste, conforme proposto por Messick (1987).

Também é importante destacar que a nota do teste está em função não somente da tarefa, mas de um conjunto de condições de aplicação da prova (Messick, 1987). Por isso, Messick (1987) adverte que, para fins de validação, a ênfase da investigação deve ser dada na nota e não no instrumento porque é por meio da análise dos resultados que podemos verificar propriedades de validade e confiabilidade da medida.

No caso do Celpe-Bras, o quanto o desenho da prova oral representa a proficiência que se quer atestar por meio da nota está relacionada com a maneira

[8] The tasks should elicit behavior which truly represents the candidates’ ability and which can be scored validity and reliably.” (HUGHES, 1989, 101p.) No terceiro capítulo discutiremos os conceitos de validade e confiabilidade.

como o desenho específico do exame está para um conjunto de outras situações ou tarefas possíveis que poderiam ou representariam a proficiência oral. Hughes (1989) dá exemplos de vários tipos de situações de tarefas que poderiam ser utilizadas para acessar a produção oral, tais como: discussão em grupo, imitação, *tape-record stimuli*, dentre outras.

A prova oral do Celpe-Bras foi objeto de discussão de Coura-Sobrinho e Dell'Isolla (2009) que analisaram a interação face a face a fim de aproximá-la a um gênero. Baseando-se em teorias da Análise do Discurso, os autores concluíram que se trata de uma conversa controlada. Underhill (1987) distinguiu a entrevista de uma conversação ou discussão quanto à postura dos interlocutores. Segundo o autor, na entrevista, o interlocutor controla a interação, ao passo que, em uma conversação ou discussão, a interação flui de maneira mais espontânea. Shohamy (2000) resenhou pesquisas sobre a entrevista de proficiência oral e concluiu que a maneira de falar nessa situação é específica e difere de uma conversa, embora ambas sejam práticas interativas. De forma geral, tais autores, partindo de diferentes perspectivas teóricas, sugerem que a situação da entrevista de proficiência oral se difere de uma conversa. Corroboro as sugestões dos autores e entendo que a interação oral face a face do exame Celpe-Bras é uma situação de entrevista de proficiência oral. Neste e em outros capítulos da tese, apresentarei argumentos e evidências empíricas que fundamentam tal posição.

A metodologia da entrevista de proficiência oral (EPO) desenvolvida pelo *Foreigner Service Institute (FSI)* do governo estadunidense para atender a demanda de certificar a proficiência de diplomatas impactou o ensino de línguas nas universidades e o debate sobre o desenvolvimento de testes. Tais entrevistas têm sido objeto de estudo de muitas pesquisas que forneceram insumos para compreensão dos parâmetros de avaliação da proficiência oral, entre outros aspectos. Brown (2005), por exemplo, ressalta que Kramsch e Savignon criticaram a grade dos testes desenvolvidos pelo FSI por não refletir o construto da competência interacional, como se a comunicação, na proposta do exame do FSI, fosse apenas uma transferência de informação. Brown (2005) sugere que tais argumentos influenciaram a formulação dos conceitos sobre competência comunicativa nas décadas de 1970 e 1980. Ainda sobre a relação entre a abordagem comunicativa e a elaboração de testes orais, Brown (2005) resume que é consenso entre os pesquisadores a ideia de que os testes devam ser comunicativos; no entanto, a autora afirma que ainda não está claro para os autores qual seria o formato destes testes ou como um teste comunicativo deveria ser. No caso da prova oral do Celpe-Bras, o Manual do examinando (2015) afirma que o exame é de natureza comunicativa.

Brown (2005), ao defender a entrevista de proficiência oral como forma de avaliação oral, argumenta que a validade do método se dá pela natureza conversacional da interação “enquanto as tarefas das avaliações de desempenho podem assumir variados formatos como *roleplays*, descrição de imagens e *information gap*; o argumento para a validade da metodologia da entrevista de proficiência oral deriva da sua natureza conversacional” (BROWN, 2005, 1p.). Além disso, a autora afirma que as entrevistas são fáceis de administrar uma vez que são estruturadas, porém sem ter um *script* definido. Como uma entrevista nunca será idêntica a outra, os tópicos selecionados para interação podem ser gerenciados em diversos contextos para diferentes examinandos sem que isso interfira na segurança do teste. No caso da tarefa oral do Celpe-Bras, os avaliadores podem variar a escolha dos EPs e das suas respectivas perguntas. Há pesquisadores, no entanto, que discutem a validade das entrevistas orais ao argumentar que a entrevista de proficiência oral não reflete os discursos da vida real.

Johnson (2001), ao investigar o tipo de evento de fala que seria a entrevista de proficiência oral, concluiu que, devido à assimetria entre os interagentes – interlocutor e examinando –, a conversa mais se assemelharia à uma entrevista sociolinguística. A autora estudou uma entrevista oral de proficiência em Língua Inglesa a fim de definir que tipo de evento de fala é a entrevista de proficiência oral, sob a perspectiva das teorias da Análise da Conversação. Dentre os aspectos investigados na pesquisa, destaca-se a mudança de turno, correção de erros, mudança de tópicos e perguntas. Ao final, a autora aproximou a entrevista oral de proficiência às entrevistas sociolinguísticas, uma vez que ambas lançam mão de estratégias de gestão de interações semelhantes, como o quebra-gelo e a mudança de tópicos (JOHNSON, 2001, p.116).

Na entrevista de proficiência oral do exame Celpe-Bras, por exemplo, a primeira parte da prova se assemelha a um quebra-gelo em que o examinando já está sendo avaliado. Na segunda parte da prova, as mudanças de tópicos são frequentes uma vez que o *avaliador-interlocutor* deve organizar a interação de forma que, após o quebra-gelo, um EP deverá pautar o assunto dos próximos cinco minutos de entrevista e, em seguida, um outro EP, que provavelmente tratará de um assunto distinto do primeiro EP, norteará a conversa e um terceiro EP será introduzido nos últimos cinco minutos do exame. Segundo Johnson (2001) o que faz com que a entrevista tenha o foco adequado ao tópico é o controle do interlocutor sobre a interação que, por sua vez, envolve tanto a escolha dos EPs como também o gerenciamento de suas perguntas, no caso do Celpe-Bras.

Brown (2005) pondera que, embora haja argumentos como os apresentados por Johnson (2001), que questionam a validade da entrevista ponderando que a entrevista de proficiência oral não reflete os discursos da vida real, este tipo de situação de prova é a mais popular para a avaliação da proficiência oral. O debate proposto Johnson (2001) nos remete aos limites da autenticidade da situação da avaliação e do desenho de tarefa. A autenticidade se refere à simulação mais próxima possível dos usos da língua na tarefa. Ressalto que em uma situação de avaliação, a autenticidade não pode ser o único ponto levado em conta no desenho da prova, porque há outros aspectos importantes, como a tentativa de controlar o tempo e a padronização do formato da interação de forma que os examinandos tenham mais ou menos a mesma oportunidade para demonstrarem o que sabem. Dessa forma, as entrevistas de proficiência oral são mesmo limitadas ao tentar simular situações de conversas da vida real, por se tratarem de contextos de avaliação.

Todo o debate sobre a natureza das entrevista orais de proficiência oral está relacionado ao argumento da validade. Julgar a validade de uma situação de prova analisando a situação da interação está associado ao que Messick (1987) afirma ser análise de validade de face de um exame. Segundo o autor, argumentar a validade de um exame a partir do seu desenho ou do padrão da conversa, sem levar em conta a nota e como ela está organizada não é suficiente para concluir algo sobre validade.

A falácia da validade de face das entrevistas de proficiência oral também está relacionada ao fato de um exame direto ser considerado automaticamente válido por gerar linguagem real e autêntica, posição adotada por muitos autores, como Hughes (1987), ao passo que o exame indireto não seria considerado válido. Fulcher (2003) refuta os argumentos de Hughes (1987), sobre a validade automática dos exames diretos, afirmando que não há uma definição de como é ou deveria ser um discurso autêntico ou real e que os descritores dos parâmetros de avaliação muito frequentemente não fornecem definições operacionais do construto. Fulcher (2003) problematiza a questão do uso da validade de face e afirma que o foco deveria ser na definição de como é ou deve ser uma situação real de fala e como os descritores deveriam fornecer definições operacionais do construto. Ainda segundo o mesmo autor, o argumento de que uma situação de entrevista seria necessariamente válido confunde a compreensão e o debate sobre a manifestação do comportamento que se quer medir com os parâmetros que refletem o construto teórico do exame. A interação verbal é um comportamento a ser medido por meio de testes que são indicadores indiretos da proficiência a ser avaliada.

Ainda sobre a validade de entrevistas orais, Fulcher (2003) afirma que o único trabalho que traz evidências empíricas de validade dessa situação de prova é o trabalho de Bachman e Palmer de, 1983. Eles compararam resultados de vários testes, bem como os efeitos de diversas situações de teste na confiabilidade da nota e concluíram que a entrevista de proficiência oral maximiza a avaliação dos construtos, ao passo que minimiza o efeito do método nas notas. O efeito do método ocorre quando a nota está sendo influenciada pela situação como um todo ou por algum de seus aspectos. É desejável, para fins de avaliação de larga escala, que o efeito do método seja minimizado.

Com base no exposto, considero que são potencialmente válidos os resultados de exames de proficiência oral que ofereçam oportunidade para o examinando demonstrar sua capacidade de interagir oralmente em Língua Portuguesa, como previsto na tarefa do Celpe-Bras. No entanto, para afirmar algo sobre a validade das inferências feitas a partir das notas é preciso analisá-las empiricamente.

Sobre a natureza da tarefa, o que se põe à prova é o significado da medida: se seria o significado de uma determinada nota oral específico do contexto de avaliação ou se o significado da nota oral de um determinando exame poderia ser generalizado para outros contextos em que demande proficiência oral (MESSICK, 1987). A questão central do debate é verificar se o tipo de desempenho na entrevista corresponde aos domínios de linguagem que se quer atestar. A entrevista de proficiência oral simula situações de uso da língua por meio da qual inferências sobre a capacidade do examinando de usar essa língua em situações fora do teste serão feitas. Nesse sentido, a validade de conteúdo, que está relacionado ao desenho da prova, deve ser analisada não só a partir dos estudos que explicitam a situação da entrevista oral a partir da análise das interações, como fez Johnson (2001), mas também a partir do estudo da nota dos parâmetros de avaliação dessas interações que compõem as grades de avaliação. A validade de conteúdo de testes, como as entrevistas de proficiência oral, está relacionada também com evidências sobre a estrutura interna do teste, ou seja, de como os itens da prova estão compondo o escore final (AERA, 2014). Como a validade está relacionada com as inferências que são feitas a partir do resultado do teste, o processo de validação deve contemplar também o estudo da composição interna da nota. A nota do teste está em função não só da tarefa, mas de um conjunto de condições de aplicação da prova que também devem ser estudadas (Messick, 1987).

Embora em um contexto de avaliação seja necessário fazer uma padronização no encaminhamento da interação, a natureza da entrevista de proficiência oral é

imprevisível e, por isso, há variáveis que podem interferir na mensuração da nota do entrevistado tais como o interlocutor, o avaliador e a tarefa (ECKES, 2015; BROWN, 2005). Brown (2005) classificou como entrevista de proficiência oral as interações face a face em que um entrevistador conduz uma conversa previamente estruturada. A autora discute especificamente a metodologia do exame *International English Language Testing System* (IELTS), que é dividido em quatro fases. A primeira fase trata-se do quebra-gelo, seguido de uma conversa sobre tópicos familiares ao examinando que podem envolver descrição, narração e explicação; na terceira fase, a partir de um texto ou imagem o examinando pode ser encorajado a propor soluções para algum problema; e, na fase final, o examinando pode ser convidado a falar de seus planos para o futuro. O exame oral do Celpe-bras inicia-se também com a fase de quebra-gelo, seguida de uma conversa sobre tópicos do cotidiano a partir dos Elementos Provocadores, que é dividida em três momentos. Nesse sentido, a tarefa de avaliação do IELTS a qual Brown (2005) se refere é ligeiramente distinta da do exame Celpe-Bras. No entanto, em ambos os exames os examinandos estão em situação de entrevista de proficiência oral, uma vez que há a fase do quebra-gelo, o controle de assuntos e o gerenciamento das perguntas por parte do interlocutor.

A seguir, farei um breve histórico das entrevistas de proficiência oral.

2.1. HISTÓRICO DAS ENTREVISTAS DE PROFICIÊNCIA ORAL

Conforme problematizei na introdução deste trabalho, o tema da avaliação no campo de pesquisa da Linguística Aplicada é relativamente recente e, segundo Fulcher (2003), a tema da avaliação oral como objeto de interesse dos linguistas aplicados é um assunto ainda mais novo. A avaliação de proficiência oral começou a ser discutida com algum interesse na década de 1920. Até a Segunda Guerra, pouco ou nada se debatia sobre avaliação oral. Fulcher (2003) conta que, antes de 1939, os linguistas acreditavam que não havia uma maneira confiável de avaliar as interações orais. Por este motivo, as provas orais se restringiam a avaliar pronúncia em testes que envolviam ditados.

McNamara (2004) corrobora Fulcher (2003) ao afirmar que os estudos sobre avaliação de línguas na área da Linguística Aplicada começaram a se desenvolver a partir de 1950 e que, por volta dos anos 1970 e 1980, os métodos comunicativos impactaram no desenho dos testes, de forma que as avaliações passaram a dar mais ênfase ao desempenho. McNamara (2004) ressalta que naquele momento a crescente

popularidade do ensino para fins específicos também incrementou o interesse por avaliações de desempenho.

Fulcher (2003) também conta que o primeiro teste verdadeiramente oral foi aplicado em 1930 pelo *College Board's English Competence Examination,* uma instituição de ensino estadunidense. O propósito do teste era selecionar futuros estudantes universitário vindos de outros países. Tratava-se de uma conversa preparada pelo examinador a partir de dez tópicos por meio da qual a avaliação era feita segundo os parâmetros de fluência, responsividade, rapidez, articulação, enunciação, comando de construção, e o uso de conectivos e vocabulário que eram graduados em uma escala de três níveis, a saber: proficiente, satisfatório e não-proficiente.

Fulcher (2003) sustenta que o exame do *College* foi criado para atender a uma demanda política. Os Estados Unidos passavam por um momento de restrição à imigração. O autor afirma que naquela época alguns estudos baseados em testes de inteligência foram feitos no sentido de buscar evidências empíricas que comprovassem a superioridade intelectual de grupos de pessoas por nacionalidade. A partir dos resultados deste tipo de pesquisa, houve uma preocupação por parte das autoridades estadunidenses em selecionar indivíduos por capacidade intelectual. Baseando-se nesses estudos, um sistema de cota imigratória por grupos foi instalado. Segundo o autor, o *Immigration Act* de 1929 é o documento que deixa clara a preocupação de não permitir a entrada no país de estrangeiros com baixa capacidade intelectual. O autor explica que uma outra maneira de conseguir entrar nos Estados Unidos sem ser pelo sistema de cotas era por meio do pedido de visto para cursar universidade. Por este motivo as autoridades demandaram dos conselhos de ensino que explicitassem o que exatamente o estudante universitário estrangeiro precisaria saber em termos de uso da língua inglesa.

É importante notar que este primeiro teste oral de língua inglesa norte-americano foi desenvolvido no contexto de uma política higienista estadunidense e foi fortemente influenciada por resultados de testes de inteligência. A definição de parâmetros na elaboração de testes é fundamental e, neste contexto, exigiu-se ainda mais rigor na explicitação do que estava sendo medido por meio das prova orais para que as autoridades tivessem algum controle sobre o significado da nota.

No contexto das primeiras propostas de avaliações orais desenvolvidas na Inglaterra, o autor salienta que não havia tanta preocupação com relação à nota, uma vez que as propostas faziam parte de um currículo de ensino de línguas e tinham um propósito de gerar insumos para melhorar o ensino e a aprendizagem. O autor cita a

proposta de avaliação oral desenvolvida pela Universidade de Cambridge, que fazia parte do exame Certificado de Proficiência em Inglês de 1913. Ela era organizada de forma que o examinando era submetido a meia hora de ditado, seguida de meia hora de atividades que envolviam leitura e conversa. Porém, o que era avaliado desta conversa era a pronúncia apenas.

Fulcher (2003) afirma que os anos da Segunda Guerra foram os mais agitados para história das avaliações orais em línguas. Os militares estadunidenses rapidamente identificaram problemas de comunicação oral entre os soldados e, por isso, foi criado um programa de treinamento para sanar o problema. Esse programa tinha como objetivo dar instrução para que o treinando aprendesse a falar a língua adicional e tivesse noções da língua em situações específicas de atuação, que poderiam ser na área da saúde, navegação, etc. Os parâmetros de avaliação oral do programa iam de *expert* a *competent* e explicitavam o que os treinandos estavam aptos a fazer na língua adicional. Fulcher (2003) explica que os *expert* seriam os que dominavam a língua no mesmo nível da sua própria língua materna e *competent* aqueles que poderiam se fazer entender por falantes nativos adultos em situações não-técnicas. O programa de treinamento em línguas adicionais teve impacto no ensino, sendo a mudança do parâmetro *conhecimento gramatical* para *habilidade para falar* o impacto mais relevante. Fulcher (2003) ressalta que foi nesse contexto que o artigo assinado por Kaulfers, publicado em 1944, lançou as bases para uma forma de avaliar a oralidade. De acordo com Fulcher (2003), o texto definia que as avaliações orais deveriam ser baseadas em exame direto, com itens variados em termos de dificuldade, sendo os interlocutores diferentes do avaliador. O artigo recomendava ainda que os interlocutores e avaliadores deveriam receber treinamento atribuindo notas à performances reais. Além disso, o texto do artigo definia que o interlocutor deveria deixar o examinando à vontade para falar, mas, ao mesmo tempo, ter postura profissional. Sobre a estrutura do teste, o texto defendia que ele deveria ser iniciado com quebra-gelo, seguido de tarefas com as seguintes temáticas: serviço de segurança, pedido de informações, dar informações.

Nesse contexto, o *Queen's College*, instituição de ensino sediada em Nova Iorque, era quem comandava o treinamento militar de língua adicional e criou um teste baseado em três tarefas: descrição de imagem, conversa sobre um tópico e conversa direcionada. A atribuição de notas tinha como parâmetros: parte 1, a comunicação e a parte 2, a gramática. De acordo com Fulcher (2003), o teste elaborado pelo *Queen's College* foi o precursor da Entrevista de Proficiência Oral (*Oral Proficiency Interview*) desenvolvido pelo *Foreign Service Institute* (FSI)

Na Inglaterra, durante a Segunda Guerra, um teste oral desenvolvido pela University of Cambridge Local Examinations Syndicate (UCLES) foi também usado pelos militares. Porém, nesta avaliação o objetivo era dar *feedback* para o ensino e encorajar os treinantes a continuarem seus estudos.

No contexto da Guerra Fria, conta Fulcher (2003), o representante do órgão responsável pelo treinamento do pessoal do governo estadunidense que ia em missão para o estrangeiro era do FSI. Por isso, houve a necessidade de se organizar um registro sobre a aptidão linguística do pessoal, e foi assim que o instrumento de avaliação foi criado. Um comitê técnico do FSI desenvolveu um teste oral cujos parâmetros de avaliação se tornaram públicos. A escala de avaliação ia do nível 1, que representava aqueles examinandos que apresentam falta de habilidade para usar a língua, ao nível 6, que representava os examinandos que teoricamente demonstravam habilidades correspondentes a de um falante nativo. O nível 4 era o mínimo exigido para que o examinando assumisse um posto diplomático. Fulcher (2003), ao analisar as grades de avaliação da prova oral do FSI, afirma que o que foi proposto foi uma escala holística com seis níveis cujos descritores eram 'fracos', uma vez que eram elaborados de forma que pouco descriminava um nível de outro. É importante lembrar que o desafio de descrever verbalmente desempenhos orais e organizá-los em uma escala é extremamente complexo e que vai além da descrição verbal de desempenho. É preciso avaliar empiricamente os níveis de descrição pelo estudo da nota e não só da maneira como o descritor está escrito. Ainda sobre os problemas da proposta do FSI, o autor explica que, após a aplicação do teste, descobriu-se que a maioria dos funcionários não tinham aptidão para as tarefas consulares e, assim, um estudo sobre o teste foi feito para verificar o porquê da reprovação dos oficiais. Os elaboradores do teste descobriram que a idade e a patente dos oficiais influenciavam na nota e identificaram, já naquela época, o que se chama hoje de *test bias,* ou viés do teste, que é quando o teste tende a prejudicar ou beneficiar sistematicamente um grupo de examinandos.

Com relação aos parâmetros de avaliação, Fulcher (2003) destaca que em 1958 a escala do teste do FSI foi modificada de forma que os avaliadores passaram a utilizar uma escala de 6 categorias, em que eram avaliados cinco fatores, a saber: sotaque, compreensão, fluência, gramática e vocabulário. Segundo o autor, a proposta foi o primeiro passo para o desenvolvimento de escalas analíticas no contexto de provas orais de línguas. O Serviço de Inteligência do governo estadunidense usou o teste do FSI, porém incrementou o processo de atribuição de notas ao incluir mais de um avaliador para melhorar a confiabilidade da nota. Fulcher (2003), ao fazer o relato das etapas de elaboração dos testes orais, concluiu que mesmo em momentos iniciais de

desenvolvimento de propostas houve um interesse em debater a organização das grades na forma holística ou analítica, bem como seus descritores e níveis, as diferenças entre parâmetros linguísticos e comunicativos, as maneiras de aumentar a confiabilidade da nota e o viés do teste (*test bias*).

O teste oral desenvolvido pelo FSI foi uma proposta que influenciou fortemente o ensino de línguas (BROWN, 2005; FULCHER, 2003). Segundo Fulcher (2003), o teste do FSI passou a ser usado fora do governo, adotado pela *Peace Corps,* agência federal estadunidense, e por universidades estadunidenses para certificar professores bilíngues. O mesmo autor conta ainda que, dentre os fatores de popularização do exame, destaca-se o fato do teste ser direto, ter comprovada confiabilidade alta entre os avaliadores, e fundamentar-se em uma abordagem nocional ou funcional[9], que era bastante difundida. Por conta da popularidade do teste do FSI, alguns estudiosos começaram a investigar a adequação da proposta do exame ao contexto de ensino universitário estatunidense, no que se referia à capacidade de a escala discriminar os níveis da população universitária. Em 1979 coube ao Conselho norte-americano de ensino de línguas estrangeiras (*American Council on the teaching of foreign language*) a tarefa de desenvolver e elaborar instrumentos e parâmetros de avaliação de línguas adicionais, e rever os parâmetros de avaliação do teste oral do FSI.

Fulcher (2003), ao fazer um histórico dos desenvolvimentos dos testes orais, defende que as críticas focavam muito mais os parâmetros, os descritores e a escala do que a tarefa em si. A problematização da tarefa surgiu depois da discussão sobre a validade da metodologia de entrevista de proficiência oral (EPO) para avaliar as habilidades de interação oral em língua adicional.

Sobre os debates em torno do processo de atribuição de notas do FSI, Fulcher (2003) também faz um pequeno resumo das discussões. Segundo o autor, os críticos à grade do FSI argumentam que os aspectos da gramática e do vocabulário tinham maior peso na nota final do examinando, dentre todos os outros aspectos. Embora muitos tenham criticado o peso destes parâmetros, Fulcher (2003) afirma que, por muito tempo, pouco progresso foi feito na discussão da escolha e do peso dos componentes de grade. Em 1958, os descritores por níveis foram elaborados para o teste do FSI, e a partir disso pela primeira vez exemplos de interações reais e sua relação com a grade entraram nas agendas de pesquisa.Nesse processo, os descritores foram definidos de maneira que fosse possível avaliar qualquer língua. Dos aspectos negativos da grade de avaliação proposta pelo FSI que influenciaram outras

[9] A abordagem nocional ou funcional teve sua base teórica desenvolvida durante as décadas de 1960 e 1970 por autores como Halliday, Hymes, Searle, entre outros; . No contexto do ensino de línguas adicionais, a abordagem enfoca o ensino da maneira como a língua é usada em diversas situações.

grades, o autor destaca a propagação do mito do falante nativo, uma vez que o nível máximo era descrito como proficiência de nativo. Além disso, como os parâmetros no teste proposto encontram-se misturados às situações das tarefas, fica difícil saber o que o examinando é capaz de fazer fora da situação de teste. Essa confusão sobre a definição dos parâmetros que são gerais e específicos das tarefas é um aspecto que perpassou várias gerações de escalas, e que foi muito criticado pelos linguistas aplicados adeptos das abordagens comunicativas de ensino de línguas.

Além dos parâmetros de avaliação, outros fatores podem interferir na nota dos examinandos em situação de entrevista oral. Discutirei, a seguir, estes outros fatores.

2.2. ENTREVISTA DE PROFICIÊNCIA ORAL: FATORES QUE PODEM INTERFERIR NO DESEMPENHO DO EXAMINANDO

No âmbito das teorias psicométricas, ou seja, que explicam o significado das medidas, os fatores que podem interferir em uma nota final de teste são chamados de facetas. Segundo Eckes (2015), facetas são sinônimos de fatores, variáveis ou componentes que fazem parte da situação de avaliação e que afetam as notas de forma sistemática. A faceta pode ser o nível de proficiência do examinando, que é o objeto de interesse do teste, e por isso é desejável que interfira na nota do teste. Outros tipos de faceta podem ser os componentes do método de avaliar, como o desenho da tarefa, os parâmetros de avaliação, o avaliador, o entrevistador, o tempo de execução da prova, etc.

Brown (2005) reconhece que, em geral, as entrevistas de proficiência oral têm objetivo de avaliar a capacidade comunicativa dos examinandos, e são desenvolvidas para que estes se engajem em uma interação que se pretende natural, espontânea e imprevisível. A autora ressalta que todos os examinandos devem ter as mesmas chances de demonstrar sua proficiência, e por isso os testes devem ter parâmetros como controle de tempo e complexidade das tarefas para que as interações sejam similares entre um e outro examinando. Segundo Brown (2005), o controle da complexidade da tarefa normalmente é resolvido pela delimitação de tópicos para discussão, de tarefas funcionais cobradas e de estratégias comuns de elicitação pelos entrevistadores. Por esta última razão, o treinamento de entrevistadores também é importante no processo. Embora a autora afirme que faltam pesquisas para demonstrar as evidências sobre formas de se hierarquizar e agrupar tópicos equivalentes em escalas de complexidade, ela ressalta que é consenso que os

tópicos devem fazer parte das condições de desempenho de testes orais, e que devem ser organizados “tópicos equivalentes” para o entrevistador selecionar. A definição da tarefa também é importante para prever o tipo de performance que se espera do examinando, lembra Fulcher (2003) O desenho da tarefa também faz parte ou está relacionado com as grades de avaliação. No caso da prova oral, especificamente, o autor afirma que os vários fatores como o tempo, a estrutura dos participantes, o modelo de atribuição de nota e o *input* influenciam no desenho da tarefa. Brown (2005) e Fulcher (2003), entretanto, problematizam a falta de evidências científicas nos estudos dos fatores que mais influenciam no desenho de tarefas orais pois, para estudar estes fatores, seria preciso discutir os elementos por meio de tarefas comparáveis, o que seria difícil de se fazer porque envolveria o controle de muitas variáveis.

Em geral, as pesquisas que analisam as facetas da metodologia de entrevista de proficiência oral que podem interferir na nota final do examinando partem do pressuposto de que a interação é co-construída entre examinandos e avaliadores e que, por isso, tanto entrevistador quanto avaliador são componentes da metodologia do exame, assim como o desenho da tarefa e os parâmetros de avaliação. Antes de abordar a questão principal deste trabalho, que é a atribuição de notas, resenharei algumas pesquisas sobre outras facetas da prova oral.

Schoffen (2003) analisou 12 entrevistas orais de examinandos falantes de espanhol como língua materna com diferentes perfis de proficiência que se submeteram ao Celpe-Bras em 2002. A autora relacionou a faixa de certificação com o desempenho do examinando na interação oral. Em 2002, o exame certificava três faixas de proficiência, a saber: avançado, intermediário e sem certificação. Em 2003, as faixas de certificação avançado superior e intermediário superior foram incorporados ao exame (SCHOFFEN, 2003). Dentre os resultados, a autora aponta que os parâmetros *pronúncia, fluência, competência interacional, adequação lexical e adequação gramatical* parecem ser eficientes para discriminar examinandos certificados entre as faixas Avançado e Intermediário, ao passo que o parâmetro *compreensão* nada contribuiu para distinguir os examinandos entre as faixas, uma vez que todos obtiveram um desempenho adequado neste quesito. A pesquisa contribui para o debate sobre a descrição do desempenho e sua relação com a escala, encaminhando sugestões de modificação na redação dos descritores. É importante fazer uma pequena ressalva quanto à montagem do corpus da pesquisa. Como a faixa de certificação do exame e não a nota da prova oral foi utilizada como critério para composição da amostra, pode ser que examinandos com baixa proficiência tenham obtido uma nota de certificação baixa pelo seu desempenho escrito e não oral, uma

vez que a menor nota entre a prova escrita e a oral é o que define o resultado para fins de certificação. Pode ser que o examinando certificado como Intermediário, por exemplo, tenha obtido tal nota de certificação devido aos parâmetros de avaliação da prova escrita e não necessariamente por causa da pronúncia, fluência, etc.

Brown (2005) pesquisou a maneira como os entrevistadores diferem ao conduzir suas entrevistas e o impacto dessas diferenças na nota final do examinando. A pesquisadora analisou a condução das entrevistas e as notas atribuídas e justificadas pelos avaliadores, a partir de uma simulação da prova oral do exame *International English Language Testing System* (IELTS). A autora concluiu que os entrevistadores variavam a forma de conduzir uma entrevista no que se refere à organização dos assuntos e na maneira como interagiam com os examinandos. Quanto à organização dos assuntos, o exame IELTS prevê que os examinandos sejam encorajados a falar de assuntos relacionados à sua experiência pessoal e também sobre temas mais abstratos. No entanto, houve entrevistadores que abordaram apenas temas da vida pessoal do examinando, tornando a entrevista fácil e dando poucas oportunidades para o examinando demonstrar seu desempenho, afirma a autora. Cabe ressaltar que no exame IELTS está previsto que a entrevista seja feita, gravada e enviada a dois avaliadores (*raters*) que as escutarão e atribuirão uma nota ao examinando a partir de uma mesma grade de parâmetros, diferentemente do exame Celpe-Bras, em que o *avaliador-interlocutor* e *avaliador-observador* atribuem cada um uma nota logo após a entrevista, utilizando grades distintas. O estudo aponta que a variação relacionada à interação entre entrevistadores e examinandos é complexa, pois envolve a forma de perguntar, a relação que é estabelecida com o examinando, bem como o estilo discursivo do entrevistador. Outro dado relevante da pesquisa de Brown (2005) é a interação entre as duas variáveis que influenciam no momento de atribuição da nota pelo avaliador: a avaliação da gestão da entrevista e a percepção do desempenho do examinando. A pesquisadora concluiu que os avaliadores (*raters*) tenderam a compensar na nota do examinando, atribuindo-lhe uma nota mais alta, quando o entrevistador pouco contribuiu para que o examinando demonstrasse sua proficiência. O trabalho de Brown (2005) jogou luz sobre as interações complexas entre as facetas da proficiência oral, principalmente, no que se refere à faceta entrevistador.

Com o objetivo de avaliar a atuação e os estilos de entrevistadores do exame Celpe-Bras, Sakamori (2006) analisou 58 provas orais realizadas em três universidades brasileiras. Ela identificou dois tipos de entrevistadores: os colaboradores e os não-colaboradores. Os colaboradores seriam os *avaliadores-interlocutores* que se envolvem na interação, por exemplo, ao fazer comentários sobre a fala do examinando, ao passo que a atuação dos *avaliadores-interlocutores* não-

colaboradores assemelhava-se a de um “perguntador” (SAKAMORI, 2006). A pesquisadora concluiu ainda que nem todos os entrevistadores seguiram as orientações quanto ao tempo e às etapas estabelecidas. Os resultados da pesquisa de Sakamori (2006) contribuem para a compreensão sobre a atuação dos *avaliadores-interlocutores* no exame Celpe-Bras.

Para além da discussão do papel dos *avaliadores-interlocutores* em entrevistas de proficiência oral, na seção seguinte, tratarei dos parâmetros de avaliação oral.

2.3. A DEFINIÇÃO E ORGANIZAÇÃO DOS PARÂMETROS PARA AVALIAÇÃO

Historicamente, as avaliações no campo da psicologia e da educação tiveram foco no desenvolvimento teórico e na sua subsequente verificação empírica; ou seja, são marcadas pela verificação dos construtos teóricos por meio de instrumentos de avaliação. Messick (1987) ressalta que a validação de um teste é um processo de pesquisa e atividade científica, uma vez que investiga-se a relação entre evidência e argumento, isto é, da relação entre a nota e a forma como alguma teoria ou conceito foi operacionalizado em instrumentos.

Bygate (2011) adverte que as teorias sobre metodologia de ensino de línguas estrangeiras (LE) fracassaram na tentativa de elaborar um construto sobre o desenvolvimento da oralidade. O autor defende ainda que, mesmo nas mais diversas formas de encarar a abordagem comunicativa do ensino de línguas, a oralidade ainda é vista mais como um meio do que um parâmetro a ser estabelecido, desenvolvido e atingido. No contexto dos estudos sobre avaliações, tais parâmetros fazem parte do construto. Construto é o que pode ser observado e medido, define Messick (1987); é o grau de habilidade, conhecimento ou capacidade que o examinando possui e que é inferido por meio do seu desempenho em um determinado teste. No documento da AERA (2014), construtos se referem a conceitos ou características que os testes medem, ressalta-se a impossibilidade do construto corresponder a um significado único atrelado aos escores ou ao padrão de resposta do teste.

Por buscar operacionalizar a natureza do construto da oralidade, testes de proficiência oral, como o Celpe-Bras, podem ser úteis para nortear questões centrais na discussão da oralidade como o construto da fala, o construto da tarefa, o parâmetro do desempenho e o construto do desenvolvimento oral da fala (BYGATE, 2011).

Segundo o autor, os testes são instrumentos valiosos de análise para discussão sobre a compreensão de parâmetros de proficiência oral.

Messick (1987) afirma que a representação do construto se refere à relativa dependência do desempenho de uma prova – a oral do Celpe-Bras, por exemplo – aos seus parâmetros de avaliação, tais como a nota do *avaliador-interlocutor*, a nota de *compreensão,* a nota de *pronúncia*, etc. Muito antes de Bygate (2011), Messick (1987) já pontuava que estudar validade de construto pode ser uma maneira de avaliar sistematicamente o papel dos construtos nas teorias não só sobre avaliação em geral, mas também sobre outros campos da educação, como o ensino de línguas.

Fulcher (2003) salienta que, no caso de ser a proficiência oral um construto a ser medido e observado, é necessário que ela esteja associada a algo que possa ser observado e mensurado. Segundo o autor, não há um construto pronto e eficiente sobre proficiência em língua adicional e não há como haver consenso entre os teóricos e professores sobre ele. A definição de um construto, segundo Fulcher (2003), é uma questão de escolher algumas teorias e tentar operacionalizá-las em um contexto de avaliação com seus propósitos específicos, providenciando uma fundamentação teórica e empírica para as escolhas feitas. Para exemplificar, podemos inferir que o conceito teórico da competência interacional foi operacionalizado nas grades de avaliação do Celpe-Bras, e pode haver questionamentos teóricos e empíricos sobre a maneira como o conceito está sendo usado na grade analítica ou holística.

O construto da proficiência oral é um fator que interfere na interpretação do desempenho do examinando. Segundo Brown (2005), no contexto de avaliações que envolvem julgamento como as metodologias EPO, são os parâmetros ou itens que definem o construto; o parâmetro ou item de avaliação seria a metalinguagem para falar da competência, ou seja, eles refletem os aspectos da proficiência que está sendo certificada.

De acordo com McNamara (2004), é a partir da performance em situação de teste que poderemos inferir se o examinando está ou não apto a interagir oralmente em situações reais de fala, ou seja, se ele tem ou não a proficiência oral que se pretende certificar. Por isso, o teste é o meio pelo qual fazemos a inferência e o critério é o alvo das inferências que são feitas por meio do teste. O parâmetro de avaliação é diferente dos critérios de avaliação. Bachman e Savignon (1986 *apud* Fulcher, 2003), afirmam que o critério não pode ser conhecido, porque não se sabe como é ou será a performance real ao usar a língua. Assim sendo, as inferências são mediadas pelos construtos, que são modelos ou conceitos teóricos usados para explicar, representar ou operacionalizar o critério. Por exemplo: se o teste prevê habilidades acadêmicas, é

preciso estudar exatamente o que é exigido dos estudantes estrangeiros na vida acadêmica e social. Assim como Fulcher (2003), McNamara (2004) ressalta que os construtos serão sempre controversos e alvos de críticas, e por isso devem estar articulados com os argumentos que defendem a validade do teste.

Os construtos podem ser baseados em teorias ou podem ser elaborados a partir de uma composição de conceitos. Messick (1987) questiona o fato dos construtos não serem baseados em uma teoria, mas em uma composição de conceitos. Para o autor, é possível investigar a eficiência da composição ao verificar até que ponto as medidas estão avaliando um mesmo construto. No caso do Celpe-Bras, como há duas grades distintas avaliando a mesma coisa, e variados parâmetros que refletem diversos conceitos, é preciso investigar como as notas atribuídas por meio da grade holística e da grade analítica estão relacionadas.

A elaboração de grades nas quais os parâmetros de avaliação são explicitados e graduados em níveis de certificação para nortear a atribuição de escores de entrevistas de proficiência oral é extremamente complexa. Para exemplificarmos essa complexidade, podemos observar o número de descritores da grade analítica do *avaliador-observador* para avaliação da competência interacional no exame Celpe-Bras, que são: desenvoltura e autonomia; contribuição para o desenvolvimento da conversa; uso de respostas breves e uso de estratégias para resolver problemas lexicais, gramaticais e/ou fonológicos.

Niederauer (2014) foi uma das que questionou a operacionalização do conceito da competência interacional que compõe a grade de avaliação da proficiência oral no Celpe-Bras. Ao falar da composição de parâmetros para a avaliação, Fulcher (2003) separa os conceitos de competência interacional e a qualidade das estratégias usadas pelos examinandos para resolver problemas linguísticos. Há um debate teórico em torno do conceito de competência interacional que é relevante para os processos de ensino e avaliação de línguas, porque questionam o peso dos conhecimentos individuais do examinando ao interagir. Há teóricos que defendem que a interação é construída conjuntamente por todos os participantes. No fórum de debate da capacitação de aplicadores do exame Celpe-Bras, por exemplo, a diferenciação entre os níveis *intermediário-superior* e *avançado* a partir dos descritores da nota de competência interacional também esteve em pauta. A moderadora do fórum esclarece a dúvida da seguinte forma:

A principal diferença entre as notas 3 e 4 na competência interacional é que no nível avançado (4), o examinando oferece mais condições para que o interlocutor conduza

a conversa da forma mais natural possível. Nesse caso, o avaliador-interlocutor tem condições de conduzir a entrevista guiando-se mais pela interação, do que pelo roteiro do Elemento Provocador. A interação é conduzida mais pelo examinando do que pelo avaliador. E isso não é observado no áudio 1, no qual a avaliadora precisa alimentar frequentemente a interação para que ela se sustente. (CEBRASPE, 2017)

No comentário acima, é possível observar que o parâmetro ou item adotado para diferenciar os dois níveis envolve fazer uma avaliação do comportamento do *avaliador-interlocutor* ao guiar a entrevista de *forma mais natural possível; mais pela interação, do que pelo roteiro* e não o de gerenciar a entrevista de maneira a *alimentar frequentemente a interação para que ela se sustente.* Alimentar frequentemente a interação, provavelmente com a inclusão de muitas perguntas no roteiro[10], seria uma evidência de que o examinando está contribuindo pouco para a interação.

McNamara (1997) define interação como algo que ao mesmo tempo é sinônimo de vários processos mentais individuais e também de natureza social ou comportamental, em que o comportamento é construído ao longo da performance. Por outro lado, Bachman (1990) se pauta no aspecto individual e psicológico para definir interação. Yoshida e Morise (1998 *apud* Fulcher, 2003) estudaram a relação entre uso de estratégias comunicativas e níveis de proficiência em contexto de EPO e não encontraram evidências de que examinandos mais ou menos proficientes usam menos ou mais estratégias de interação. Até mesmo as paráfrases, consideradas estratégias de compensação, e esperadas para falantes menos proficientes, diminuíam do nível 1+ para o nível 2+ e depois aumentava do 2+ para 3. Os pesquisadores interpretaram que o uso de estratégias de interação está relacionado com a gestão da interlocução e com a natureza assimétrica da interação na EPO.

Nesse sentido, Niederauer (2014) questionou a maneira como o parâmetro competência interacional está graduado na grade analítica do Celpe-Bras, a partir da evidência de uma pesquisa exploratóriarealizada com estudantes de 4 níveis de estudo de português língua estrangeira. Ela chega à conclusão de que os estudantes iniciantes são os que utilizam com mais frequência as estratégias comunicativas. Fulcher (2003) problematiza o uso das estratégias nos descritores das grades por considerar que testar seu uso envolve um processo de inferência sobre os motivos pelos quais o examinando usou ou deixou de usar um certo tipo de estratégia. O autor se baseia em análises de interações em situação de entrevista de proficiência oral a fim de investigar estratégias usadas por diferentes perfis de proficiência de examinandos para sugerir que, caso se opte pelo uso de estratégias comunicativas

[10] Para um debate sobre o uso das perguntas do roteiro na gestão da entrevista, ver Bottura (2014)

como parâmetro de avaliação, é preciso definir bem o construto e pesquisar os procedimentos de atribuição de notas que podem influenciar os variados fatores que afetam o uso de estratégias. Na grade do exame Celpe-Bras, consta em *competência interacional* para os níveis avançados o uso frequente dessas estratégias. Niederauer (2014) contesta a forma como o descritor aparece ao longo dos níveis e propõe uma reformulação da grade. O trabalho contribui principalmente para o debate acerca da graduação dos parâmetros de avaliação, da escolha dos descritores e sua relação com o conceito teórico sobre *competência interacional*. Vale ressaltar, no entanto, que são necessárias pesquisas que verifiquem empiricamente a implicação da proposta da autora na interpretação da grade pelos avaliadores por meio da atribuição de notas, e que levem em consideração os resultados de pesquisa de Fulcher (2003), que diz respeito aos fatores que interferem no uso de estratégias e sua relação com o construto do exame.

Fulcher e Davidson (2007), ao tratar do sistema de atribuição de notas em testes de línguas, afirmam que o processo de julgamento da nota é o que conecta a evidência de performance à tarefa e ao construto. Por isso, assim como Messick (1987), os autores concluem que a questão da validade da tarefa está também relacionada ao processo de atribuição de notas e não apenas com o seu formato. McNamara (1998 *apud* Brown, 2005) ressalta que os fatores envolvidos no processo de atribuição de notas devem ser pesquisados empiricamente com instrumentos de análise adequados.

As grades e descritores desempenham um relevante papel no processo de atribuição de notas, porque dizem respeito ao significado de uma determinada nota é extremamente importante. As grades e descritores são desenhados por um grupo de especialistas com o objetivo de explicitar os componentes de uma determinada habilidade para instrumentalizar o avaliador.

Ao problematizar a construção de grades e graduação dos descritores, Fulcher (2003) afirma que na maioria das vezes as grades são feitas baseadas no argumento de autoridade de um grupo de especialistas e poucas passam por processos de análises empíricas. Sobre o argumento de autoridade, Messick (1987) afirma que em todas as etapas de elaboração do teste

> o julgamento de especialista é com certeza um elemento importante no que diz respeito ao conteúdo e ao formato. Embora a relevância deste tipo de julgamento seja rotineiro no processo de desenvolvimento de testes, a sistematicidade na documentação dos consensos a partir dos vários

> julgamentos está longe de ser um lugar comum nesta etapa de construção de testes. (MESSICK, 1987, 56p.)[11]

Fulcher (2003) ressalta ainda que, muitas vezes, as grades são construídas sem levar em consideração exemplos reais de performance, uma vez que os descritores são definidos, em geral, a partir das expectativas dos elaboradores. Fulcher (2005) divide em duas as metodologias de elaboração de grades, sendo a primeira intuitiva e a segunda empírica. O autor chama de intuitiva as grades elaboradas por especialistas ou comitês e que depois passam por revisão à medida em que foram sendo usadas em processos de avaliação. A segunda metodologia de elaboração de grades de avaliação, denominada empírica, é organizada de três maneiras distintas. A primeira compreenderia as grades baseadas em evidências, ou seja, por meio da análise de respostas às tarefas são feitas descrições de elementos-chave que podem ser observados para fazer as inferências relacionadas ao construto. Outra maneira de elaborar empiricamente uma grade é definindo os limites. Avaliadores experientes dividem performances em boas e ruins e as razões para categorização são registradas e usadas para escrever sequências de perguntas binárias, de sim e de não, que guiam a nota. Uma terceira maneira seria por meio do escalamento de descritores. Os descritores são coletados isoladamente e avaliadores experientes os colocam em ordem de dificuldade, criando as escalas.

Fulcher e Davidson (2007) explicam que a definição dos descritores dos parâmetros graduados nos níveis de certificação é elaborada para dar mais detalhes ao avaliador sobre a habilidade que o teste pretende mensurar. Para detalhar uma grade de avaliação, Fulcher (2003) prevê a descrição da proficiência em níveis para o julgamento do desempenho. Ao operacionalizar o construto linguístico, organiza os descritores por níveis, partindo do mais baixo ao mais alto, de forma a considerar o que o falante é capaz de fazer pela competência de uso dos elementos da língua. Além disso, a grade é usada conjuntamente com a tarefa e com o propósito do teste, sendo que, para ser usada é preciso que os avaliadores recebam treinamento.

Fulcher e Davidson (2007) consideram que uma forma de organizar as grades seria dando uma nota única para cada uma das performances por tarefa, chamada de *primary trait score*. Fulcher (2003) defende que as grades organizadas de forma a gerar uma nota única pressupõem que o julgamento só pode ser feito a partir do contexto de avaliação e que os parâmetros deveriam levar em conta as

[11] “expert judgment is clearly an important ingredient in attesting to content and format relevance. Although relevance judgments are routinely made as an ongoing part of the professional test development process, systematic attempts to documents the consensus of multiple judges are far from commonplace at this test construction stage” (MESSICK, 1987, 56p.)

especificidades da tarefa. Outra maneira seria a *multiple trait score*, cuja atribuição de nota se dá ao dividir cada uma das performances por tarefa ou em um conjunto de performances orais de diferentes tarefas em mais de um parâmetro, e graduando seus descritores ao longo dos níveis de proficiência que se quer atestar. A fragmentação da nota, prevista pelo *multiple-trait score,* pode ser feita por tarefas ou por parâmetros. A definição de *multiple-trait score* de Fulcher (2003) está relacionada com o estabelecimento de um modelo de atribuição de nota que em Messick (1987) é definido como modelo combinado. Nesta situação, a nota de cada uma das tarefas poderia compor a nota final ou a nota poderia ser fragmentada a partir da divisão dos parâmetros de avaliação de uma mesma tarefa. Por exemplo, no caso do Celpe-Bras, se considerássemos que o quebra-gelo é uma tarefa distinta dos outros quinze minutos, uma nota poderia ser dada esse momento da entrevista. Segundo Fulcher (2003), cada nota representaria um aspecto da performance que traduziria a capacidade de falar de si e a capacidade de falar sobre os assuntos dos EPs, no caso do Celpe-Bras. Mesmo sendo uma nota única para o desempenho durante toda a entrevista, o modelo que vigora para atribuição de nota da prova oral no exame Celpe-Bras é o composto ou *multiple-trait,* uma vez que a nota final é composta por outras notas atribuídas por cada um dos sete parâmetros de avaliação. Outra forma de separar a nota seria atribuindo diferentes valores a diferentes aspectos do construto que subjaz à proficiência oral. No caso da avaliação do Celpe-Bras, o *avaliador-observador* atribui uma nota para cada aspecto da proficiência oral. Por isso, a nota analítica é um exemplo do que o autor chama de *multiple-trait score*.

Fulcher (2003) salienta que, por ser impossível atribuir nota para todo e qualquer construto da proficiência oral já teorizado, as escalas múltiplas ou analíticas oferecem a oportunidade de incorporar diferentes construtos na nota e, por isso, podem ser bastante úteis para o fornecimento de informações diagnósticas para o examinando. O desafio de se usar uma escala analítica como a do *avaliador-observador*, segundo o mesmo autor, é atribuir diversas notas a uma mesma performance. Isto pode ser um complicador para o avaliador pela dificuldade de distinção dos construtos, e por isso se justifica o investimento em treinamento de avaliadores para que eles saibam distingui-los. No comentário abaixo do fórum de dúvidas do curso de capacitação, um cursista levanta a questão de como relacionar a evidência da performance *interferências do francês na pronúncia e na regência dos verbos* nos parâmetros de avaliação da grade.

Também tive problema nesse áudio. Dei 5 para a Adequação Lexical e 4 para a Adequação Gramatical, quando a resposta da AL era 4 e, da AG, 5. A justificativa dada para que a nota da AL fosse 4 foi que o examinando tem interferências do francês na

*pronúncia e na regência dos verbos - mas entendo que a pronúncia deve ser avaliada no quesito Pronúncia e, a regência, no quesito Adequação Gramatical... No entanto, na AG foi avaliada como 5 e a justificativa é que as inadequações linguísticas são raras. Ora, isso me parece contraditório: como um tópico gramatical interfere sobre a nota da Adequação Lexical, mas não interfere sobre a nota da própria Adequação Gramatical? Ademais, apesar de ocorrer pequena interferência, acho que a pronúncia do examinando é adequada - o que também achei contraditório, já que, no quesito Pronúncia, é avaliada como 5, ao contrário do que é dito na justificativa da avaliação do quesito AL. (INEP/*CEBRASPE, 2017)

No comentário acima, a avaliadora-cursista questiona a justificativa de atribuição de nota para *adequação lexical, adequação gramatical e pronúncia*. Segundo o comentário, uma evidência de desempenho que se refere à *adequação gramatical*, uso de regência verbal, justificaria a interpretação de uma nota no parâmetro da *adequação gramatical* e não em outro parâmetro, como o da *adequação lexical,* conforme ela questiona. Além disso, ela questiona o fato de que, na justificativa de atribuição da nota fornecida pelos organizadores do curso, a *pronúncia* ter sido parâmetro para avaliar a *adequação lexical*. É importante ressaltar que a cursista apresenta uma divergência de apenas um ponto entre a nota atribuída pelos organizadores do curso, ou seja, não chega a configurar uma discrepância que envolva reavaliação, por isso as notas são convergentes.

Eckes (2015) afirma que, no contexto de performance em que a atribuição de nota é mediada pelos avaliadores, os parâmetros de avaliação das grades analíticas, bem como seus níveis de proficiência são também uma questão que pode interferir sistematicamente na atribuição de notas. Segundo o autor, os parâmetros podem ser definidos de forma que seja pouco provável que os examinandos atinjam o nível máximo, por exemplo, ou de forma que os avaliadores não distingam ou confundam um construto de outro, como o comentário acima exemplificou.

Além das duas formas de organizar a composição da nota, *primary trait score* e *multiple-trait score,* há também a atribuição de notas de maneira holística, na qual uma impressão geral da performance guia a nota que pode ser ou não norteada por uma escala de descritores. Segundo Fulcher (2005), a proficiência oral é complexa demais para caber em uma descrição geral, por isso as grades holísticas podem ser potencialmente problemáticas por não levar em conta os construtos que constituem a proficiência oral, mas apenas a noção geral de proficiência oral que, segundo o autor, não está resolvida do ponto de vista teórico. Por outro lado, Fulcher e Davidson (2007) ressaltam que as grades holísticas são mais práticas, pois exigem menos tempo de

avaliação. Já as grades múltiplas ou analíticas são mais dispendiosas, porém geram mais informações sobre a performance ao avaliar distintos construtos separadamente. Fulcher (2003) sugere que, em uma avaliação diagnóstica, usar escala de níveis, como são as grades holísticas, talvez não faça sentido e neste caso seria preferível a descrição de uma somatória de coisas que o examinando pode fazer com seus devidos pesos para o cálculo da nota final.

No caso do Celpe-Bras, há duas grades de avaliação para o julgamento da proficiência oral: a grade do *avaliador-interlocutor* que prevê atribuição de uma nota única para a performance, ou seja, trata-se de uma grade composta por um item, e a grade do *avaliador-observador,* composta por seis parâmetros de avaliação ou seis itens. São eles: *compreensão oral, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia* (ANEXO 1). Os pesos dos parâmetros são distintos para o cálculo da nota final do *avaliador-observador,* conforme já discuti em seção anterior.

Vale ressaltar que os componentes das habilidades refletidas nas grades são embasados no construto teórico que fundamenta o exame. Fulcher e Davidson (2007) problematizam que como os descritores são elaborados de forma a tentar prever todos os tipos de performance, frequentemente os avaliadores têm dificuldade de relacionar um determinado desempenho a uma nota. Os autores afirmam que a compreensão coletiva dos descritores é o que assegura a validade da maneira como o construto é interpretado na organização da grade. Nesse sentido, a confiabilidade de atribuição de notas entre os avaliadores é uma forte evidência para validade de um teste. Por outro lado, Fulcher (2003), ao questionar a confiabilidade da grade desenvolvida pelo FSI, aponta que não basta os avaliadores se tornarem adeptos à maneira como se utiliza a grade, é preciso avaliar também a qualidade da grade em si, ou seja, da graduação dos descritores ao longo dos níveis de proficiência. No caso do Celpe-Bras, avaliar a qualidade da grade em si implica em avaliar também as suas qualidades psicométricas ou o significado de cada uma dessas notas por meio de modelos estatísticos. Isso quer dizer: é preciso analisar a capacidade de cada um dos parâmetros de discriminar os examinandos ao longo das faixas de proficiência que se pretende certificar.

O debate sobre as qualidades de uma grade estão diretamente relacionados aos conceitos de validade e confiabilidade dos quais tratarei a seguir.

3. VALIDADE E CONFIABILIDADE EM TESTES DE LÍNGUAS

Retomando a metáfora de McNamara (2000) que apresentei na introdução desta tese, o autor aproximou o processo de elaboração e verificação da qualidade de um instrumento de avaliação à construção de um carro. Para o autor, colocar um teste para funcionar é como colocar um carro na rua: envolve muitas etapas de elaboração com inúmeras possibilidades de averiguar a eficiência do instrumento. Para os psicometristas, a averiguação da eficiência do método é chamada de validade. No contexto das pesquisas psicométricas é também concenso que as evidências de validade são variadas e trata-se de um processo interminável.

Especialistas em avaliação de línguas tendem a separar as noções de validade das de confiabilidade no contexto dos exames. Neste trabalho, apoio-me na Teoria de Validade proposta por Messick (1987) na qual a validade de construto é o foco das discussões de validade e confiabilidade. McNamara (2004), ao tratar das teorias que fundamentam o campo da avaliação em línguas, cita a importância de Messick (1987) nesses estudos de Linguística Aplicada que tem como tema a validação de exames de línguas. Neste capítulo, trato da teoria de validade de Messick (1987) bem como a relação de suas ideias com as dos especialistas da área de avaliação em línguas.

Messick (1987) define validade como “o julgamento integrado e avaliativo baseado em estudos empíricos e teóricos que demonstram o quanto as inferências e ações são

adequadas e apropriadas a partir das notas do teste."[12] (MESSICK, 1987, 6p.). A validade é sinônimo de julgamento feito com base em evidências diversas que podem advir de estudos tanto teóricos quanto empíricos para demonstrar a coerência de uma ação tomada a partir da nota de um teste. O que é notável na teoria de Messick (1987) é que a noção da validade não se esgota no estudo da nota, nem na sua relação com a forma como o teste é organizado internamente. Além disso, a teoria amplia o debate para o quanto o instrumento gera de informação pertinente para que as ações sejam tomadas. Ao considerar a pertinência das ações a serem tomadas a partir da nota do teste, o autor inclui os aspectos externos ao exame na agenda de pesquisa sobre validade.

Messick (1987) discute a relação entre *teste* e *nota.* Em sua teoria de validade, ele enfatiza que a nota do teste deve ser o foco da investigação da validade porque o que precisa ser validado não é o teste como ele é, mas as inferências que são feitas a partir dos seus resultados. Para chegar nas inferências é preciso coletar evidências que, para o autor, estão concentradas nas notas. A ênfase é dada à nota porque as propriedades de validade e confiabilidade estão presentes nas respostas ao teste e não no teste em si. Estudar o teste em si, bem como os argumentos de especialistas que fundamentam o uso de um determinado desenho de avaliação é relevante. No entanto, para fins de validação, na perspectiva do autor, tais argumentos devem ser combinados com evidências empíricas, por meio do estudo da nota. Este é um ponto de divergência na maneira de compreender a validade por alguns especialistas em avaliação de línguas da qual trato a seguir.

Para exemplificar a importância do estudo da nota em pesquisas sobre avaliação de línguas, retomamos o trabalho de Brown (2005). Embora o foco de Brown (2005) tenha sido na condução da entrevista, a autora investigou o impacto que causam na nota algumas maneiras de entrevistar. Além de analisar a estrutura da entrevista sob a luz da Análise da Conversação, a autora conduziu uma análise das notas, utilizando-se do Rasch Multifacetado, em que são calculadas estatisticamente as variações de atribuição de nota de um específico avaliador e dos avaliadores entre eles. Por meio da análise das notas, foi possível avaliar a variação de notas de um determinado avaliador e a variação de notas entre diferentes avaliadores. A pesquisa não termina na análise das entrevistas, nem na argumentação da pesquisadora sobre a qualidade da condução da entrevista, porque o trabalho alia as duas metodologias e avalia empiricamente o comportamento de condução da entrevista e seu impacto na nota final do examinando. Nesse sentido, a pesquisa de Brown (2005) também traz

[12] "an integrated evaluative judgement of the degree to which empirical and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". (MESSICK, 1987, p.6)

contribuições sobre a validade da EPO, porque analisa a faceta entrevistador no processo de atribuição de nota.

Outro aspecto a se considerar na teoria de Messick (1987) diz respeito à validade das ações a serem tomadas e que são externas à estrutura do teste. O autor amplia o conceito de validade para além dos instrumentos de avaliação e inclui nos processos de validação o debate sobre a pertinência das consequências sociais geradas pelo instrumento. Nesse sentido, o autor diferencia a interpretação do teste do seu uso. Interpretação do teste diz respeito à eficiência do teste ao certificar a proficiência dos examinandos. O uso do teste diz respeito à pertinência dessas certificações para processos de decisões sobre a vida do examinando. No caso do Celpe-Bras, o nível de proficiência do candidato pode ser um dos motivos ou requisitos em um processo de pedido de reconhecimento de diploma para exercer sua profissão no Brasil.

Tradicionalmente, o estudo da validade abarca a relação entre significado da nota até sua relação com o arcabouço teórico que a fundamenta. No entanto, Messick (1987) estende ainda mais o conceito de validade de forma que o uso do teste seja também justificado: "o significado da medida e portanto, a validade do seu construto, devem sempre ser justificada – não apenas para fundamentar a interpretação do teste, mas para justificar o seu uso"[13] (MESSICK, 1987, 16p). Se a validade verifica a pertinência do instrumento para um determinado uso, Messick (1987) alega que chega a ser irônica a forma como historicamente os estudos de validade não contemplavam as consequências sociais das avaliações analisadas.

Assim como não há teorias isentas de valores, não há construtos e instrumentos neutros. Embora questões políticas e sociais determinem também os valores implicados na interpretação do resultado do teste, no contexto da proposta de Messick (1987), tais implicações de valores presentes na interpretação do teste e suas potenciais consequências sociais estão fortemente relacionadas com o significado do construto no contexto do exame. Novamente, a questão central é investigar o significado da nota pois os valores sociais não estão subordinados ao resultado, mas integrados aos escores. Dito de outra forma: as potenciais consequências sociais estão integradas ao significado da nota.

O processo de validação de testes é contínuo porque as evidências são sempre incompletas. Para Messick (1987), a validade é uma questão de fazer o maior esforço possível para se entender o significado da nota, apoiando-se em evidências para guiar a interpretação e o uso do teste. Além disso, a validade não é questão de

[13] "The meaning of the measure, and hence its construct validity, must always be pursued – not only to support test interpretation but to justify test use." (MESSICK, 1987, 16p.)

tudo ou nada, de forma que um instrumento validado hoje pode ser contestado por novas evidências no futuro e é por isso, também, que a validade é um processo contínuo.

Nesta perspectiva teórica, a validade é sinônimo de investigação científica da inferência. Para Messick (1987), a inferência é uma hipótese e, por isso, validar uma inferência é verificar uma hipótese. Cabe ressaltar que, ao revisar debates sobre os paradigmas da filosofia da ciência e sua relação com as teorias sobre validade, Messick (1987) afirma que a validade não está atrelada a uma perspectiva específica, e exemplifica como sua teoria incorpora perspectivas de variados paradigmas científicos. Ao assumir que há variadas maneiras de fundamentar a validade e que os significados são construídos a partir de uma teoria que não é neutra, o autor incorpora novos debates da perspectiva do paradigma relativista das ciências. “A validade abarca tanto os meios experimentais e estatísticos quanto os filosóficos pelos quais hipóteses e teorias científicas são formuladas”.[14] (MESSICK, 1987, p.6)

Embora o autor contemple várias fontes e evidências para fundamentar as inferências feitas a partir da nota, a validade é definida como um conceito único e pode ser fundamentada por evidências fortes e fracas, a depender do procedimento de análise.

3.1. PERSPECTIVA HISTÓRICA DO CONCEITO DE VALIDADE E SUA RELAÇÃO COM A AVALIAÇÃO EM LÍNGUAS

Historicamente, havia variados tipos de validade, a saber: validade de conteúdo, validade de critério, validade de previsão, validade concorrente ou validade paralela e validade de construto. Alguns linguistas aplicados, como Hughes (1989) se fundamentam nesta perspectiva da validade. Descrevo abaixo cada um desses tipos de validade, relacionando com alguns autores especialistas em avaliação de línguas e apresentando a perspectiva da validade de Messick (1987), que se difere da definição tradicional ou histórica do termo e para quem a validade é um conceito unitário.

A validade de conteúdo ou validade de face diz respeito à relação do conteúdo do teste com as situações da sala de aula ou com as situações às quais são importantes para as conclusões a serem tiradas a partir do teste. O julgamento de especialistas sobre a relevância do conteúdo do teste e sobre a pertinência ou representatividade de cada uma das tarefas do teste fornece evidências que fundamentam a relevância e

[14] “test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated”. (MESSICK, 1987, 7p.)

a representatividade da tarefa para gerar informações pertinentes. A finalidade desse julgamento é a de atestar a proficiência que o teste define. No entanto, tais julgamentos não geram evidências que fundamentam as inferências feitas a partir da nota. Isto ocorre porque eles não contemplam na análise a resposta ao teste, a estrutura interna e externa do teste, as diferenças no desempenho e as consequências sociais da nota, dentre outros fatores. Ainda que a relevância e a representatividade da tarefa influenciem a natureza da nota, Messick (1987) considera que a validade de conteúdo não é qualificada como um estudo de validade propriamente dito, uma vez que determinar o que está sendo medido exige outros tipos de análises.

Essa visão não é unânime entre os especialistas em avaliação em línguas. McNamara (2000), por exemplo, alinha-se à perspectiva de Messick (1987) de que a avaliação do conteúdo sem levar em conta as notas não constituem exatamente uma validação. Para outros autores, como Hughes (1989), Fulcher (2003) e Brown e Abeywickrama (2010) a aparência do teste, ou seja, a validade de face constitui um aspecto da validade. Hughes (1989) afirma que a validade de face se trata do quanto o teste parece medir o que ele se propõe. Na afirmação, o próprio autor sinaliza que embora um teste pareça medir algo, isto não especifica o quê o exame esteja medindo na prática. Brown e Abeywickrama (2010) apresentam diretrizes para investigar a validade de face, a saber: boa elaboração das tarefas, plausibilidade de execução da tarefa no tempo previsto, clareza dos itens, clareza dos comandos e instruções dos itens, ineditismo da tarefa, relação da tarefa com o conteúdo do curso ou com o uso futuro da língua, nível de dificuldade apropriado. Tais diretrizes baseiam-se em análises descritivas, ou seja, no argumento de especialistas. Normalmente um ou mais especialista analisa e julga os aspectos envolvidos na validade de face. A análise descritiva do instrumento pode sinalizar o que o teste parece ser. No entanto, para avaliar o que o teste é realmente é preciso verificar seus resultados e relacioná-los às avaliações descritivas. Exemplificando, quanto ao nível de dificuldade das tarefas, muito pouco se pode afirmar a partir do aspecto da tarefa; é preciso pilotar os itens. Por isso, recomenda-se uma aplicação piloto para coletar notas e consequentemente realizar ajustes não só de comandos, mas também de todos aspectos citados acima.Ou seja, é a partir da averiguação das respostas e das notas que se pode analisar as medidas relacionadas à dificuldade do item, coeficientes de confiabilidade, etc.

Fulcher (2003) afirma que o argumento de especialistas quanto ao aspecto do teste seria um argumento de validade, porque é preciso que alguma comunidade de especialistas esteja de acordo com o desenho da prova. No entanto, o autor problematiza o uso exagerado do recurso de validade de face.Há quatro argumentos

que endossam a validade de face da EPO que merecem ser discutidos, a saber: (1) o fato dos testes orais serem diretos; (2) o fato de envolverem contexto natural com tarefas da vida real que podem conduzir a uma conversa natural; (3) a afirmação de que a escala descreve e discrimina os níveis de desempenho oral; e (4) a afirmação de que a experiência dos avaliadores conduz à aplicação consistente dos parâmetros de avaliação. Quando Messick (1987) propõe que os argumentos fundamentados no aspecto do teste não devam ser compreendidos como uma validação, no sentido estreito do termo, o autor está propondo que outros estudos sejam conduzidos para averiguar empiricamente, por meio da análise das notas, a pertinência das afirmações fundamentadas na avaliação da aparência do exame. Nesse sentido, ao elencar os pressupostos que carecem de mais evidências, o próprio Fulcher (2003) sinaliza a fragilidade da qualidade de evidências coletadas para sustentar o que ele denomina validade de face ou de validade de conteúdo.

Um segundo tipo de validade na perspectiva tradicional do conceito é a validade de critério. A validade de critério tem o foco na relação entre a nota do teste com uma ou mais variáveis externas consideradas relevantes para o que se quer medir, que são chamadas de critério. Cabe diferenciar aqui que critério se refere ao aspecto da proficiência a ser mensurada e não ao parâmetro de avaliação[15]. Segundo McNamara (2004), o critério não pode ser conhecido, porque não se pode determinar com certeza o uso da língua daquele examinando na vida real, mas é possível prevê-la ou apontar probabilidades de sucesso ou fracasso no desempenho da língua. Por isso, as inferências são mediadas pelos construtos, que são modelos teóricos para explicar, representar e operacionalizar o critério. Os parâmetros de avaliação operacionalizam o construto que, por sua vez, operacionaliza o critério. Por exemplo: a escrita acadêmica em inglês poderia ser o critério de um exame que tenha como propósito avaliar a aptidão de brasileiros para participar em programas de mobilidade acadêmica em países cuja língua oficial é diferente do português. Vamos supor que, em geral, os discentes já apresentassem um bom domínio oral dessa língua ou que não fosse possível avaliar a habilidade oral, e a escrita acadêmica fosse o foco da inferência para decidir quem estaria apto a estudar em uma universidade fora do país. Os elaboradores teriam que estudar e escolher uma teoria ou perspectiva teórica de aquisição de língua estrangeira, ou ainda conceitos de escrita e/ou desenvolvimento de escrita acadêmica para fundamentar o desenho e os parâmetros de avaliação. Os desenhos e os parâmetros de avaliação devem estar coerentes com a perspectiva teórica, ou seja, com o construto que está representando o critério de saber escrever

[15] Para um debate sobre a diferenciação entre parâmetro e critério, veja a seção 2.3 desta tese.

satisfatoriamente nas situações determinadas. McNamara (2004) resume a relação entre teste, construto e critério no quadro a seguir:

QUADRO 2

Teste, construto e critério

teste		construto		critério
desempenhos e respostas às tarefas ou itens	desenho do teste	caracterização dos parâmetros essenciais do desempenho, teoria sobre o domínio		desempenho no mundo real, o que o examinando realmente faz no mundo real
observável	inferências	via modelos teóricos	sobre	não-observável

Fonte: McNamara, 2004, 765p.

De acordo com o quadro acima, as respostas às tarefas são manifestações observáveis para gerar inferências via modelos teóricos sobre o critério, que, por sua vez, seria algo não observável, porque se refere à potencialidade do examinando fazer algo na língua no futuro. A mediação entre a caracterização dos parâmetros essenciais do desempenho – ou seja, do construto – e o teste é feita por meio das respostas e, em última instância, por meio das notas.

Outro tipo de validade tradicionalmente classificada é a validade de previsão, que indica em que medida o uso futuro do que foi medido foi previsto pelo desempenho que o examinando teve em um determinado teste. Hughes (1989) associa este tipo de validade aos testes de nivelamento, no sentido de avaliar o quão eficaz é o instrumento para nivelar os estudantes. Segundo o autor, analisar a validade de previsão implica verificar o quanto os estudantes foram nivelados via teste de maneira equivocada para, em seguida, verificar se vale a pena continuar com o processo de nivelamento.

Por usa vez, a validade concorrente indica em que medida a nota está relacionada com o que o examinando de fato faz quando comparado ao seu desempenho em variados formatos de teste. A validade concorrente pode ser interpretada como um tipo de validade de critério ou estar sobreposta a ela.

O último tipo de validade, na perspectiva tradicional, é a de construto, que é estudada investigando-se as qualidades da medida. Isso significa procurar saber quais são as qualidades do desempenho que estão sendo medidas, determinando o quanto cada conceito ou construto representa ou explica o desempenho do examinando no

teste. Avaliar as qualidades psicométricas da medida está diretamente relacionado com o aspecto da validade de construto. Qualquer evidência que esteja relacionada com a interpretação e significado da nota diz respeito à validade de construto: o construto remete à proficiência ou ao fator que é responsável pelas relações entre os indicadores. As medidas são entendidas como um conjunto de indicadores que estão relacionadas com o construto. Tais indicadores muito frequentemente estão em uma relação probabilística entre eles mesmos e entre o construto. Messick (1987) entende que não tem como o construto ser explicitamente definido, pois trata-se de um conceito aberto uma vez que há uma dependência com a maneira como ele está sendo operacionalizado no contexto de um instrumento.

Messick (1987), ao revisar variados documentos que estabeleciam diretrizes para validação de testes no contexto educacional e psicológico dos Estados Unidos, verificou que o foco das discussões era a validação empírica e sua relação com as evidências relacionadas ao critério e ao construto. O autor enfatiza que, ao longo da discussão, os termos sobre o tipo de validade (concorrente, de previsão, etc.) foram sendo substituídos nestes documentos pela exploração dos procedimentos de validação. Ele explica que, muitos autores, em muitos documentos institucionais, chegaram a conclusão de que todos os procedimentos de validação discutiriam a validade de construto. Messick (1987), portanto, fundamenta sua visão de validade como um conceito unitário a partir da revisão histórica do conceito de validade, e chega à conclusão de que se trata de uma coisa só, embora envolva variados procedimentos. Nesse sentido, ao pressupor que o significado da nota está implicado pela validade de construto, a validade se torna um conceito único. A validade de construto, que subjaz às inferências que são feitas a partir do resultado do teste, está incorporada no significado da nota.

Entre os especialistas em avaliação de línguas, parece não haver um consenso sobre o que é, como se investiga e como se estrutura a validade. Hughes (1989) se pauta na divisão tradicional do conceito de validade e cita os tipos de validade da perspectiva tradicional. Bachman e Palmer (1996) colocam a validade como uma das qualidades dos testes de línguas, ao lado de aspectos como praticidade, confiabilidade, autenticidade das tarefas e do formato do teste, etc. Bachman e Palmer (1996) ao afirmarem que a validade de construto 'está relacionada à pertinência das interpretações que fazemos com base nas notas do teste"[16] (BACHMAN E PALMER, 1996, 21p.) se aproximam da proposta de Messick (1987). Assim como Bachman e Palmer (1996), Brown e Abeywickrama (2010) apresentam a validade como um dos princípios que devem nortear a avaliação dos teste. Tais princípios envolvem a análise

[16] "pertains to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores". (BACHMAN E PALMER, 1996, p.21).

da praticidade, confiabilidade, autenticidade do desenho, efeito retroativo do exame no ensino da língua e a validade. Ainda que os autores definam a validade citando a definição de Messick (1987), os especialistas em avaliação de línguas apresentam dentre os tipos de validade não só a de construto, mas também a de critério, conteúdo, de face e a validade consequencial, que diz respeito às consequências sociais do teste. Como Messick (1987) adota uma perspectiva conceitual unitária do termo de validade, de forma que as evidências se referem à validade de construto, Bachman e Palmer (1996) assim como Brown e Abeywickrama (2010) parecem se fundamentar apenas em partes na teoria de validade proposta por Messick (1987). Já Fulcher (2003), embora também considere a análise do aspecto do teste como validade de face, se fundamenta fortemente na teoria de validade de Messick (1987) e inclusive resenha os aspectos da validade de sua proposta.

No contexto da teoria de Messick (1987), a validação de construto é o ponto em que convergem as evidências de validade. Tais evidências podem estar fortemente ou fracamente relacionada ao construto. O construto e a sua operacionalização estão entrelaçados, por isso as evidências que fundamentam as notas embasam também o instrumento.Qualquer evidência de que as notas não fundamentam o construto dizem respeito também à ineficiência do instrumento; e neste caso, é preciso decidir se é o construto e ou a medida que deverão ser revistas.

Deve-se levar em conta que os testes são imprecisos por possibilitarem erros aleatórios. Testes são exemplares imperfeitos de um construto que se quer validar. Em assim sendo, variados tipos de evidências são necessárias para sua validação. Messick (1987) as divide em duas, sendo que a primeira evidência se refere ao grau de implicação do construto na nota, e a outra diz respeito ao grau de implicação entre nota-construto não estar relacionada com outros construtos alternativos.

A sub-representação do construto se conecta a esta primeira evidência que ameaça a validade do teste, porque se refere à possibilidade do teste excluir da relação construto-nota aspectos importantes do construto. Para exemplificar, podemos considerar um teste oral comunicativo em que apenas a pronúncia seja avaliada. Nele é provável que o construto da proficiência oral esteja sub-representado porque há outros aspectos relevantes do construto que não estavam sendo considerados nos parâmetros de avaliação como o da fluência, da competência interacional, dentre outros. Neste caso, pode haver uma sub-representação da teoria ou conceitos que dizem respeito à proficiência oral, no contexto das discussões sobre ensino de línguas na perspectiva comunicativa, em função da operacionalização desses construtos nos parâmetros de avaliação. Outra ameaça à validade de construto é o que Messick

(1987) denomina de variância irrelevante para o construto do teste. Trata-se de uma variação na nota que não tem a ver com a proficiência oral que o examinando demonstra, mas com aspectos ocultos no instrumento. Variância irrelevante para o construto do teste diz respeito à contaminação da nota por causa de algum aspecto que não interessa ser medido, que é irrelevante para o construto que o teste está operacionalizando. Na situação da EPO, por exemplo, pode ser que a simpatia ou timidez possam estar influenciando a nota.

No contexto da discussão de validade por especialistas em avaliações de línguas, Bachman (1990) também afirma que uma das preocupações fundamentais a respeito do desenvolvimento de avaliações é a identificação de fontes potenciais de falhas em um determinado processo de atribuição de notas e a criação de estratégias para minimizar os efeitos desses erros nos escores. No entanto, diferentemente de Messick (1987), Bachman (1990) aproxima o estudo do significado da nota ao conceito de confiabilidade, ao afirmar que quanto mais os efeitos das falhas são minimizados, mais confiável é um teste. Bachman (1990) coloca os conceitos de validade e confiabilidade nas extremidades de um contínuo, ao passo que Messick (1987) determina que o foco das discussões sobre a nota é a validade de construto, incluindo aspectos sobre a confiabilidade.

De acordo com Bachman (1990), em pesquisas cujo foco seja a compreensão da confiabilidade de testes, a pergunta que se busca responder é: *até que ponto a performance do indivíduo no teste é fruto de uma falha na medida ou de outros fatores externos à habilidade linguística que se quer avaliar?* Os trabalhos com o foco na confiabilidade devem envolver análise lógica e pesquisa empírica de maneira a primeiro identificar as fontes de falhas e, em seguida, estimar o quanto essas falhas influenciam nas notas finais. Nas palavras do autor, "estritamente falando, a confiabilidade se refere à nota atribuída e não ao teste em si."[17] (BACHMAN, 1990, p.171, grifo nosso). Por isso, ele sugere que as investigações devam analisar se as interpretações e usos das notas são válidos.

Cabe ressaltar que Bachman (1990) utiliza-se de duas perguntas distintas para diferenciar os conceitos de validade e confiabilidade, que referem-se às falhas na medida e sobre como a medida está relacionada ao construto. No entanto, as respostas a essas duas perguntas compõem o que entendo por validade de construto. Além disso, as duas perguntas estão diretamente relacionadas à busca de duas evidências cruciais para validação de construto na teoria do Messick (1987), sendo a primeira a sub-representação do construto, e a segunda, a variância irrelevante ao

[17] "Strictly speaking, reliability refers to the test scores, and not to test itself" (BACHMAN, 1990, p.171).

construto. Vale destacar que na revisão recente do conceito de validade publicada em 2014 pela AERA, as noções de sub-representação do construto e variância irrelevante ao construto compõem as diretrizes para compreensão do processo de validação do testes. A evidência importante no estudo da validade está relacionada com o que Bachman (1990) classifica como estudo de confiabilidade, que se refere à possibilidade da medida estar contaminada por construtos alternativos ao que se quer atestar, ou seja, trata-se da variância irrelevante ao construto. Para Bachman (1990), ao definir análise de nota, o autor pontua que esta diz respeito à confiabilidade. Em outros momentos do seu texto, o autor explica que uma das formas de analisar a validade também é por meio de análise de notas que envolvam lidar com outras evidências que devem emergir de estudos correlacionais e experimentais. Neste caso, o que se buscaria estabelecer é a relação entre as tarefas de um mesmo teste, a relação entre um teste e outro, a correlação de mais de um teste que mede a mesma habilidade, dentre outros.

Hughes (1989), Bachman e Palmer (1996), Brown e Abeywickrama (2010), todos linguistas, relacionam também análise de notas com o conceito de confiabilidade e distinguem confiabilidade da validade, seguindo a mesma linha de raciocínio de Bachman (1990). No entanto, Bachman (1990), sobre a validade de construto, parece corroborar com as ideias de Messick (1987) ao afirmar que “quando nós definimos os construtos operacionalmente, nós estamos estabelecendo hipóteses sobre a relação entre os construtos e os escores do teste que podem ser vistos como a manifestação do construto" (BACHMAN, 1990, 257p.).[18] Hughes (1989), Bachman e Palmer (1996), Brown e Abeywickrama (2010) também relacionam a validade de construto com a interpretação da proficiência, e Fulcher (2003) se apoia explicitamente na Teoria de validade de Messick (1987) ao afirmar que a validade é um argumento. Na esteira das ideias de Messick (1987), Fulcher (2003) afirma que

> a validade não é uma questão de ‘tudo ou nada’, é uma atividade constante para fundamentar um argumento e reunir provas para apoiar o argumento. Decidir ‘se um teste é apropriado para um propósito específico’ envolve avaliar criticamente tanto a plausibilidade do argumento como as evidências usadas para fundamentá-lo (FULCHER, 2003, p.171).[19]

[18] “When we operationally define construct as measures of language ability, we are making hypotheses about the relationship between constructs and test scores, which can thus be viewed as behavioral manifestation of the construct.” (BACHMAN, 1990, 257p.)

[19] “Validity is not an ‘all or nothing’ affair, it is an ongoing activity to improve an argument and gather evidence to support the argument. Deciding ‘if a test is appropriate for a particular purpose’ involves critically evaluating the plausibility of the argument and the evidence used to support the argument.”(FULCHER, 2003, p.171)

Fulcher (2003) entende o processo de validade como uma construção de uma argumentação que é contínua e incompleta, cujas evidências que a fundamentam devem ser avaliadas. Messick (1987) enfatiza o fato de que a validade seja uma avaliação integrada ao julgamento do quanto determinadas evidências fundamentam o teste e o uso do mesmo. Ressalto, portanto, a ênfase dada, na perspectiva de Messick (1987), na qualidade das evidências e, por isso, o foco da análise deve ser o da qualidade da evidência e não do argumento em si. O atual documento da American Educational Research Association (AERA, 2014), revisando o conceito de validade, salienta que o tipo de evidência para fundamentar o uso que se faz do teste está relacionada com os pressupostos implicados na interpretação de seus resultados.

3.2. INTERPRETAÇÃO DO TESTE E SEU USO COM BASE NA TEORIA DE VALIDADE DE MESSICK

Messick (1987) apresenta um esquema (QUADRO 3) de como a interpretação do teste e seu uso interagem com as bases evidenciais e consequenciais nos estudos de validade. A interpretação do teste está relacionada às bases evidenciais da validade, pois geram informação sobre a validade de construto. As bases evidenciais apoiam o uso do teste, uma vez que se referem à relevância ou utilidade do instrumento. Já as bases consequenciais do uso do teste estão relacionadas às consequências sociais. As bases consequenciais da interpretação do teste são os valores implicados na formulação de tais construtos, uma vez que as teorias não são neutras. No caso do exame Celpe-Bras, as bases evidenciais que apoiam a interpretação do teste podem ser a eficiência de discriminação das faixas, que diz respeito à sua validade de construto. A validade de construto está, por sua vez, atrelada a um conjunto de valores sobre qual a Língua Portuguesa que está sendo avaliada, bem como e o quê dela deve ser avaliado. O estudo das faixas de definições da proficiência e sua relação com a composição da nota com o construto referem-se à interpretação do teste. E como cada definição contribui para informar às instituições brasileiras de ensino se o examinando está apto ou não a seguir seus estudos no Brasil, diz respeito também ao uso do teste.

McNamara (2000) corrobora a noção de validade de Messick (1987) e afirma que as avaliações, ao exercerem um papel e um propósito que refletem os objetivos da política institucional, são institucionais.. Nesse sentido, o teste não é fruto de atividade técnica e científica apenas, mas também deve ser elaborado tendo em conta seu

efeito na vida dos examinandos, os seus impactos na sala de aula, bem como no sistema de ensino como um todo. É preciso levar em conta os efeitos desejados e não-desejados na elaboração e na aplicação. Numa perspectiva crítica da avaliação de línguas, McNamara (2000) afirma que as avaliações são usadas como instrumento para as elites se manterem no poder e o teste é um projeto intelectual de dominação. O processo de desenvolvimento das grades de avaliação das EPO no âmbito das políticas higienistas norteamericana exemplificam como os testes podem estar a serviço das políticas de Estado. O próprio Celpe-bras também está a serviço do Estado Brasileiro, ao fazer parte de um conjunto de instrumentos de promoção da língua portuguesa e programas de cooperação educacional, especialmente na América do Sul. Messick (1987) afirma que os valores implicados na interpretação e uso do teste estão relacionados não só com a validade de construto, mas também com seu contexto institucional e político. Para fins de validação, é preciso avaliar o peso das consequências sociais da interpretação e uso do teste e da falta dele.

As distinções sobre interpretação do teste e seu uso, propostas por Messick (1987), ajudam a compreender a dimensão funcional dos aspectos da validade. O autor argumenta que a relevância ou utilidade do teste e sua adequação para um determinando uso depende do significado da nota – mais especificamente da relação construto-nota. Por exemplo, se houver problema na operacionalização do construto ao se detectar uma falha em como uma nota está sendo composta, os argumentos que fundamentam o uso do teste estarão também fragilizados. O uso do teste, bem como suas consequências sociais, são dependentes da validade de construto, ou seja, de como o construto está sendo operacionalizado no instrumento. Por isso, mais uma vez, a validade de construto é o foco da análise, mesmo quando se quer afirmar a validade do uso do teste. No quadro 3, retomo o quadro organizado por Messick (1987) sobre a relação entre evidências, consequências e interpretação e uso do teste, discutidas nesta seção.

QUADRO 3

Facetas da validade

	interpretação do teste	uso do teste
Bases evidenciais	Validade de construto	Validade de construto + relevância e utilidade do teste
Bases consequenciais	Implicações de valor	Consequências sociais

Fonte: Messick (1987), 17p.

3.3. ASPECTOS DA VALIDADE DE CONSTRUTO

Diferentemente da perspectiva tradicional em que a validade era dividida em tipos, na teoria de Messick (1987), a validade de construto é um conceito unitário, porém composto por vários aspectos. Segundo o autor, a validade de construto pode ser dividida nos seguintes aspectos: de conteúdo, substantivo, estrutural e os aspectos externos à validade de construto. No recente documento da AERA (2014), o conceito de validade é tido também como conceito unitário. Embora o documento não utilize a mesma nomenclatura sobre os aspectos da validade utilizada por Messick (1987), o documento aponta para as variadas fontes de evidência para o estudo da validade. Farei algumas ponderações sobre os encontros e desencontros entre a Teoria de Messick e o documento recente da AERA (2014) quanto aos aspectos da validade.

Para Messick (1987), os aspectos da validade de construto que tratam do conteúdo estão relacionados com o que tradicionalmente se chamava de validade de conteúdo. Cabe aqui relembrar que há limitações já discutidas anteriormente sobre as maneiras de coletar evidências que relacionem o conteúdo dos testes com o significado da nota. A análise descritiva do instrumento, que normalmente é feita baseando-se no argumento de um especialista, pode sinalizar o que o teste parece ser ou não. Já argumentei anteriormente que, para avaliar o que o teste é, é preciso verificar seus resultados e relacioná-los às avaliações descritivas e aos argumentos dos especialistas.[20]

No documento mais recente da AERA (2014), quanto à polêmica sobre ser ou não uma evidência de validade o aspecto do conteúdo do teste, assume-se que a análise de juízes ou especialistas podem fornecer evidências de validação. Segundo o Standards (AERA, 2014), os especialistas podem oferecer informações sobre o quanto o teste como um todo ou cada um de seus itens pode prever em que medida o instrumento representa o critério, ou a habilidade avaliada, ou seja, os especialistas podem julgar a representatividade dos itens. O documento, assim como Messick (1987), também aponta para a possibilidade de investigar a nota para avaliar aspectos da relação entre conteúdo e construto.

O aspecto substantivo da validade de construto é a confrontação entre os julgamentos de especialistas. Essas comparações podem fundamentar a relevância e a representatividade dos conteúdos da tarefa a partir de evidências empíricas que refutam ou endossam tais argumentos. É uma questão que envolve a verificação dos aspectos de conteúdo. O aspecto substantivo da validade de construto, assim, está

[20] No capítulo quatro apresento uma discussão sobre os argumentos de validade das entrevistas orais, problematizando o uso deliberado da validade de conteúdo.

relacionado com a tarefa e sua maneira de organizar a nota, e também com a consistência da nota. Tradicionalmente, o que faz uma tarefa entrar ou não no teste é a sua pertinência com as especificações. De acordo com Messick (1987), porém, nas abordagens empíricas de construção de teste, os itens deveriam entrar na composição do teste depois de feitas análises de dados; sejam estes dados internos ao teste, que demonstrem a homogeneidade do item ou suas cargas fatoriais, seja por meio de análises externas de dados, que envolvam o estudo da correlação do parâmetro de avaliação ou a correlação da discriminação do critério com relação a um conjunto de outros parâmetros. No documento da AERA (2014), por aspecto substancial, entende-se evidência baseada na estrutura interna do teste. Cabe ressaltar que tanto Messick quanto o Standards (2014) aponta para a análise fatorial como forma de analisar a unidimensionalidade da medida, ou seja, se os itens estão medindo o mesmo construto. Por exemplo, o quanto cada parâmetro de avaliação exemplifica o construto que está sendo medido é uma questão de aspecto substantivo da validade de construto.

Messick (1987) aponta que a análise fatorial é recomendável para avaliar essas relações por meio da análise das cargas fatoriais. As cargas fatoriais são valores que permitem avaliar o quanto cada parâmetro de avaliação está compondo a nota final de uma avaliação. O mesmo autor afirma que a análise fatorial é recomendada quando se quer combinar a avaliação de teorias e a construção de escalas para interpretar as consistências das respostas. Trato com mais detalhes da análise fatorial mais adiante neste capítulo e com mais profundidade no capítulo da metodologia. O objetivo da abordagem substantiva da validade de construto pode ser o de confrontar as informações sobre a estrutura da EPO com sua consistência na atribuição de notas, para poder avaliar se algum parâmetro de avaliação das grades apresenta propriedades empíricas insuficientes que podem distorcer a representatividade do construto medido. Caso aconteça de algum parâmetro se apresentar empiricamente pouco representativo, Messick (1987) sugere a sua substituição por outro em consonância com as especificações do teste.

O aspecto estrutural da validade de construto está relacionado com a maneira de compor a nota. O modelo de atribuição de notas deve ser coerente com a forma como as manifestações do construto se estruturam. A questão central é, pois, investigar como os parâmetros combinam entre si para produzir determinados efeitos. No documento da AERA (2014), o aspecto é abordado quanto às evidências que dizem respeito às relações entre o instrumento e o critério (*test-criterion relationships)*. A pergunta principal que as evidências deveriam demonstrar é o quanto a nota do teste prevê o desempenho real da performance avaliada, ou seja, o quanto a estrutura

interna da combinação dos parâmetros do Celpe-Bras é similar às relações com outras formas de manifestação do construto pode ser um tipo de análise. Então, avaliar o componente estrutural da validade de construto está relacionado a atribuição de notas ser coerente com outras formas da proficiência fora do teste, e o grau de relação entre estes parâmetros dentro do próprio teste. A correlação pode ser um recurso para estudar esse aspecto. É importante ressaltar que a correlação entre os parâmetros ser alta ou baixa não é algo bom ou ruim por si só. A interpretação do que é positivo ou não no contexto do teste depende do construto avaliado. Por exemplo, se a prova mede conhecimentos gramaticais de forma isolada, é provável que os itens estejam fortemente correlacionados entre si e, por isso, é provável que haja pouca discrepância de avaliação e que as notas sejam consistentes. A consistência da nota diz respeito ao fato da nota não se alterar com as condições de aplicação. O fator responsável pela alteração da nota deve ser o desempenho do examinando. Se a prova exige capacidade de usar estruturas gramaticais e lexicais para escrever um texto coeso e coerente, a tarefa é mais complexa do que a avaliação do conhecimento isolado de gramática. Por isso, deve-se aumentar a quantidade de textos e/ou de corretores nesta situação, de modo que se garanta mais consistência entre as notas. Estatisticamente, há maneiras de avaliar o quanto de tarefas e de corretores são necessários para garantir um padrão de consistência adequado. A consistência da nota depende da natureza do construto e, em sendo assim, um teste curto com alta consistência é o suficiente. Para construtos mais complexos, como os que envolvem análise de desempenho, a consistência da nota pode não ser tão alta, e por isso é preciso avaliar o modelo de atribuição de notas e sua relação com o construto para averiguar se as condições de aplicação e julgamento da nota são suficientes para assegurar a consistência dos resultados. Na especificação do comando de uma tarefa, além de ela estar fundamentada no construto, é necessário que a tarefa seja possível de ser feita e avaliada a partir de julgamentos consensuais e relevantes para o domínio, e é preciso também informar e fundamentar o modelo de atribuição de notas de forma a garantir a consistência das medidas.

No contexto do estudo do componente estrutural e substancial da validade de construto do Celpe-Bras, a representação do construto se refere à relativa dependência da resposta à tarefa na situação de prova oral na metodologia de EPO à *adequação gramatical, fluência, compreensão*, etc.. Analisar como o construto da proficiência oral está sendo representado na tarefa envolve analisar também a relação entre cada parâmetro de avaliação.

O componente externo à validade de construto se refere às relações entre o que é avaliado em um teste e em outros testes que se referem ao mesmo construto. No

documento da AERA (2014) salienta-se também a necessidade de investigar evidências para validade e consequências do teste. O componente externo à validade serve para medir, por exemplo, em que medida as notas de um mesmo examinando na prova oral em dois exames distintos que avaliam a proficiência oral estariam relacionadas. Messick (1987) afirma que antes da análise da correlação da nota da prova com outras medidas externas ao teste, e que mediriam o mesmo construto, é preciso entender a correlação entre a nota final e as notas que compõem a nota final. O modelo estrutural da composição da nota influencia a natureza e a dimensão da sua correlação externa, e também a interpretação da nota. Messick (1987) afirma que a validação de construto deve, em algum momento, confirmar o que está sendo medido em um teste. A validação de construto está relacionada com o que o examinando faz fora do teste, e por isso, em termos estatísticos, a correlação entre as medidas do teste e fora dele deve ser diferente de zero.

Alguns desses aspectos estão relacionados com a generalização da nota para além dos limites do teste. Trata-se de um aspecto valorizado pelos estatísticos, tendo sido criadas, a partir dele, várias teorias estatísticas. Generalizar uma nota de um teste significa dizer que as características dos examinandos, quanto ao seu nível de proficiência do Celpe-Bras, por exemplo, podem ser interpretadas separadamente das características dos itens ou do teste como um todo. As características dos itens de um teste ou de um teste como um todo não mudam se as características dos examinandos mudarem. No contexto da avaliação oral, as notas podem variar de acordo com o avaliador, entrevistador, tarefa e interação tarefa e avaliador. Ferramentas e modelos estatísticos que permitem isolar as fontes de erros sistemáticos do teste e a relação entre elas são as mais adequadas para este tipo de análise. Voltarei ao tema da generalização dos resultados do teste ao falar da Teoria de Resposta ao Item no capítulo da metodologia.

Messick (1987) também discorre sobre a validade de construto e considera que as análises de correlação, *multitrait-multiplemethod*, e a análise fatorial e de dificuldade de item podem fornecer evidências fortes para fundamentar as inferências que são feitas a partir da nota. O autor explica que a análise correlacional ou de covariância para validação de construto é útil para verificar a estrutura do teste ou dos construtos representados pela nota e suas relações internas. A correlação é uma medida de associação entre variáveis que pode ser usada não só para interpretar notas de diferentes testes que medem o mesmo construto, mas também para avaliar a consistência de atribuição de notas entre avaliadores. A convergência de indicadores aponta para uma evidência positiva quanto à validade de construto. Por exemplo, analisar a correlação de notas atribuídas para cada um dos parâmetros analíticos da

prova oral do Celpe-Bras é uma maneira de investigar a convergência ou divergência entre os indicadores do construto da proficiência oral nesta grade. Messick (1987) problematiza que, quando as dimensões do construto são variadas, pode ser difícil avaliar sistematicamente o padrão de correlação, porque haverá muitas variáveis intervenientes na composição da nota. Neste caso, Messick (1987) novamente sugere que a análise fatorial é uma alternativa para estudar a correlação.

Alguns especialistas em avaliação de línguas também se referem a algumas metodologias de análise da validade. Fulcher (2003), por exemplo, faz um resumo das possibilidades metodológicas para a análise do argumento de validade no contexto dos testes orais. Essas metodologias são igualmente úteis para a investigação de testes de performance em geral, e algumas delas podem inclusive servir para investigar testes de múltipla escolha ou instrumentos estruturados na forma de escala Likert, como o são as grades de avaliação do exame Celpe-bras. As escalas Likert são muito usadas em pesquisas de opinião e frequentemente estão presentes em avaliações de proficiência. A escala do Celpe-Bras segue a mesma organização de uma escala tipo Likert. O parâmetro *compreensão*, por exemplo, é descrito em uma escala de zero a cinco em que no ponto máximo consta “Compreensão do fluxo natural da fala. Rara necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala” e no mínimo consta “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” (ANEXO 1, p. 144). Entre as duas descrições há uma escala que gradua a compreensão no contexto da performance, assim como na lógica das escalas de opinião em que algo pode ser avaliado como sendo: muito bom, bom, mediano, ruim, muito ruim, por exemplo.

Fulcher (2003) ressalta que avaliar um teste é complexo e a análise pode ser feita a partir de evidências que fundamentam o argumento da validade. Assim como Messick (1987), o autor afirma que como variados aspectos da validade podem ser investigados, a qualidade de uma pesquisa está diretamente relacionada com a qualidade da evidência, que pode ser resultado de uma análise qualitativa ou quantitativa. Nesse sentido, os métodos se complementam e o valor de cada técnica está relacionada com a eficiência de responder as perguntas de pesquisa colocadas. Fulcher (2003) adverte que a validade deve ser analisada a partir de diferentes métodos, os quais trato a seguir, pois uma pesquisa que se fundamenta em um só método pode ser facilmente questionada. No contexto das pesquisas sobre avaliação, o autor afirma que os métodos quantitativos, em geral, investigam a nota e são potencialmente úteis para identificar fontes de falhas sistemáticas em um sistema de

avaliação. Tais falhas sistemáticas, detectáveis por meio de análises empíricas são o foco da discussão da validade de construto. Como já discuti anteriormente, o uso do teste está em dependência da qualidade da relação construto-nota.

Fulcher (2003) lista alguns dos métodos quantitativos mais populares entre os estudiosos das avaliação em línguas, a saber: correlação, análise fatorial, *multitrait-multiplemethod*, estudos de generalização e *multifaceted* Rasch. A lista de metodologias é semelhante à lista de Messick (1987). Estudos de generalização e os modelos Rasch incorporam o conceito da dificuldade de item, também mencionado por Messick (1987). Ao falar da correlação, Fulcher (2003) adverte que a medida é de difícil interpretação, por isso recomenda-se associá-la com outras análises. Messick (1987) sugere associá-la à análise fatorial. Assim como Messick (1987), Fulcher (2003) afirma que a análise fatorial pode servir para investigar a validade de construto, verificar hipóteses teóricas e resumir um grande volume de dados. Há um consenso entre os estatísticos sobre a importância da análise fatorial para estudos sobre validade de construto (MESSICK, 1987; BROWN, 1960/2015; KIM e MUELLER, 1978; THOMPSON, 1951). Voltarei na discussão sobre as potencialidades da análise fatorial para a investigação da validade dos testes no próximo capítulo.

Multitrait-multiplemethod também pode fornecer evidências de divergência e convergência para fundamentar a validade de construto. O método define o teste como um *trait-method unit*, ou seja, uma combinação entre construto de interesse e metodologia ou tarefa. A análise prevê a comparação de notas atribuídas em testes que têm ou não o mesmo construto e que podem ou não ter a mesma metodologia. É uma metodologia útil para avaliar a metodologia mais adequada no momento de construção de um instrumento, por exemplo.

A *Generalizability theory* ou Teoria G, que fundamenta os estudos de generalização, é um desenvolvimento da Teoria estatística clássica, que pressupõe que a nota atribuída a partir do desempenho em um determinando teste e a nota de erro[21], que pode ser algum efeito do método, por exemplo, seja igual à nota real. Isso significa que a nota realmente diz respeito à proficiência real do examinando. Na Teoria clássica, a nota de erro pode ser interpretada como aleatória ao passo que, na Teoria G, separa-se o erro aleatório do sistemático. As utilidades desta teoria são diversas, sendo possível, por exemplo, avaliar a probabilidade de um mesmo examinando tirar a mesma nota no mesmo teste, o quanto cada faceta do método contribui para a variância da nota, quantos avaliadores ou quantas tarefas são

[21] Nota de erro é um termo da teoria clássica. Trata-se de uma abstração do que seria um valor de uma nota que supostamente está contaminada por fatores aleatórios, como o mau humor do avaliador; ou sistemáticos ao instrumento, como o erro de redação do comando, por exemplo.

necessárias para que os resultados sejam significativos, etc. A Teoria G impactou positivamente a maneira de estudar a confiabilidade das avaliações e, principalmente, a maneira de avaliar a pilotagem dos testes, pois os resultados das análises poderiam ser interpretados independentemente do público do teste pilotado. Bachman (1990) apresenta as metodologias baseadas na Teoria G como uma das possibilidades para avaliação do que ele chama de confiabilidade.

Como os resultados de análise de pilotagem feitos a partir da Teoria Clássica eram dependentes do perfil dos examinandos, era necessário controlar com muito cuidado o perfil dos examinandos do teste piloto que deveriam ter as mesmas características dos examinandos para o qual o teste foi desenvolvido. As teorias mais modernas, como a Teoria de Resposta ao Item é um desdobramento dos estudos de generalização. A *multifaceted* Rasch é uma metodologia amplamente utilizada em estudos sobre avaliações na área educacional, e permite investigar a influência de mais de uma variável na nota final atribuída pelo avaliador. Os modelos Rasch pertencem a uma família de modelos estatísticos baseados na Teoria de Resposta ao Item (TRI), que surge como alternativa à Teoria clássica. Ambas têm o objetivo de analisar as qualidades psicométricas dos testes, ou seja, o quanto o sentido das notas geradas por um sistema de avaliação são válidas para um determinado propósito de avaliação, por meio da análise do significado da nota no contexto da avaliação. A TRI, assim como a Teoria Clássica, viabiliza ferramentas estatísticas para calcular a dificuldade e a discriminação dos itens de um teste. Messick (1987) explica que a dificuldade do item é matematicamente modelada a partir das respostas a um conjunto de tarefas ou subtarefas. A modelagem matemática da dificuldade do item em modelos Rasch são independente do perfil dos examinandos, tornando possível o acesso a informações sobre os itens mesmo quando o teste é pilotado em um contexto distinto da aplicação real, porque a dificuldade do item depende mais dos itens em si do que do perfil da amostra (DEMARS, 2010). Além disso, os modelos estatísticos Rasch permitem colocar diversas variáveis em uma escala única e avaliá-las ao mesmo tempo. É possível, por exemplo, avaliar a probabilidade de um examinando específico ter uma nota alta ou baixa em um determinado parâmetro a partir de sua localização na escala do construto (proficiência oral), calculada pelo modelo. Diferentemente dos estudos de generalização, os modelos matemáticos baseados na TRI permitem identificar os parâmetros de avaliação mais difíceis e também se essa dificuldade ocorre em algum ponto específico da escala de nota. Além disso, a TRI é um instrumento poderoso para análise de processos de atribuição de nota, ao identificar avaliadores que tendem a dar notas mais altas ou mais baixas. No contexto da avaliação oral do Celpe-Bras, a análise do comportamento de avaliação dos *avaliadores-observadores* e dos

avaliadores-entrevistadores pode ser estudada em detalhe para gerar informações sobre a qualidade dessas atribuições de notas que estão sendo feitas nos postos aplicadores. Uma aplicação pragmática dos resultados deste tipo de análise seria nos cursos de aperfeiçoamento, ao permitir focar o debate e a formação nos problemas recorrentes empiricamente identificados.

O uso dos métodos quantitativos em estudos de avaliação é recente e tem contribuído para compreensão da validade dos testes. Assim como nos métodos quantitativos há algum elemento qualitativo, quando se faz a interpretação dos resultados das análises estatísticas é comum fazer uso também de análise qualitativa para um aprofundamento da interpretação dos dados. Dentre os métodos qualitativos, Fulcher (2003) destaca o julgamento de especialistas, questionários e entrevistas, análise do discurso e análise de protocolo. As pesquisas que utilizam o julgamento de especialista são comuns. Uma questão problemática deste tipo de pesquisa é que nem sempre os trabalhos oferecem uma descrição detalhada dos processos usados, destaca Fulcher (2003). Questionários e entrevistas são técnicas de coleta de dados tanto da perspectiva do examinando quanto dos avaliadores e podem contribuir para compreensão da validade interna, e também da validade externa ao sistema de avaliação.

O discurso é parte do construto da proficiência oral, e por isso análises que buscam defini-lo podem ser relevantes, principalmente, para testes de línguas para fins específicos, comenta Fulcher (2003). A análise de protocolo gera dados por meio da introspecção e pressupõe que é possível verbalizar processos cognitivos. São coletadas retrospectivamente, pois se referem a algo que já aconteceu. No processo de atribuição é comum o uso deste tipo de metodologia para estudar como os avaliadores tomam decisões enquanto estão no processo de atribuição de notas. Normalmente, a técnica do protocolo é usada de forma a complementar outras análises.

As questões de pesquisa que dizem respeito ao objeto de estudo da avaliação são variadas e, felizmente, há diversas técnicas disponíveis para investigação. Ainda sobre a questão da escolha metodológica, o uso de uma técnica específica é uma questão de julgamento.

Neste trabalho, analisarei a relação construto-nota. Conforme a perspectiva de validade de Messick (1987), propus uma discussão sobre a validade de construto do teste oral do Celpe-Bras ao investigar o modelo de atribuição de notas do instrumento. Com o objetivo de avaliar a correlação entre os parâmetros de avaliação e o quanto cada um dos parâmetros da grade analítica contribuem para a composição da nota do

observador e da nota final da prova oral, utilizei a análise fatorial. Como discutido neste capítulo, embora a análise fatorial cumpra um papel importante no fornecimento de evidências relacionadas ao construto, é desejável contrastá-la com informações vindas de outras formas de análise. Por este motivo, avaliei também o quanto a grade do *avaliador-entrevistador* e cada parâmetro analítico da grade do *avaliador-observador* contribuem para a definição de cada um dos níveis de proficiência certificados pelo Celpe-Bras. Apresentarei também uma análise de discriminação dos critérios por meio do modelo Rasch, que pertence à família de modelos psicométricos fundamentados nos princípios da TRI.

Discutirei no próximo capítulo a metodologia adotada nas análises.

4. METODOLOGIA

Com o objetivo de responder às perguntas de pesquisa, a coleta e análise de dados foi organizada em duas etapas. Na primeira etapa, o objetivo geral foi o de analisar empiricamente a relação entre os parâmetros de avaliação das grades de avaliação. Inicialmente, apresentarei a correlação entre as medidas e os valores das cargas fatoriais por meio dos quais novos pesos para cada um dos parâmetros foram propostos e contrastados com os valores que hoje vigoram no modelo de composição da nota do teste. Os dados utilizados foram as notas de 1.000 examinandos que dizem respeito à prova oral, das quais faziam parte as avaliações do *avaliador-observador* e do *avaliador-interlocutor* relativas à primeira edição de 2016. As notas foram selecionadas pelos servidores do Inep e disponibilizadas para esta pesquisa. Por meio do pacote de notas, não é possível identificar de onde vieram as notas, se vieram de muitos ou poucos postos de aplicação, se os examinandos eram em sua maioria falantes de espanhol ou de inglês, etc. Cabe ressaltar que a não identificação da origem dos dados faz parte da política de acesso aos dados do INEP.

Os parâmetros da grade analítica recebem pesos diferenciados para o cálculo da nota final do *avaliador-observador*, conforme já apresentado. Por meio do cálculo fatorial exploratória, apresento uma verificação empírica dos pesos com a finalidade de identificar os parâmetros ou itens de maior relevância ou que mais determinam ou explicam a nota analítica e a nota final da prova oral. Além da análise fatorial, avaliei o grau de discriminação de cada um dos itens da grade analítica e a grade do *avaliador-*

interlocutor por meio da análise da dificuldade de item, utilizando o modelo Rasch para itens politômicos na especificação *Partial Credit Model*.

O desenho metodológico da pesquisa passou por várias revisões após a coleta piloto até chegar na análise das cargas fatoriais e da discriminação de itens. A seguir descrevo as principais contribuições metodológicas a partir do experimento piloto e, em seguida, detalho as etapas de coleta e análise.

4.1. A COLETA PILOTO

Comecei esta pesquisa com o interesse de analisar o processo de julgamento da nota do examinando. A partir de minhas experiências com avaliação das provas oral e escrita do exame Celpe-Bras, e também elaborando Elementos Provocadores a serem utilizados na prova oral entre outros instrumentos, fui me tornando cada vez mais interessada nos aspectos intervenientes na atribuição de notas. Inicialmente, achei que o comportamento dos avaliadores poderia trazer evidências sobre como o construto do exame estava sendo interpretado. Por isso, com o objetivo de investigar o julgamento do *avaliador-observador,* fiz uma coleta de dados a partir de um experimento piloto realizado no Cefet-MG, durante o mês de junho de 2016. O projeto piloto consistiu na simulação de quatro entrevistas de proficiência oral de acordo com o padrão do exame Celpe-Bras. Além dos quatro informantes estrangeiros, cinco informantes participaram como *avaliadores-observadores* em mais de duas entrevistas. Os *avaliadores-observadores* atribuíram e justificaram as notas analíticas para cada um dos seis parâmetros da grade previstos no exame, e também justificaram a nota atribuída para cada parâmetro de avaliação tendo como base os descritores da grade analítica.

Após a reflexão sobre as limitações e as dificuldades da coleta piloto, foi preciso refazer o desenho metodológico da pesquisa. A primeira limitação se deveu ao formato de coleta de dados. Além de atribuir notas, foi pedido aos avaliadores para justificarem as mesmas. Cada entrevista, atribuição e justificativa de nota durou em média 45 minutos, sendo 20 para cada entrevista propriamente dita e 20 minutos para que os informantes presentes justificassem as notas atribuídas. A simulação de entrevistas de 20 minutos de duração se mostrou inviável quando o objetivo é o de coletar uma grande quantidade de notas analíticas e suas justificativas por causa da disponibilidade dos informantes. Outra questão importante foi a dificuldade de controlar o nível de proficiência do examinando nas simulações de entrevista, o que teve como consequência a impossibilidade de controle das faixas de notas por

parâmetros em algum nível de proficiência para análise e discussão. Como o objetivo era discutir os descritores da grade analítica por meio das justificativas das notas, para a análise de protocolo seria necessário que o experimento contemplasse diversos níveis de proficiência. Encontrei examinandos disponíveis a participar do experimento apenas com proficiência alta, que obtiveram notas entre 4 a 5, segundo os avaliadores que participaram das simulações.

Ao analisar os dados, percebeu-se que o experimento piloto foi pouco eficiente no que diz respeito não só à metodologia de coleta das notas, como também com relação à coleta das justificativas. Nas justificativas de notas, os avaliadores copiaram os descritores da grade. A expectativa seria que houvesse, por parte dos avaliadores, uma reflexão sobre os descritores. Minha expectativa era que os avaliadores problematizassem em algum momento a maneira como os parâmetros estão descritos nas grades. De alguma forma, a organização dos instrumentos da coleta não foi suficiente para explicitar e discutir o processo de julgamento do desempenho oral dos examinandos.

A partir da experiência com a coleta piloto e de muitas leituras, percebi que era preciso estudar o modelo de composição da nota antes de fazer o experimento porque a qualidade do julgamento da avaliação também está relacionada como a qualidade das grades – não só no que se refere à maneira como os parâmetros estão descritos, mas à qualidade das medidas. Por isso, após o piloto, o foco do trabalho se voltou para o estudo das qualidades psicométricas das grades de avaliação do Celpe-Bras.

Como discuti no âmbito da Teoria de Validade, optamos pela análise fatorial e pela análise Rasch por se mostrarem eficientes para responder às perguntas de pesquisa deste trabalho. A seguir, são apresentados detalhes destas metodologias de análise.

4.2. ANÁLISE FATORIAL

Messick (1987) cita em variados momentos a análise fatorial como uma forma eficiente de coletar evidências para a validade, uma vez que sua finalidade é chegar num número limitado de variáveis que compõem as inter-relações entre os parâmetros de avaliação. A análise das inter-relações investiga a relação de covariação a partir de medidas observáveis, que são as notas atribuídas para cada um dos parâmetros de avaliação. Dizendo de outra forma, a análise fatorial tenta investigar a(s) causa(s) ou fator(es) relacionado(s) a um conjunto de medidas.

Brown (1960/2015) afirma que “a intenção fundamental da análise fatorial é determinar o número e a natureza das variáveis latentes, ou fatores, que explicam a variação e covariação entre um conjunto de medidas observáveis, comumente referidas como indicadores.”[22] (*op.cit.* , p.10). Os teóricos da área da estatística utilizam o termo variável latente para se referir a habilidades, proficiências, conhecimentos etc.; ou seja, ao que se quer medir a partir de um instrumento. No nosso caso, a variável latente é a proficiência oral, e os indicadores são as notas a elas relacionadas. A análise fatorial pode ter o objetivo de investigar os componentes e a natureza do construto da proficiência oral a partir da análise de resultados de instrumentos de avaliação oral, por exemplo. Por meio da análise, é possível investigar qual nota dos parâmetros de avaliação pesa mais na nota final da prova oral, porque o cálculo da covariação do conjunto de notas fornece informações sobre a contribuição de cada um dos itens para a composição da nota final.

As variáveis observáveis se referem ao conjunto de medidas teoricamente relacionadas ao construto da proficiência oral, ou seja, às dimensões da proficiência oral que, no caso do Celpe-Bras, são os parâmetros de avaliação da prova, a saber: *nota do avaliador-interlocutor, nota de compreensão, competência interacional, fluência, adequação lexical, adequação gramatical* e *pronúncia.* Como os dados utilizados foram as notas da prova oral de 1.000 examinandos, analisamos 1.000 notas de cada uma das sete variáveis, ou seja, 1.000 *notas do avaliador-interlocutor,* 1.000 *notas de compreensão,* 1.000 *notas de competência interacional,* 1.000 *notas de fluência,* 1.000 *notas de adequação lexical,* 1.000 *notas de adequação gramatical* e 1.000 *notas de pronúncia.*

Ao falar da análise fatorial, Fulcher (2003) retoma a discussão sobre a divisão fictícia entre os métodos qualitativos e quantitativos ao afirmar que a definição do fator é uma interpretação que deve ser teoricamente fundamentada. O cálculo da relação entre as variáveis e sua relação com os fatores só faz sentido quando a teoria sobre um determinado construto for levada em consideração ao coletar e ao interpretar os dados. Messick (1987) ressalta que, quando os construtos teóricos são formados a partir de uma composição de conceitos, como acontece em avaliações orais em geral, a validação de construto se torna ainda mais necessária para verificar empiricamente a pertinência da proposta de composição de tais conceitos. O autor afirma que, quando o construto é formado por uma teoria consistente, será menos penosa a verificação empírica de sua validade. Brown (1960/2015), em seu livro sobre análise fatorial confirmatória para pesquisas aplicadas, também faz a mesma ressalva:

[22] “The fundamental intent of factor analysis is to determine the number and nature of latent variables or factors that account for the variation and covariation among a set of observed measures, commonly referred to as indicators.” (BROWN, 1960/2015, p.10)

“fundamentação teórica consistente e experiência prévia com as variáveis farão com que o pesquisador interprete os fatores e avalie o modelo fatorial geral com mais facilidade.”[23] (BROWN, 1960/2015, p.21). Sinaliza-se, portanto, um desafio para a análise empírica da relação construto-nota no contexto da prova oral do Celpe-Bras.

Em termos estatísticos, Brown (1960/2015) define o fator, que é o construto operacionalizado como “uma variável não observável que influencia mais do que uma medida observável e que representa as correlações entre essas medidas observáveis. Em outras palavras, as medidas observáveis estão relacionadas entre si porque partilham uma causa comum.”[24] *(*BROWN, 1960/2015, p.10). Essa causa comum é o fator ou a proficiência oral, no nosso caso. Thompson (1951/2004), muito antes de Brown (1960/2015), já havia afirmado que as variáveis a serem estudadas por meio da análise fatorial deveriam ser interconectadas; caso contrário, a possibilidade de interpretação dos dados seria nula. No nosso caso, as variáveis são as notas atribuídas para cada parâmetro de avaliação em uma situação controlada de avaliação, por isso pressuponho que as variáveis estejam fortemente interconectadas e que, por isso, a análise fatorial contribuirá para o presente estudo.

Kim e Mueller (1978) explicam com mais detalhes o porquê das relações entre variáveis observáveis (ou as notas dos parâmetros de avaliação) e o fator (construto da proficiência oral) poderem ser estudadas pela análise fatorial. Segundo os autores,

> (…) as variáveis observáveis são combinações lineares de alguns fatores subjacentes (hipotéticos e não-observáveis). Alguns destes fatores são comuns a duas ou mais variáveis e alguns são únicos para cada variável. (...) Os fatores únicos não contribuem para a covariação entre as variáveis (...) somente fatores comuns (...) podem contribuir para a covariação entre as variáveis observáveis.[25] (KIM e MUELLER, 1978, p.8)

[23] “A firm theoretical background and previous experience with the variables will strongly foster the interpretability of factors and the evaluation of the overall factor model.” (BROWN, 1960/2015, 21p.)

[24] “(…)factor is an unobservable variable that influences more than one observed measure and that accounts for the correlations among these observed measures. In other words, the observed measures are intercorrelated because they share a common cause (…)” (BROWN, 1960/2015, p.10)

[25] “Factor analysis assumes that the observed variables are linear combinations of some underlying (hypothetical and unobservable) factors. Some of these factors are assumed to be common to two or more variables and some are assumed to be unique to each variable. (…) the unique factors do not contribute to the covariation between variables (…) only common factors (…) contribute to the covariation among the observed variables.” (KIM; MUELLER, 1978, p.8)

Trazendo esta ideia para o presente estudo, considera-se a proficiência oral como um fator subjacente às variáveis observáveis, hipotético e não-observável. A proficiência oral é o construto formado por uma composição de conceitos teóricos e abstratos tais como *competência interacional, fluência* etc., que almejamos analisar empiricamente por meio das variáveis observáveis, ou seja, por meio das notas atribuídas que são as variáveis observáveis do fator proficiência oral. No nosso caso, é provável que o estudo das variáveis observadas como as notas atribuídas pelo entrevistador e pelo observador estejam relacionados a um só fator que é o da proficiência oral, justamente por se tratar de uma avaliação que tem como propósito certificar a proficiência da oralidade. Nesse sentido, as medidas devem indicar uma relação linear entre as notas de cada um dos parâmetros. Caso a análise demonstre que a nota de algum parâmetro esteja fracamente relacionada com as outras notas e com a nota final, pode ser um sinal de que este parâmetro esteja pouco relacionado com o construto da proficiência oral que foi operacionalizada no exame.

De acordo com Thompson (1951/2004), o surgimento da análise fatorial tem como pano de fundo uma discussão iniciada no começo do século 20 sobre a natureza da inteligência. Alguns teóricos acreditavam que o indivíduo inteligente poderia ter alta performance em diversas tarefas, ao passo que outros argumentavam que nem sempre um indivíduo muito capaz de fazer determinadas atividades poderia necessariamente fazer outras com a mesma desenvoltura. Thompson (1951/2004) revisitou o trabalho de Spearman de 1904, no qual ele criou métodos para investigar o construto da inteligência por meio da análise de escores em instrumentos de avaliação, e que hoje chamamos de análise fatorial. Thompson (1951/2004) citou também o trabalho de Guilford, de 1967, que tinha o objetivo de desenvolver uma teoria relacionada ao construto da inteligência. Por meio da análise fatorial dos escores de vários testes, Guilford (1967 *apud* THOMPSON, 1951/2004) concluiu que a inteligência consiste em mais de 100 habilidades diferentes que são independentes umas das outras.

A análise fatorial pode servir tanto para investigar a validade de construto, quanto para verificar hipóteses teóricas e resumir ou agrupar um grande volume de dados. No campo da estatística, o substantivo *validade* era seguido do termo *fatorial*. Messick (1987), revisando os conceitos de validade nos documentos que estabelecem diretrizes de avaliação educacional e psicológica estadunidense, encontrou o termo validade fatorial para se referir a evidências de correlação de uma causa, ou fator comum, que estão influenciando as medidas. Thompson (1951/2004) também revisitou o texto de Nunnally de 1978 e afirma que o termo histórico para validade de construto é validade fatorial. Thompson (1951/2004) sugere que ao se desenvolver documentos

de especificações relacionados a uma medida, como grade de avaliação e graduação de descritores por níveis de proficiência, a análise fatorial deveria ser utilizada para verificar a validade da nota que, no contexto da teoria de Messick (1987), seria a validade de construto. Thompson (1951/2004) explica que, se o pesquisador tem o objetivo de responder perguntas relacionadas a o quê o teste mede, a resposta deveria ser em termos fatoriais.

Brown (1960/2015) é também um entusiasta do uso da metodologia da análise fatorial para verificar a validade de construto em pesquisas da área de ciências sociais e comportamentais. De acordo com o autor, a análise pode oferecer evidências empíricas sobre validade convergente ou discriminante em relação aos construtos teóricos. Pode também mostrar evidências empíricas de forte inter-relação entre as variáveis observáveis que são similares ou sobrepostas do ponto de vista teórico, ou de fraca inter-relação quando as variáveis observáveis fazem parte de construtos teóricos distintos. No caso da metodologia de avaliação do exame Celpe-Bras, por exemplo, a análise fatorial pode apontar quais variáveis são observáveis, ou seja, quais parâmetros de avaliação oral estão mais fortemente relacionados entre si.. Quanto mais os parâmetros estiverem relacionados entre si, mais há evidências de que a avaliação está sendo feita a partir de um mesmo construto teórico ou de uma composição de conceitos empiricamente coerente – no caso do presente trabalho, o construto seria o da proficiência oral. Messick (1987) corrobora a afirmação de Thompson (1951/2004) que, embora o conceito de validade não inclua a validade fatorial, a análise fatorial continua sendo ferramenta útil na construção de questões relacionadas à validade de construto.

Segundo Fulcher (2003), a análise fatorial é o ponto de partida de muitas pesquisas que analisam o significado das notas atribuídas para retratar algum tipo de desempenho. Brown (1960/2015) explica que a análise fatorial é a ferramenta primeira para estudos empíricos sobre o significado das medidas de instrumentos de avaliação, tais como questionários, escalas, dentre outros. Fulcher (2003) exemplificou a metodologia ao citar o trabalho de Hinofotis de 1983, no qual ele analisou 12 critérios para avaliar a comunicação de Professores Assistentes na Universidade da Califórnia com seus estudantes em situações de interação oral de sala de aula. Por meio da análise fatorial, Hinofotis (1983 *apud* FULCHER, 2003) investigou a relação entre os parâmetros de avaliação: *vocabulário, gramática, pronúncia, fluência, contato visual, aspectos não-verbais, segurança, presença, desenvolvimento de argumentação, uso de evidências ao argumentar, clareza* e *relacionamento com os estudantes*. O pesquisador partiu da hipótese de que os parâmetros poderiam ser agrupados em cinco fatores e, após a interpretação dos dados, concluiu que o fator (1) *Comunicação*

e informação é fortemente influenciado pelos parâmetros: *desenvolvimento de argumentação*, *uso de evidências ao argumentar*, *clareza e relacionamento com os estudantes*; o fator (2) *expressão*, por *fluência* e *habilidade de se relacionar com os alunos*; o fator (3) *aspectos não-verbais*, por *habilidade de se relacionar com os alunos*; o fator (4) *proficiência linguística*, por *vocabulário* e *gramática;* e o fator (5) *pronúncia* pelo parâmetro *pronúncia* apenas. Fulcher (2003) chama atenção para o fato da *pronúncia* não fazer parte empiricamente do fator que diz respeito ao construto da proficiência linguística e também para o fato do parâmetro *relação com os estudantes* estar presente em dois fatores: o da *comunicação* e *informação* e o da *expressão*. Ao final, Fulcher (2003) concluiu sobre o método da análise fatorial que, se o argumento baseado na análise for plausível, então o pesquisador teve êxito ao apresentar evidências para fundamentar uma inferência sobre o significado da nota.

Ainda no campo da avaliação de línguas estrangeiras, Kunnan (1992) também utilizou a metodologia para analisar as qualidades psicométricas de um teste de nivelamento da Universidade da Califórnia. Dentre outros métodos, a análise fatorial exploratória foi utilizada em quatro grupos de estudantes de inglês como segunda língua, para investigar a validade de instrumento, que avaliava separadamente as habilidades de leitura, compreensão oral e gramática. Ao final do estudo, Kunnan (1992) concluiu, a partir da análise fatorial, que os estudantes com baixa proficiência tendem a ter notas baixas em diferentes habilidades, uma vez que a carga fatorial deste grupo pesa para um só fator, ao passo que estudantes mais proficientes podem ter variação no domínio das habilidades de ler, compreender oralmente e usar a gramática da língua, porque a carga fatorial é distribuída. A nota total no teste era o indicador para o encaminhamento dos discentes para as disciplinas no âmbito do programa de estudos de línguas analisado pelo autor. Baseando-se nas análises, o autor sugere que a nota por habilidade, ou seja, separada por seção do teste (nota de leitura, nota de compreensão oral e nota de gramática), e não a nota total, é que deveria ser usada para nivelar os discentes e encaminhá-los às disciplinas de línguas.

Além de ser útil para avaliar as notas de um instrumento que já está elaborado, Brown (1960/2015) afirma que a análise fatorial é uma ferramenta popular para o desenvolvimento e construção de escalas de avaliação. Por meio do cálculo das cargas fatoriais referente a cada um dos parâmetros de avaliação é possível definir como a nota final da prova oral deve ser composta, por meio de um instrumento de avaliação que prevê um conjunto de itens politômicos. Itens dicotômicos são aqueles em que a resposta é sim ou não ao passo que os politômicos permitem uma gradação na resposta. A escala Likert é um exemplo clássico de itens politômicos. As grades de avaliação da prova oral são organizadas de forma que cada parâmetro de avaliação é

um item politômico com seis possibilidades de resposta, porque a escala da proficiência do exame é de seis pontos, indo do zero ao cinco.

Brown (1960/2015) também afirma que a análise fatorial pode ser usada para verificar se os parâmetros de avaliação estão ou não relacionados com a proficiência que se quer atestar, bem como os padrões de relação de cada um dos parâmetros que a compõem. Brown (1960/2015) argumenta que, quando a estrutura latente, ou seja, a estrutura da proficiência geral em uma língua, é multifatorial, isto é, formada por um ou mais fatores (que podem ser a proficiência oral, escrita, etc.), o padrão das cargas fatoriais calculadas, utilizando algum modelo de análise fatorial confirmatória, pode determinar como o teste deverá ser organizado em termos de atribuição de nota, que pode ser distribuída em subescalas. No caso do Celpe-Bras, a proficiência geral da língua é dividida em proficiência oral e escrita, com modelos de atribuição de notas distintos. Neste trabalho, conforme já assinalado, o foco é no modelo de atribuição de notas da prova oral e suas subescalas. As subescalas, explica Brown (1960/2015), correspondem ao número de fatores e a relação entre eles, ou seja, correspondem ao número mínimo de dimensões da proficiência e a maneira como diferentes tipos de proficiência se relacionam, além de indicarem a quantidade e os pesos de parâmetros presentes no modelo de atribuição da prova para composição da nota final. Dessa forma, no contexto de avaliação do Celpe-Bras, por se tratar de um instrumento de avaliação que já está elaborado e conta com um número pequeno de parâmetros, é provável que todas as variáveis estejam relacionadas a um só construto – que é o da proficiência oral –, e por isso teríamos apenas um fator para explicar as variáveis ou parâmetros. A análise fatorial confirmará a relação de todos os parâmetros a um único fator e fornecerá insumos para a discussão dos pesos destes critérios na composição da nota oral final do examinando. Cabe ressaltar que como o construto é formado por uma seleção de conceitos teóricos, a etapa de confirmação da relação nota-construto se faz necessária.

A seguir, fundamentarei a escolha da metodologia para analisar a capacidade de discriminação de cada um dos parâmetros a partir do método Rasch básico.

4.3. RASCH E TEORIA DE RESPOSTA AO ITEM (TRI)

Nesta seção, trato da metodologia da segunda etapa de análise do trabalho. Inicialmente aponto como os estudos sobre avaliação em línguas incorporaram a TRI para investigar principalmente aspectos da validade de construto. Com o desenvolvimento dos modelos estatísticos, as possibilidades de análise se ampliaram

para discussão de aspectos como a influência do avaliador no julgamento da performance em testes de desempenho, como a prova oral do Celpe-Bras. Após a discussão sobre a relação entre a TRI e o campo de avaliação em línguas, foco o debate nos aspectos fundamentais do modelo Rasch Básico na especificação *Partial Credit Model* para sustentar a análise de dados que apresentarei no capítulo seguinte.

4.3.1. TEORIA DE RESPOSTA AO ITEM E ESTUDOS SOBRE AVALIAÇÃO EM LÍNGUAS

Szabó (2007) afirma que, embora os estudos sobre o significado da nota em teste educacionais no final dos anos sessenta já começassem a utilizar a TRI nas análises, os estudos sobre exames de línguas começaram a incorporar essa metodologia apenas nos anos oitenta. Desde então, muitas foram as pesquisas que se utilizaram da TRI para o estudo de diversos aspectos relacionados aos testes de línguas. A aplicação da TRI na análise das avaliações não se restringe a testes de larga escala, mas também a testes realizados em contextos que variam de pequenas edições de exames, como testes de nivelamento e de sala de aula, a teste padronizados. Segundo os estudos resenhados por Szabó (2007), que estão organizados no quatro 4, o objetivo da análise das pesquisas que utilizam a TRI são também variados e podem: (a) focar o estudo da dificuldade dos itens para grupos específicos; (b) regular ou calibrar os itens, baseando-se nos parâmetros de discriminação; (c) avaliar o impacto das condições de aplicação na nota final; (d) verificar o quanto o teste gera de informação para que a inferência sobre o desempenho seja feita; e, por último, (e) avaliar vários aspectos da construção de testes mediados por computador. Dentre os modelos estatísticos usados, o Rasch é o mais popular entre os especialistas em avaliação de línguas. No quadro 4, resumo algumas pesquisas que utilizam a TRI para ilustrar seu uso no campo do estudo de línguas.

A presença da TRI nos estudos de avaliação em línguas se intensificou a partir dos anos 2000, ampliando-se os objetivos desses usos. Em 2008, por exemplo, os elaboradores do Quadro europeu comum de referência utilizaram o Rasch no desenvolvimento da escala de descritores das habilidades para cada um dos níveis de estudo de línguas sugeridos pelo documento (ECKES, 2015; FULCHER e DAVIDSON, 2007).

QUADRO 4

TRI em estudos de avaliações em línguas

Autores e ano	Modelo da TRI utilizado	Contexto	Objetivos e outras observações sobre a análise
Chen e Henning (1985)	Rasch	Nivelamento inglês	Detectar itens do teste que eram mais fáceis para um grupo de hispanos e de chineses
Madsen e Larson (1986)	Rasch	Três subtestes de inglês (gramática, compreensão oral e leitura)	Verificar o princípio da unidimensionalidade (ver se o item poderia ser respondido baseado nas informações da primeira língua e não do inglês, como deveria ser)
DeJong (1986)	Rasch	Compreensão oral	Calibrar um banco de itens baseado em repostas de nativos e não-nativos analisadas com base no fit e parâmetros de discriminação
Henning, Hudson e Turner (1985)	Rasch (dados multidimen-sionais)	Compreensão oral, leitura, gramática, vocabulário e escrita	Os dados foram analisados como um só conjunto e estimado a dificuldade dos itens, o princípio da unidemensionalidade não foi violado
Choi e Bachman (1992)	Vários modelos TRI	Testes de leitura da University of Cambridge Local Examination Syndicate's (UCLES), FCE e TOEFL	Verificar o pressuposto da unidimensionalidade (resultados são inconsistentes)
Hennings (1991)	Rasch	Compreensão oral do TOEFL	Condições de aplicação e impacto na qualidade do teste. Repetição do áudio reduz dificuldade do item (mas não discriminação e o valor do fit), tamanho do teste de leitura aumenta a confiabilidade e opções curtas de respostas de múltipla escolha produzem um melhor fit.
Fulcher (1997)	Modelos do TRI complemen-tando estudos clássicos de confiabilidade e validade	Nivelamento inglês da Universidade de Surrey	Avaliar a qualidade do teste em termos de confiabilidade, estimar o coeficiente de confiabilidade e o modelo Rasch para calcular a test information function
McNamara e Lumley's (1997)	Multi-faceted rasch	Occupational English Test, nivel avançado do teste australiano para profissionais da saúde	Verificar a variabilidade do interlocutor na condução da entrevista
Henning (1986)	TRI	programa de inglês como segunda língua da UCLA	Construção de banco de itens do programa de inglês como segunda língua da UCLA
Tung (1986)	TRI - CAT		Construção de banco de itens mediado por computador (cat) no contexto de avaliação de línguas
Canale (1986)	TRI - CAT	Teste de compreensão escrita	Validade de construto e efeito da metodologia do teste (sugere cautela no uso da ferramenta CAT uma vez que um teste de leitura pode não ser unidimensional)
Madsen (1991)	TRI - CAT	Compreensão oral e leitura	Compara testes mediados por computador e impressos (conclui que os mediados por computador são mais eficazes e menores)
Brown (1997)	TRI - CAT	CAT e testes de línguas	Discute problemas como pilotagem, mensuração de escores e questões logísticas relacionadas aos testes mediados por computador

Fonte: Elaborado pela autora a partir das pesquisas resenhadas por Szabó (2007)

McNamara e Knoch (2012) analisaram como o modelo Rasch foi incorporado pelos pesquisadores da Linguística Aplicada com o objetivo de desenvolver e validar instrumentos de avaliação de línguas. Os autores analisaram artigos publicados em periódicos especializados entre 1984 e 2009 que utilizaram o modelo Rasch e dividiram esse período em três momentos. O primeiro momento foi caracterizado pela incorporação do modelo Rasch na área, que ocorreu quando o professor Ben Wright se correspondeu com Georg Rasch, matemático que desenvolveu a teoria que sustenta o modelo. O segundo momento foi o da resistência tanto dos especialistas em estatística quanto dos estudiosos da avaliação em línguas adicionais. No terceiro momento, ocorreu um desenvolvimento significativo da ferramenta de análise nos trabalhos da área, na medida em que a própria metodologia foi atualizada a partir do trabalho de Mike Linacre, que incorporou multi-facetas ao Rasch e criou o software FACETS. McNamara e Knoch (2012) citaram um trabalho de Linacre e McNamara em que estudavam o efeito da variabilidade de atribuição de notas no IELTS, exame de proficiência em inglês do Governo Britânico, e concluíram que "FACETS, na verdade, revelou um problema recorrente em todos os testes de desempenho: a vulnerabilidade da nota sobre os efeitos do processo de atribuição de notas"[26] (MCNAMARA; KNOCH, 2012, p.12). De acordo com os autores, a entrada da ferramenta de análise proposta pelo modelo *Multi-facet Rasch* ecoou no crescente movimento comunicativo de ensino e aprendizagem de línguas estrangeiras da década de 90, que incorporou as tarefas de produção oral e escrita por simularem contextos de utilização da língua na vida real. Os testes também aderiram à perspectiva ao propor avaliações baseadas em desempenho, afastando-se dos itens de múltipla escolha em que as notas eram mensuradas dicotomicamente, em respostas de sim e não. Neste contexto, o modelo se mostrou uma ferramenta potencial para analisar o efeito da atribuição da nota final, principalmente no que se refere ao comportamento do avaliador. Os autores justificaram as potencialidades do modelo ao afirmar que

> as características dos avaliadores, que foram abertas para a pesquisa detalhada usando o método Rasch, incluem relativa leniência ou severidade, grau de consistência nas notas, influência do treinamento de avaliadores, influência do *background* profissional e consistência da atribuição de notas com o passar do tempo [...] É como se os pesquisadores da área tivessem em mãos um poderoso microscópio para examinar a complexidade do processo de atribuição de notas[27] (MCNAMARA; KNOCH, 2012, p.13)

[26] "In fact, FACETS had revealed a problem that was common to all performance tests, the vulnerability of the score to rater effects." (McNAMARA; KNOCH, 2012, p.12)

[27] "Rater characteristics which were now open for detailed research using Rasch methods included relative severity or leniency; degree of consistency in scoring; the influence of rater training; the influence of professional back-ground; and consistency over time. […] It is as if researchers in this field had been handed a very powerful microscope to examine the complexity of the rating process." (McNAMARA; KNOCH, 2012, p.13).

Eckes (2015) afirma que em exames de desempenho, como o do Celpe-Bras, o avaliador media a avaliação e as chances do examinando atingir uma determinada nota. O examinando dependeria não só de suas habilidades e da dificuldade da tarefa, mas também das inúmeras características do avaliador, que pode tender a fazer julgamentos mais ou menos severos. A pesquisa de Brown (2005) é um exemplo de aplicação deste tipo de análise. A autora investigou a relação entre características do avaliador ao fazer o julgamento de avaliação e seus impactos na nota. Para Eckes (2015), não só o comportamento de avaliador, mas também a escala é uma questão a se estudar, principalmente no que diz respeito à capacidade de discriminação de cada uma das faixas de proficiência respectiva a cada um dos parâmetros de avaliação. As faixas de proficiência da escala, assim como os níveis de proficiência que elas representam, podem ser definidas de forma a tornar pouco provável que os examinandos atinjam notas altas ou notas medianas, a depender de como a grade está organizada, por exemplo.

Eckes (2015) questiona até que ponto é possível supor que os avaliadores vão usar a escala da mesma forma ou seguirão um próprio estilo de avaliação. O autor discute a interação entre grade de avaliação e avaliador para sugerir que é possível investigar como os avaliadores estão utilizando a escala por meio dos modelos estatísticos do Rasch. Na especificação *Partial Credit Model* do modelo Rasch está pressuposto que os avaliadores usam a escala de forma distinta, permitindo ao pesquisador verificar com precisão como cada item e cada categoria de resposta está sendo usada pela maioria dos avaliadores. Tendo em vista as potencialidades do modelo, o presente trabalho adotará a ferramenta Rasch na especificação *Partial Credit Model* para analisar a capacidade de discriminação das grades orais do exame Celpe-Bras com base nas respostas dos avaliadores em situações reais de avaliação.

Detalharei as características do Rasch a seguir.

4.3.2. O MODELO RASCH BÁSICO

O Rasch faz parte de uma família de modelos estatísticos desenvolvidos a partir da TRI, teoria psicométrica que tem como objetivo avaliar a qualidade dos itens de testes em geral. Segundo DeMars (2010), os modelos da TRI analisam a relação entre a variável latente ou habilidade medida por um instrumento e a resposta ao item ou à tarefa.

O construto relacionado à variável latente ou habilidade pode ser o da proficiência oral, o da proficiência acadêmica, a aptidão ou até mesmo uma determinada crença. Segundo Eckes (2015), no contexto da estatística aplicada aos testes, a habilidade ou *trait* é um termo amplo e genérico que se refere à variável latente que representa o construto de interesse. A variável é latente porque não é diretamente observável, mas se manifesta nas respostas, que são observáveis.

As tarefas ou itens podem ser dicotômicos ou politômicos. Conforme já dito anteriormente, os itens dicotômicos são aqueles em que duas categorias de resposta são possíveis. Itens dicotômicos geram dados dicotômicos. Os itens politômicos permitem mais de duas categorias de resposta e são resultado do uso de escalas e grades de avaliação. As grades de avaliação graduam, granulam ou dividem a resposta ao item ou ao parâmetro de avaliação em mais possibilidades. Segundo Eckes (2015), há uma relação direta com a escala de Likert, que segue a seguinte lógica de possibilidades de resposta: bom, mais ou menos bom, mais ou menos ruim, ruim. A TRI permite avaliar as qualidades psicométricas tanto de itens dicotômicos quanto de itens politômicos e há modelos e especificações estatísticas diferentes para cada tipo de dado.

De forma geral, a TRI estuda a relação empírica entre habilidade do examinando e as respostas aos itens por meio de um conjunto de notas de um teste. É possível determinar empiricamente o perfil da habilidade dos examinandos e a dificuldade de cada um dos itens que compõem o teste. Segundo Szabó (2007), como só as respostas dos examinandos constituem a informação disponível para o cálculo, é a partir delas que é estimada a dificuldade do item em função da habilidade do examinando. Cabe ressaltar que a métrica que compõe as faixas de proficiência de um determinando exame, que representam um conjunto de conceitos teóricos sobre a natureza do conhecimento sobre a língua, não são diretamente equivalentes ao que se chama de habilidade ou *latent trait* ou *latent dimension.* Por exemplo, no contexto da análise de dados desta pesquisa, um examinando com habilidade 5.0 localizado na métrica da dimensão latente não significa necessariamente que sua nota final tenha sido 5 na prova oral do Celpe-Bras e que, por isso, seria classificado na faixa *avançado-superior* na prova oral do Celpe-Bras. A habilidade do examinando no contexto da análise da TRI não é produto da teorização sobre as medidas, mas de uma análise empírica em que a habilidade do examinando é estimada pelo modelo a partir do conjunto de notas total em relação a diversos perfis de examinandos. Tais perfis de examinandos são agrupados a partir do modelo e, ao organizar os examinandos em graus de habilidade, o modelo os coloca em uma métrica própria chamada de *latent trait.* A partir da análise empírica da habilidade dos examinandos e

da dificuldade dos itens é possível relacionar o quanto cada item discrimina determinados perfis de examinandos.

De acordo com DeMars (2010), a escala de habilidade é construída em intervalos, ao passo que a escala de número de resposta correta está relacionada com números ordinais. Os modelos Rasch organizam esses intervalos na métrica *log-odds* ou *logits.* Em um teste de 20 itens dicotômicos, por exemplo, vamos supor que haja um conjunto de notas em que alguns examinandos acertem um item, outros que acertem todos e a grande maioria acerte entre 7 e 16 itens. Se compararmos a escala de números de acerto e a escala de intervalo estimada pelo Rasch, teríamos a seguinte comparação no figura abaixo:

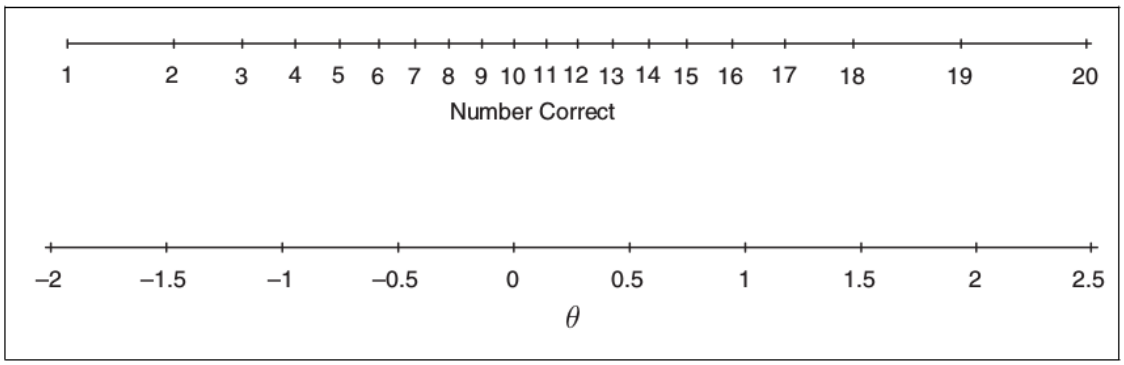


FIGURA 2 - Relação entre as métricas habilidade (θ) e escores crus (number correct)
Fonte: DeMARS, 2010, 17p.

Na primeira escala, em *number correct,* temos a distância entre o número de acertos dos examinandos em um teste de um a vinte itens. Na escala abaixo, representada pela letra grega θ, temos a mesma informação, porém na escala de intervalo e na métrica *logits.* Na primeira escala, o intervalo entre os números de respostas corretas não é espaçado em partes iguais, porque o examinando que acertou 10, em termos de habilidade está muito próximo do que acertou 11 itens. Por outro lado, o examinando que acertou 18 está mais distante em termos de habilidade do que o examinando que acertou 19 itens. Dizendo de outra forma, é provável que examinandos com habilidades próximas tenham acertado de 10 a 11 itens e examinandos de habilidades não tão próximas acertaram de 18 a 19 itens. Na parte de baixo da figura, a mesma informação é organizada em uma métrica em que os valores da habilidade do examinando foram organizados em intervalos de meio ponto.

O cálculo da habilidade pode ser feito conjuntamente ou separadamente do cálculo da dificuldade de item. Há procedimentos estatísticos que calculam primeiro a habilidade e há outros que calculam tanto a habilidade quanto a dificuldade em conjuntos. De forma geral, o modelo calcula a probabilidade de acerto de cada um dos

itens por examinandos de diferentes perfis de habilidade. DeMars (2010) explica que a dificuldade de um item é medida a partir das respostas corretas e não em termos de quantidade de esforço ou percepção da dificuldade.

Retomarei o conceito de dificuldade do item com mais detalhes ao longo do texto. A seguir, discutirei a relação entre Teoria Clássica e a TRI.

4.3.3. TEORIA CLÁSSICA E TRI

A TRI é um desenvolvimento da Teoria Clássica da Medida, que também é utilizada para analisar itens de testes. A TRI deve ser entendida não em oposição, mas como um desdobramento da Teoria Clássica (DEMARS, 2010).

Os modelos estatísticos da TRI, quando comparados à Teoria Clássica, analisam de forma mais detalhada as qualidades psicométricas dos testes (HAMBLETON *et al.* 1991; ECKES, 2015; DEMARS, 2010). De acordo com Hambleton (1991), a Teoria Clássica é limitada porque as habilidades dos examinandos e as características do teste não podem ser separadas, ou seja, as duas só podem ser interpretadas conjuntamente. A habilidade na Teoria Clássica é sinônimo de *true score,* que seria uma abstração do desempenho esperado real ou verdadeiro do examinando em um sistema de avaliação livre de impactos e interferência nesta nota. Na Teoria Clássica, quando o exame é difícil, a habilidade do examinando será interpretada como sendo baixa; se for fácil, será interpretada como sendo alta. Na TRI, os parâmetros estatísticos de dificuldade do item ou do teste são independentes da população de examinandos e essa característica é chamada de invariância (HAMBLETON *et al.*, 1991; DEMARS, 2010).

O segundo problema da Teoria Clássica é que a confiabilidade está associada à correlação de medidas de testes paralelos, ou seja, testes que medem a mesma coisa, mas de forma diferente. Segundo DeMars (2010), na Teoria Clássica, a nota verdadeira é a média hipotética da nota atribuída, que seria obtida se a prova fosse aplicada várias vezes nas mesmas condições. Nesta perspectiva tradicional, o cálculo da confiabilidade se dá a partir da replicação dos testes, ou seja, de notas dadas ao mesmo examinando a partir da aplicação do mesmo teste em duas situações diferentes ou do desmembramento de um mesmo teste, quando o teste possui perguntas duplicadas que medem o mesmo construto. Hambleton *et al.* (1991) problematizam o conceito de paralelismo para o cálculo da confiabilidade ao questionar as condições de testes paralelos. Segundo o autor, para que dois testes

sejam paralelos é preciso que eles respeitem condições que são muito difíceis senão impossíveis de serem satisfeitas. No caso de exames de desempenho como o Celpe-Bras este tipo de condição seria impossível de ser satisfeita porque teríamos que controlar o comportamento dos entrevistadores, dos avaliadores e o perfil exato da proficiência dos examinandos, dentre outras variáveis.

O terceiro problema da Teoria Clássica, ainda de acordo com Hambleton *et al.* (1991), é que o foco está no teste e não no comportamento individual de cada item, o que permitiria entender como o conjunto de itens poderia contribuir para a qualidade do teste. Além disso, na Teoria Clássica não se analisa como os examinandos respondem a um determinado item e isso faz com que os elaboradores não consigam prever como um grupo de examinandos vai responder a ele. Para o autor, saber a probabilidade que um perfil de examinandos tem de acertar um item fornece subsídios relevantes para os elaboradores, administradores e construtores de políticas a serem conduzidas a partir do resultado do exame. Por fim, Hambleton *et al.* (1991) expõem as limitações da Teoria Clássica dizendo que os modelos não são replicáveis, uma vez que dependem dos examinandos.

Na TRI, a análise é replicável, pois os parâmetros dos itens são independentes da população que se submete ao teste. A TRI leva em conta tanto a dificuldade do item quanto o seu grau de discriminação e essas duas características do item são usadas para o cálculo da confiabilidade.

Segundo Hambleton *et al.* (1991), a TRI está fundamentada em dois pressupostos básicos: o da unidimensionalidade e o da dependência local. O primeiro pressuposto implica que a performance do examinando pode ser prevista a partir de um conjunto de fatores ou de um só fator chamado de variável latente, que corresponde à proficiência oral, no caso do Celpe-Bras. O segundo pressuposto, da dependência local, implica em uma relação entre a quantidade de erros e acertos nos itens do teste e a capacidade ou proficiência do examinando, que pode ser descrito a partir de uma função monotônica crescente chamada de *função da característica de item* ou *curva da característica do item*. Segundo Hambleton *et al.* (1991) a unidimensionalidade, de forma geral, permite afirmar que apenas uma habilidade está sendo avaliada, mas na prática podem haver outras variáveis interferindo, como personalidade, ansiedade, etc. O princípio da unidimensionalidade pressupõe que um fator está predominando na avaliação da performance. O pressuposto da independência local diz respeito à independência entre habilidade avaliada e item. O pressuposto está sendo violado quando algum item do teste avalia uma habilidade diferente da que se pretende medir em algum item específico ou no teste como um todo. Por exemplo, o pressuposto

estaria sendo violado em um teste de matemática em que examinandos tiram notas baixas porque o enunciado está em uma língua estrangeira, independentemente do conhecimento do conteúdo específico do testes.

Segundo DeMars (2010), de forma geral, o cálculo da habilidade é feito da seguinte forma: primeiro estima-se a proficiência ou como os níveis de proficiência estão distribuídos no corpus, depois a proficiência individual de cada examinando é calculada a partir das notas atribuídas, ou seja, analisa-se o significado daquela nota em termos de proficiência a partir de um conjunto de notas. A análise gera a função da informação relacionada com a construção de intervalos confiáveis no cálculo das faixas da proficiência. DeMars (2010) afirma que a função de informação do teste está atrelada à proficiência e aos itens e, por isso, pode variar ao longo das faixas que caracterizam os níveis de proficiência. Quanto mais informação se tem do teste, menor o desvio padrão e maior o grau de confiabilidade. Como é possível medir o comportamento de cada item isoladamente, é possível avaliar também a função de informação de cada item isoladamente e analisar o quanto cada item contribui para o teste como um todo.

Segundo Hambleton *et al.* (1991), há três famílias de modelos estatísticos da TRI, que variam com relação à quantidade e tipo de característica que se acredita estar afetando o desempenho: de um, dois ou três parâmetros. Trato neste trabalho dos aspectos relacionados ao modelo escolhido na análise: o Rasch, que está relacionado ao modelo de um parâmetro da TRI. DeMars (2010) explica que o modelo Rasch é matematicamente equivalente aos modelos de um parâmetro da TRI, porém como foi desenvolvido à parte, a maioria dos usuários deste modelo preferem não denominá-lo como TRI, mas apenas como Rasch.

Para o cálculo estatístico tanto da habilidade quanto da dificuldade do item, é suficiente saber apenas quantas vezes o examinando acertou no teste como um todo, não importando qual item acertou, nem seu perfil. Segundo DeMars (2010), a medida *dificuldade do item* identifica o perfil de habilidade na métrica *logit,* organizado na escala de intervalo em que é esperado que 50% dos examinandos respondam um item corretamente. A medida é representada em um gráfico em que, no eixo horizontal encontra-se a métrica da habilidade ou variável latente, e no eixo vertical está a probabilidade de acerto do item. Tomemos por exemplo, uma análise que apresenta uma métrica da habilidade entre -2.5 e 2.5 pontos na escala *logit.* Neste contexto de análise, o perfil de examinandos encontra-se localizado no ponto 1 da métrica de habilidade em que apresentam 50% de chance de acerto no item A e outro perfil de examinandos encontra-se localizado no ponto 2.5 com 50% de chance de acerto no

item B. Neste caso, pode-se afirmar que o item A é mais fácil que o item B, porque examinandos com habilidades baixas, localizados no ponto 1.0 da métrica *logit,* apresentam chances de acerto no item A, ao passo que examinandos localizados no ponto 2.5 têm chances de acerto do item B. Dizendo de outra forma, o item A discrimina examinandos de baixa habilidade e o item B, examinandos com perfil mais alto de habilidade.

Segundo DeMars (2010), a medida de dificuldade é uma medida ß , que corresponde a um ponto na métrica em que, por exemplo, 50% dos examinandos com uma proficiência ß =0.2 vão acertar o item, e uma porcentagem bem maior do que 50% corresponderia as chances de acerto dos examinandos com um perfil um pouco maior de habilidade. Outra medida que pode ser avaliada é a capacidade de discriminação do item. Segundo a autora, uma alta discriminação significa dizer que um item ou um teste discrimina entre diferentes perfis de examinandos.

No contexto da análise de itens politômicos em modelos Rasch, como as respostas de escalas e grades de avaliação, a *dificuldade do item* é o *threshold, limite* ou *nível limiar.* Eckes (2015) afirma que o parâmetro *threshold* "representa o ponto de transição em que o examinando tem a probabilidade de 50% de chance de responder ou estar em uma de duas categorias adjacentes, pressupondo que o examinando está em uma dessas duas categorias."[28] (ECKES, 2015, 27p.). Ou seja, a medida relacionada ao nível limiar representa um ponto de transição em que o examinando tem 50% de chance de estar em uma categoria de resposta ou em outra. No caso das escalas do Celpe-Bras, as categorias adjacentes se referem às faixas de proficiência, e o *threshold*, ao momento em que, nos gráficos da curva de informação de cada faixa de proficiência as curvas de duas faixas de proficiência da escala se cruzam.

Por exemplo, vamos supor novamente que, após uma análise, temos uma métrica da habilidade entre -2.5 a 2.5 pontos na escala *logit* na nota correspondente ao parâmetro de avaliação *compreensão* e *fluência*, numa prova oral como a do Celpe-Bras. O perfil de examinandos localizados no ponto 1 da métrica da habilidade indica que eles têm 50% de chance de estar na faixa *avançado* ou *avançado superior*. Em resposta ao parâmetro *fluência* os examinandos localizados no mesmo ponto 1 têm 50% de chance de estar na faixa *intermediário* e *intermediário superior*. Neste caso, poderíamos concluir que o parâmetro *compreensão* é mais fácil que o parâmetro *fluência* porque, com relação ao perfil de examinandos localizados no ponto 1, é provável que em *compreensão* os examinandos sejam classificados em faixas altas de

[28] "In other words, τ k (threshold parameter or category coefficient) represents the transition point at which the probability is 50% of an examinee responding in one of two adjacent categories, given that the examinee is in one of those two categories." (ECKES, 2015, p.27)

proficiência ao passo que, quanto à nota de *fluência*, o mesmo perfil de examinandos representa classificação em faixas mais baixas da proficiência. Dizendo de outra forma, é mais provável os examinandos tirarem nota mais alta em *compreensão* do que em *fluência*.

Segundo DeMars (2010) é desejável que os itens diferenciem variados níveis de proficiência, pois quanto mais um item discrimina, mais confiável é o teste. Para que um teste meça realmente o que tem que medir, a dificuldade dos itens devem estar relacionadas com o propósito do teste. No caso do Celpe-Bras, que se propõe a diferenciar e certificar várias faixas de proficiência, esperam-se que estejam contemplados nos itens a diferenciação tanto do nível *avançado* para o do *avançado superior* quanto do *sem certificação* para o *intermediário.*

Além dos modelos da Rasch serem replicáveis e independentes da amostra de examinandos, seus modelos permitem estudar não só a qualidade psicométrica do teste como um todo, mas também a qualidade de cada item individualmente e o quanto cada item está contribuindo para discriminar os examinandos em diferentes faixas de proficiência. Para a análise de testes politômicos, como é o caso do Celpe-Bras, o uso dos modelos Rasch permite avaliar a relação da dificuldade e discriminação em função do nível de proficiência relacionadas a cada uma das seis categorias de resposta ou nota da escala de avaliação (*sem certificação, básico, intermediário, intermediário superior, avançado, avançado superior)* para cada parâmetro de avaliação (*nota do entrevistador, compreensão, competência interacional, fluência, adequação gramatical, adequação lexical e pronúncia*).

Na perspectiva das teorias psicométrica, Eckes (2015) afirma que o estabelecimento do intervalo entre as faixas de proficiência é uma questão que deve ser empiricamente investigada. O estabelecimento do intervalo entre as possibilidades de resposta, ou interpretação do desempenho, e parâmetros de avaliação do Celpe-Bras, por exemplo, deve ser empiricamente analisado a partir de um conjunto de notas atribuídas em situação de avaliação. A pergunta que se coloca é até que ponto um desempenho avançado em *compreensão* corresponde a um desempenho avançado em *adequação lexical,* por exemplo. Na perspectiva do Eckes (2015), o nível avançado em todos os parâmetros de avaliação deveriam ser similares ou correspondentes. A expectativa deve ser de que a escala como um todo e os parâmetros como um todo consigam discriminar tanto examinandos de baixa quanto de alta proficiência, ou seja, o intervalo dos valores de *threshold* entre as faixas de certificação deveriam ser homogêneo. Verificar a homogeneidade da escala é uma tarefa empírica possível de ser feita a partir da análise da relação da medida *dificuldade do item* em função dos

perfis de habilidade. Eckes (2015) afirma que organizar a estrutura das categorias, faixas ou nível de resposta de uma escala deve ser uma hipótese elaborada a partir da análise da estrutura de um conjunto de observações, do qual a análise da medida de *dificuldade do item* em função dos perfis de habilidade faz parte. Como a grade do Celpe-Bras é composta por sete itens politômicos, é possível estudar a dificuldade de um examinando ser classificado em cada uma das faixas de certificação de cada um dos itens utilizando o modelo Rasch na especificação *Partial Credit Model.* Cabe ressaltar que em modelos Rasch Básico, o parâmetro de discriminação (*a-parameter*) é fixado em 1 (DEMARS, 2010), no entanto, segundo Eckes (2015), a especificação *Partial Credit Model* permite o cálculo da dificuldade de transição de uma faixa de proficiência para outra e, ao fazê-lo, pressupõe-se ser possível estimar diferentes valores de *threshold* ou de dificuldade para itens distintos.

Ao considerar as potencialidades de análise do conjunto de respostas da prova oral do Celpe-Bras com a utilização do modelo Rasch na especificação *Partial Credit Model*, justifica-se sua pertinência para o debate sobre a relação entre o construto e a nota, no que se refere à organização dos intervalos de faixas de certificação previsto no exame e na escala.

Neste capítulo, introduzi como a coleta piloto desencadeou uma série de questionamentos que culminaram na necessidade de estudar empiricamente o significado das medidas atribuídas à prova oral e, em seguida, fundamentei a escolha metodológica para condução do estudo. Apresentarei, a seguir, a análise dos dados.

5. ANÁLISE E DISCUSSÃO

Antes de tratar da análise fatorial e da análise da TRI, descrevo o corpus de forma a apresentar o conjunto de dados que correspondem às notas da parte oral do exame Celpe-Bras. Em seguida, será apresentada o resultado da análise de correlação entre os parâmetros de avaliação das grades de proficiência oral e, ao final, serão apresentados os resultados da análise fatorial e do Rasch.

Os dados correspondem à avaliação de desempenho oral de 1.000 examinandos que se submeteram à avaliação na edição do primeiro semestre de 2016. O conjunto de dados analisados apresenta 9 variáveis: seis notas referentes aos seis parâmetros avaliados na grade do observador; uma nota total denominada *nota do observador*; uma nota total denominada *nota do entrevistador* e a nota final do examinando denominada *nota prova oral*. Na análise foram utilizados apenas os dados referente às notas dos 6 itens que compõem a grade analítica e a nota do entrevistador. A nota final do observador e a nota final da prova não foram consideradas nos cálculos que apresento à seguir. Ou seja, apenas 7 itens foram levados em conta na análise.

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

permite diversos cálculos estatísticos. Para a análise fatorial, foi utilizado o pacote Psych versão 1.8.4 (REVELLE, 2018) .

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

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

	Nota do observador	Nota do entrevistador	Nota final prova oral
0%	0.250	0	0.4000
2.5%	1.709	2	1.8373
5%	2.170	2	2.0900
25%	3.250	3	3.1500
50%	3.855	4	3.9300
75%	4.500	4	4.2900
95%	5.000	5	5.0000
97.5%	5.000	5	5.0000
100%	5.000	5	5.0000

Nas tabela 2 e 3 está descrita a composição do corpus a partir das categorias analíticas. Ao analisar como estão distribuídas as notas da parte analítica na amostra do estudo, quando comparadas ao conjunto de parâmetros de avaliação, observamos a predominância de notas elevadas em alguns parâmetros. *Compreensão* e *competência interacional* apresentam notas elevadas quando comparados com *pronúncia*, *adequação gramatical* e *adequação lexical*. Na tabela 3, temos que quando até 50% de examinandos que tiraram nota igual ou menor a 3 para os parâmetros *adequação gramatical* e *adequação lexical,* temos até 50% de notas 5 em *compreensão* e até 50% de notas igual ou menor a 4 em *pronúncia, fluência e competência interacional*. Por exemplo, o avaliador pode organizar sua avaliação de forma a atribuir para um mesmo examinando nota 5 em *compreensão*, 4 em *pronúncia* e em *fluência* e nota 3 em *competência interacional,* em *adequação gramatical* e em

adequação lexical. É possível afirmar que há um pouco mais de notas elevadas atribuídas aos parâmetros *compreensão, competência interacional e fluência* do que aos parâmetros *adequação lexical, adequação gramatical* e *pronúncia* na composição da amostra. A predominância de notas elevadas é mais evidente para a nota de *compreensão*, em comparação com os outros parâmetros, uma vez que 70% das notas de *compreensão* do corpus se concentram na nota 5 (TABELA 3). A partir dos dados apresentados na tabela 2, com relação ao parâmetro *compreensão*, até 50% das notas do corpus se concentram nas notas 5 ou menores, ou seja, as notas desta categoria na amostra são as mais elevadas quando comparadas aos outros parâmetros. Ao longo das análises, discutirei, dentre outras coisas, a relação da nota de compreensão com os demais parâmetros.

Tabela 2 - Distribuição normal padrão acumulada das notas analisadas da parte analítica

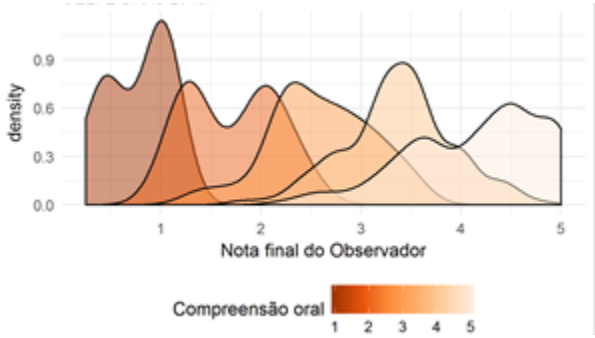
	compreensão	competência interacional	fluência	adequação lexical	adequação gramatical	pronúncia
0%	1	0	0	0	0	0
2.5%	2	2	1	1	1	1
5%	3	2	2	2	2	2
25%	4	3	3	3	3	3
50%	5	4	4	3	3	4
75%	5	5	5	4	4	4
95%	5	5	5	5	5	5
97.5%	5	5	5	5	5	5
100%	5	5	5	5	5	5

Tabela 3 - Proporção de cada nota por parâmetro analítico

	0	1	2	3	4	5
compreensão	0.000	0.010	0.023	0.077	0.190	0.700
competência interacional	0.004	0.018	0.066	0.195	0.295	0.422
fluência	0.001	0.030	0.084	0.236	0.283	0.366
adequação lexical	0.004	0.037	0.163	0.303	0.320	0.173
adequação gramatical	0.007	0.040	0.168	0.310	0.303	0.172
pronúncia	0.003	0.027	0.125	0.258	0.379	0.208
entrevistador	0.001	0.021	0.082	0.293	0.398	0.205

Ao analisar a distribuição das notas por parâmetro de avaliação em relação à nota analítica organizada na tabela 3, temos algumas considerações a fazer. Com relação às notas de *compreensão*, cabe destacar a predominância na distribuição das notas 4 e 5, quando a nota analítica geral é a partir de 3. Isso quer dizer, por exemplo, que é provável que um examinando que obtenha 3 na nota do observador possa ter tirado uma nota 4 ou 5 em *compreensão*. No gráfico 1, é possível visualizar esses valores normalmente mais altos da nota de compreensão em relação à nota do observador. Abaixo apresentamos o gráfico que se refere à densidade das notas atribuídas pelo observador para cada parâmetro e a sua relação com a nota final analítica.

Gráfico 1 - Densidade das notas de compreensão atribuídas pelo *avaliador-observador*



O gráfico acima representa a relação entre a *nota compreensão* e a *nota final do observador*, ou seja, a nota analítica. No gráfico de densidade, cada uma das montanhas representam a distribuição de uma nota de compreensão e como ela está distribuída no eixo y em função da nota do observador. Por exemplo, quando temos uma nota 1 em *compreensão* é provável que o examinando tenha tirado uma nota analítica 1 ou 2. A nota 5, por sua vez, ocupa uma faixa maior do eixo y e isso quer dizer que quando temos uma nota 5 em *compreensão*, pode ser que o examinando tire uma nota analítica 3 ou 4.

As representações das notas por parâmetros em função da nota geral do observador nos gráficos de curva de nível e densidade complementam a análise das tabelas 2 e 3 analisadas acima. Os demais gráficos de curva de nível e densidade dos outros parâmetros se encontram em anexo e nada dizem de novo sobre o que foi anteriormente explorado nas tabelas. Trata-se apenas de uma outra forma de representar as mesmas informações das tabelas acima. Sobre os gráficos em anexo (ANEXO 3, p.148 em diante), pode-se dizer que as representações das notas por parâmetros em função da nota geral do observador nos gráficos de curva de nível e

densidade complementam a análise das tabelas de distribuição normal padrão acumulada. De maneira geral, quando comparamos os gráficos de densidade dos parâmetros de *fluência* (ANEXO 3, p.150), *competência interacional* (ANEXO 3, p. 147) *e pronúncia* (ANEXO 3, p.153) temos aspectos semelhantes quando a nota final do observador que, em geral, fica em torno do mesmo valor da nota do observador. As notas de *adequação gramatical* e *adequação lexical* tendem a ser mais baixas do que a nota analítica a partir de 3 (ANEXO 3, p.151-152). Poderíamos inferir, a partir da distribuição e concentração das notas analíticas, que é mais difícil um examinando tirar notas altas nos parâmetros *adequação gramatical* e *adequação lexical* quando comparado aos outros parâmetros analíticos. Os demais gráficos de curva de nível e densidade dos outros parâmetros se encontram em anexo e nada dizem de novo sobre o que foi anteriormente explorado nas tabelas 2 e 3. São apenas uma forma gráfica de representar as informações das tabelas acima, cujas informações foram tratadas.

5.1. MATRIZ DE CORRELAÇÃO

Retomando as discussões sobre as maneiras de verificar empiricamente a validade, alguns autores tais como Messick (1987), Fulcher (2003) e McNamara (2000) apontam a correlação como um das possibilidades. No contexto da validação do construto, Messick (1987) afirma que é preciso entender a correlação entre a nota final e as notas que a compõem, porque o modelo estrutural da composição da nota influencia a natureza e a dimensão da sua correlação interna e também da interpretação da nota. Estudar a correlação entre os parâmetros de avaliação significa analisar em que medida as notas atribuídas para cada um dos parâmetros analíticos e a nota do avaliador-entrevistador estão relacionadas. Se as notas estiverem correlacionadas, significa dizer que, quando o examinando tira uma nota alta em um determinando parâmetro, a nota dos outros parâmetros também serão altas, e quando uma nota é baixa, as outras notas também serão baixas. Trata-se de uma correlação linear entre as variáveis nota do avaliador-entrevistador, nota de compreensão, nota de fluência etc.

Apesar do algortmo computacional utilizado no pacote Psych (REVELLE, 2018) oferecer diversos métodos de cálculo das correlações, optou-se por calcular as medidas de correlação entre os itens, que apresento na matriz abaixo, pelo método de Pearson. Na tabela a seguir, estão organizados os valores que correspondem à correlação e à dispersão entre as notas de cada um dos seis parâmetros da grade

analítica, da nota final do avaliador-observador, da nota do avaliador-entrevistador e da nota final da prova oral. Cabe ressaltar que a composição da nota do avaliador-observador, bem como a nota final da prova oral, foi feita a partir dos pesos que atualmente vigoram: os três primeiros parâmetros correspondem à 50% da nota total da prova oral. Os 100% da nota do avaliador-observador estão organizados de forma que 16.6% correspondem às notas de *compreensão, competência interacional, fluência*; *adequação lexical* e *adequação gramatical* correspondem à 42%, sendo 21% para cada parâmetro; e 8% corresponde à *pronúncia*. A nota do avaliador-entrevistador compõem 50% da nota e a nota do avaliador-observador os outros 50%.

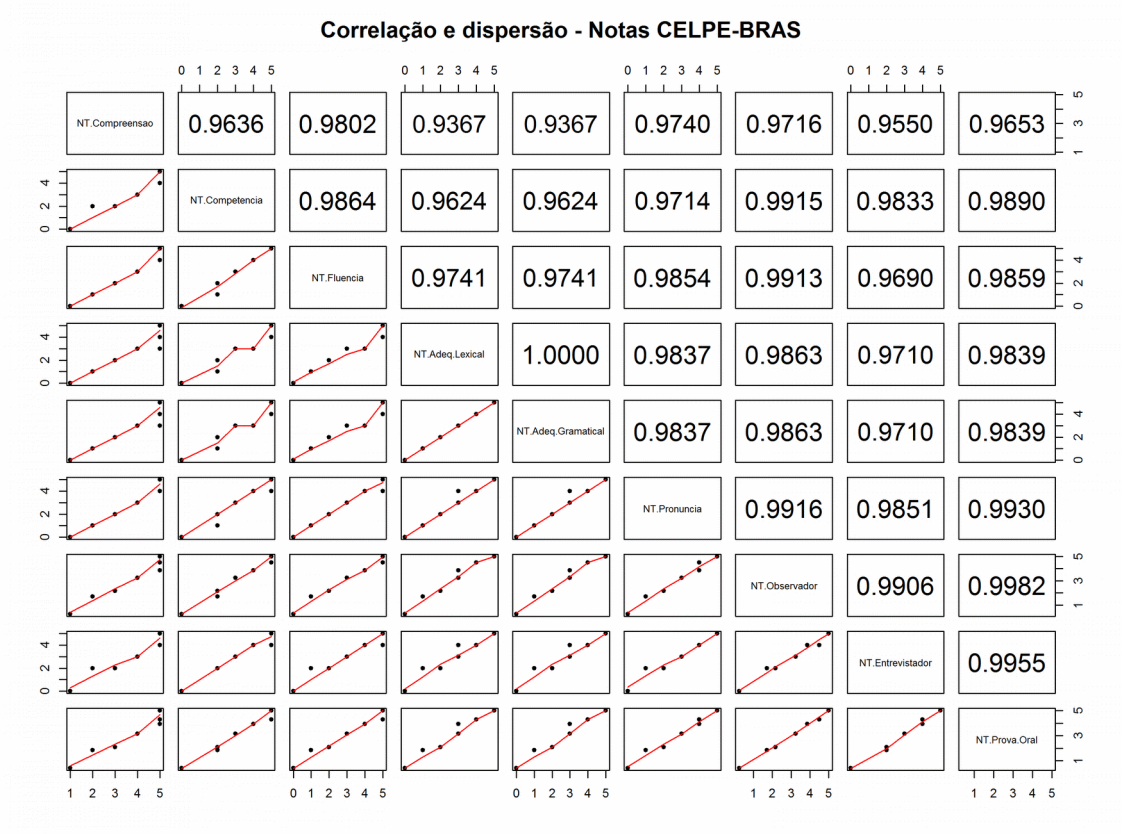


FIGURA 3 - Matriz de correlação entre variáveis

Na figura acima, está organizada a matriz de correlação das medidas por cada uma das variáveis. Na diagonal, atravessando o meio da matriz, estão denominadas as variáveis. Por exemplo, no primeiro quadrado à esquerda temos a *NT. Compreensão,* ou seja, nota de compreensão, e todos os números dispostos à direita são valores que correspondem à correlação entre a variável *NT. Compreensão* e as demais variáveis. Por exemplo, o valor 0,9636 se refere à correlação entre *NT. Compreensão* e *NT. Competência Interacional;* 0,9802 à *NT. Compreensão* e *NT.*

Fluência, e assim por diante. Os gráficos à esquerda se referem à representação gráfica da mesma informação expressa pelos valores de correlação à direita. Por exemplo, o valor 0,9636, que se refere à correlação entre *NT. Compreensão e NT. Competência,* está representado graficamente no primeiro quadrado da segunda linha.

Por meio da análise dos valores de correlação podemos avaliar a plausibilidade da validade de construto das medidas. Quanto maior o valor da correlação, mais provável que as medidas estejam relacionadas e, por isso, sugere-se que possam estar medindo o mesmo construto. Em uma escala de 0 a 1, os valores de correlação da matriz na tabela acima são maiores que 0,9, por isso é possível afirmar que as notas estão fortemente correlacionadas entre si. Na análise de correlação das notas dos parâmetros de avaliação previstos nas grades da prova oral do Celpe-Bras, verifica-se, por meio dos valores, que é provável que as notas estejam organizadas de forma a refletir aspectos de um mesmo construto - a proficiência oral.

Ao analisar a correlação entre os parâmetros da adequação lexical e da adequação gramatical, encontramos uma forte correlação, porque o valor 1,0 correspondente à relação entre essas duas variáveis. Uma correlação forte entre variáveis sugere que tais aspectos possam estar medindo a mesma dimensão do construto. O valor de correlação talvez possa indicar que a nota da adequação lexical e a nota da adequação gramatical possam estar sobrepostas. Os resultados das análises fatoriais da TRI fornecerão mais evidencias sobre o significado da relação entre estas duas variáveis.

A seguir, apresentarei o resultado da análise fatorial.

5.2. RESULTADOS DA ANÁLISE FATORIAL

Conforme já discutido em capítulos anteriores, a análise fatorial é uma ferramenta potencial para verificar a relação entre construto e nota. Por meio do estudo da estrutura fatorial das variáveis é possível encontrar evidências empíricas sobre o significado do que está sendo medido.

A análise fatorial exploratória pode ser utilizada para reduzir variáveis ou agrupar variáveis que explicam um ou mais de um fator. A exploratória é também usada quando se quer estudar a estrutura fatorial de um determinado instrumento. Por meio da análise fatorial exploratória, é possível estudar um conjunto de variáveis que foram medidas de uma mesma maneira a fim de verificar como elas se agrupam em termos de importância. Como o objetivo era o de ‘raquear’ a nota da prova oral do Celpe-Bras,

ou seja, entender como as seis variáveis da nota analítica e a nota do avaliador interlocutor estão compondo o fator *proficiência oral,* na prática e com o uso da grade pelos avaliadores, foi feito uma análise fatorial exploratória. Todas as notas atribuídas aos sete itens foram levados em conta no cálculo que teve como base a análise fatorial dos eixos principais (*Principal Axis Factoring, PAF*). Nesta análise, pode-se concluir que as sete variáveis podem ser representadas por apenas um fator. Não foi necessária a rotação de fatores porque a análise se reduziu a um fator. No anexo quatro, coloco os valores de ajuste de modelo quando foram testadas as hipóteses sobre a estrutura fatorial das notas estarem organizadas em um ou dois fatores. Cabe ressaltar que foram feitas análises preliminares utilizando equações estruturais e análise confirmatória, porém detectou-se problemas de convergência.

Por questões de reprodutibilidade, o algoritmo para cálculo dos intervalos de confiança para os pesos e cargas fatoriais, utilizando *bootstrap* (DAVISON, HINKLEY, 1997; CANTY, RIPLEY, 2017), está incluído no anexo quatro. A estimação por meio de *bootstrap* faz uso de conceitos do teorema central do limite. Independente da forma da distribuição dos dados, a distribuição amostral dos parâmetros de interesse consegue assumir uma distribuição normal. As notas orais do Celpe-Bras são assimétricas, porém a aplicação dos conceitos do teorema pode garantir resultados precisos para os valores calculados, independente da forma que os dados se apresentam. A adoção de estimação por reamostragem ou *bootstrap* se fez necessária para garantir a correta aplicação do teorema central do limite aos dados. Este método consiste em fazer sucessivas amostragens nos dados disponíveis e calcular os valores de interesse. Após as sucessivas amostragens, o valor final será a média dos valores calculados. No caso em estudo, foram retiradas 10000 amostras com reposição, de tamanho 1000 dos dados em estudo.

A análise que apresentarei está organizada em duas partes. Na primeira parte, analisarei a estrutura fatorial das medidas que correspondem à nota final do observador, que é composta por seis notas atribuídas aos seguintes parâmetros: *compreensão, competência interacional, fluência, adequação lexical, adequação gramatical, pronúncia.* Na segunda parte da análise, acrescento a nota do entrevistador aos seis parâmetros da grade analítica, de forma a analisar a estrutura fatorial da nota final da prova oral.

5.2.1 RESULTADOS DA ANÁLISE FATORIAL DA NOTA ANALÍTICA

Na primeira análise fatorial, pressupõe-se que um fator *proficiência oral* esteja sendo explicado por seis variáveis, que seriam os seis itens que compõem a nota analítica. Para fazer a avaliação de ajuste local do modelo, analisou-se o coeficiente de determinação, o R^2, que se refere à porcentagem de variação das variáveis, notas da prova oral, que estão sendo explicadas pela estrutura fatorial calculada. A estrutura fatorial apresentada mostrou um R^2 de 0.9409718. Isso indica que que os dados se adequaram ao modelo de análise e por ele podem ser explicados. Para fazer a avaliação do ajuste de modelo, apresento o cálculo do RMSEA (*root square erros of approximation*) cujo valor foi de 0.23. Sugere-se como índice de bom ajuste que o valor fique em torno de 0.5. No caso da análise, uma hipótese possível para o alto valor do índice de RMSEA pode ser o fato das notas estarem muito fortemente correlacionadas. O Tucker Lewis Index (TLI) é outra maneira de avaliar a confiabilidade dos resultados calculados pelo modelo da análise. No caso da análise, o valor foi o de 0.896, trata-se de um valor satisfatório, sugerindo que os dados possam ser explicados pela método de análise adotada. Outros valores de ajuste podem ser avaliados no anexo.

Com o objetivo de analisar quais parâmetros analíticos são mais importantes para construção da nota do avaliador-observador e também para saber com quantos fatores ou dimensões do construto estamos lidando, analisamos as cargas de fatores, peso e valores de comunalidades. Na tabela abaixo, apresento o resultado da análise:

Tabela 4 - Valores da análise fatorial dos parâmetros analíticos

	carga	peso	comunalidade
compreensão	0.6565503	0.0767320	0.4310583
competência Interacional	0.8029686	0.1535918	0.6447585
fluência	0.8827109	0.2557081	0.7791786
adequação lexical	0.9073271	0.2980357	0.8232424
adequação gramatical	0.8871079	0.2110563	0.7869604
pronúncia	0.7982720	0.1051046	0.637238

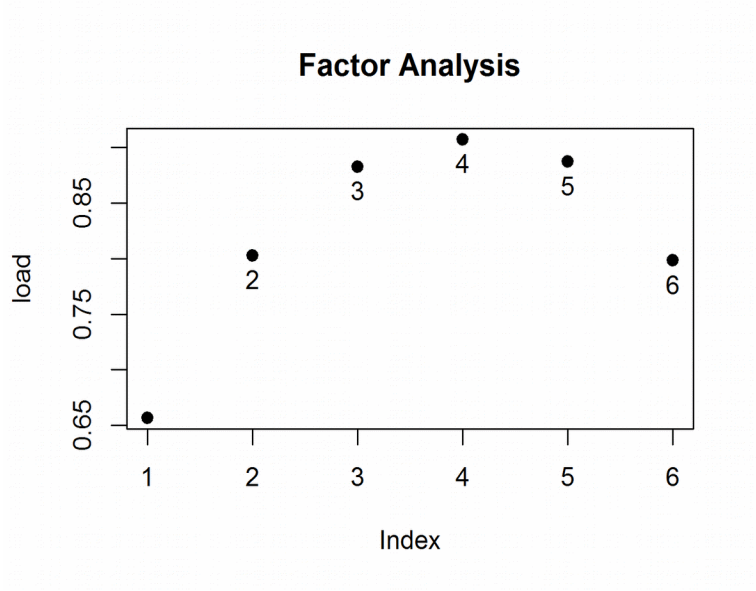
De acordo com Kim e Mueller (1978), o cálculo da comunalidade se dá a partir da correlação de cada variável com o restante do conjunto das variáveis, ou seja, ela tenta quantificar como a nota de compreensão, por exemplo, está correlacionada com

o conjunto das outras notas. Figueiredo Filho e Silva Júnior (2010) explicam que é a partir do valor de comunalidade que podemos inferir que uma variável está linearmente correlacionada com as outras. Os autores afirmam que valores baixos de comunalidades (menores que 0.50) significam que elas possam não estar linearmente correlacionadas. Quanto aos valores de comunalidade da tabela acima, temos um valor ligeiramente baixo para a nota de compreensão, sugerindo que o parâmetro possa estar menos relacionado com outros. Como há um intervalo de confiança na estimativa do valor de 0.43 para comunalidade, não é tão problemático o valor deste item. Os demais valores de comunalidade são acima de 0.5, por isso, parece provável que eles estejam medindo a mesma coisa, ou relacionados a um mesmo fator, que seria a proficiência oral.

A partir dos valores da carga na tabela acima, é possível afirmar novamente que as variáveis estão relacionadas ou que explicam um mesmo fator. Por meio da análise dos valores de carga fatorial, sugere-se que um só fator esteja influenciando os valores das notas atribuídas aos parâmetros, e por isso podemos afirmar que a medida é unidimensional. Afirmar que a medida é unidimensional, no nosso contexto, é o mesmo que dizer que as notas estão relacionadas a uma coisa só, que é a proficiência oral. De modo geral, a análise fatorial da grade analítica corrobora a análise da matriz de correlação, uma vez que as medidas estão fortemente correlacionadas. No entanto, a nota de compreensão na análise fatorial demonstra relação fraca com os outros parâmetros. Embora a matriz de correlação apresente uma correlação perfeita entre os parâmetros da *adequação lexical* e da *adequação gramatical*, seus valores na análise fatorial são distintos, sugerindo que tais parâmetros possam estar medindo dimensões diferentes do construto da proficiência oral.

Os itens que têm a maior carga fatorial na nota final analítica, do maior para o menor com o intervalo de confiança, são: *adequação lexical,* 0.90 (0.89-0.92), *fluência,* 0.88 (0.87-0.90), *adequação gramatical,* 0.88 (0.89-0.92), *competência interacional,* 0.80 (0.77-0.83), *pronúncia 0.80 (0.77-0.83)* e *compreensão,* 0.65 (0.60-0.69). Por meio da representação gráfica da carga fatorial de cada um dos parâmetros, é possível afirmar que *compreensão* explica pouco a nota analítica, quando comparado aos demais parâmetros. *Competência interacional* e *pronúncia* são variáveis que contribuem de maneira aproximadamente igual para o fator proficiência oral, assim como *fluência* e *adequação gramatical*, por apresentarem valores aproximados de carga. A *adequação lexical se s*obressai com um valor de 0.90.

Gráfico 2 - Análise fatorial dos parâmetros analíticos



O gráfico acima organiza os valores da carga fatorial ou *loadings*, explicitados na tabela 4, no eixo x em função dos seis parâmetros analíticos, sendo o 1 a *compreensão*, o 2 *competência interacional,* o 3 *fluência*, o 4 *adequação lexical*, o 5 *adequação gramatical* e o 6 o da *pronúncia*. Graficamente percebe-se como os valores da carga de fatorial para as variáveis *fluência*, *adequação lexical*, *adequação gramatical* são superiores e por isso estão representados nos pontos superiores do gráfico. No extremo, temos o valor para *compreensão* abaixo e à esquerda, quase colado ao valor 0,65 de carga fatorial.

A partir da análise empírica, novos valores aproximados para o peso dos parâmetros foram calculados com base no valor de pesos da análise fatorial (tabela 4). Cabe ressaltar que os valores referente ao intervalo de confiança dos pesos estimados podem ser verificados em anexo. Os valores em pesos representam o quanto cada parâmetro deve compor a nota do observador. Os valores em peso calculados pela análise são aproximados e, ao somá-los, chegaríamos a um valor aproximado de 107%. Assim sendo, recalculei os valores de forma a acomodá-los na métrica de 100% e apresento na tabela em seguida (Tabela 5). Nesta tabela, comparo os pesos que vigoram na composição da nota e os pesos aproximados propostos, fundamentados na análise fatorial para composição da nota final a partir do recálculo.

Tabela 5 - Comparação entre peso atual e peso estimado pela análise fatorial para composição da nota final do observador

parâmetros analíticos	peso atual	valor do peso estimado na análise fatorial
compreensão	16.67%	6.34%
competência interacional	16.67%	13.23%
fluência	16.67%	25.57%
adequação lexical	21%	27.09%
adequação gramatical	21%	18.80%
pronúncia	8%	8.97%

Ao comparar os novos pesos, os valores de peso para *adequação gramatical, competência interacional e pronúncia* são aproximados aos que vigoram, conforme pode ser observado na tabela 5. O valor que sofreu maior redução corresponde à *compreensão,* que no recálculo perde 10% de peso na composição da nota analítica, indo de 16.6% para 0.63%. Os parâmetros que tiveram seus valores aumentados na proposta empírica de composição da nota analítica são o da *fluência* e o da *adequação lexical*. O novo valor dos pesos propostos indicam que os valores para *adequação lexical* e *fluência* representam um pouco mais que 50% da nota analítica. Ao somarmos os pesos de *adequação lexical, fluência e adequação gramatical* temos aproximadamente 70% da nota analítica. Isso significa dizer que empiricamente 70% da nota analítica pode ser explicada por esses três parâmetros.

De forma geral, podemos dizer que as variáveis *adequação lexical, fluência e adequação gramatical* são os parâmetros da grade analítica que mais explicam ou pesam na avaliação da proficiência oral do observador da forma como está organizada a situação da entrevista e as tarefas na proposta de avaliação do Celpe-Bras. Se compararmos o resultado da análise com o estudo de Hinofotis de 1983 (*apud* Fulcher 2003) sobre os parâmetros de avaliação da proficiência oral de docentes em situação de sala de aula, os parâmetros vocabulário e gramática também estavam relacionados entre si e compunham o que o autor chamou de fator linguístico da proficiência oral. Além disso, outro resultado do presente trabalho que corrobora as conclusões de Hinofotis é o fato dos valores da análise fatorial da grade analítica sugerirem que a pronúncia esteja menos relacionada aos parâmetros gramaticais e lexicais. Hinofotis também encontrou que a pronúncia era uma variável que empiricamente não estava relacionado com o que ele denominou de fator linguístico da proficiência oral.

Curiosamente, na descrição geral sobre as faixas de certificação do Celpe-Bras, a faixa de proficiência referente ao intermediário superior é definida da seguinte

forma “Conferido a examinandos/as que preenchem as características descritas no nível Intermediário, com a diferença de que, nesse nível, as inadequações e interferências de língua materna e/ou de outra(s) língua(s) estrangeira(s) na pronúncia e na escrita devem ser menos frequentes que naquele nível.” (BRASIL, 2015, grifo meu). Embora a pronúncia seja o parâmetro que menos explique ou contribua para a composição da nota da proficiência oral tanto nos pesos atuais quanto nos pesos propostos empiricamente, o parâmetro foi o escolhido para descrever a proficiência da faixa intermediário. Mais coerente com a presente análise seria descrever a proficiência oral com os parâmetros que estão empiricamente mais relacionados com o construto operacionalizados no exame. Segundo os valores de carga fatorial dos parâmetros analíticos, estes parâmetros seriam as adequações lexical e gramatical e a fluência.

Na presente análise, o parâmetro *fluência* está mais relacionado com a*dequação lexical e adequação gramatical* do que com a *competência interacional.* Hinofotes, em 1983, concluiu que *competência interacional* e *fluência* seriam variáveis próximas, com cargas fatoriais semelhantes, ao passo que neste trabalho contatou-se que os valores de cargas fatoriais da *fluência* está mais próxima das *adequações gramatical e lexical* do que da *competência interacional.* Uma hipótese para que a nota de fluência esteja próxima às adequações gramatical e lexical poderia estar associada à maneira como a fluência está descrita na grade. De acordo com a grade do observador, a fluência está relacionada à pausas e hesitações para organização do pensamento e resolução de problemas linguísticos. Ressalta-se que o que é graduado nos descritores são as frequências de resolução de problemas linguísticos que podem estar fortemente associadas ao uso de estruturas lexicais e gramaticais.

5.2.2. ANÁLISE FATORIAL DA NOTA DO OBSERVADOR E DA NOTA DO ENTREVISTADOR

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

avaliação do ajuste de modelo, apresento o cálculo do RMSEA (*root square erros of approximation*) cujo valor foi de 0.18. Sugere-se como índice de bom ajuste que o valor fique em torno de 0.05. No caso da análise, uma hipótese possível para o alto valor do índice de RMSEA pode ser o fato das notas estarem muito fortemente correlacionadas. O Tucker Lewis Index (TLI) é outra maneira de avaliar a confiabilidade dos resultados calculados pelo modelo da análise. No caso da análise, o valor foi o de 0.896, trata-se de um valor satisfatório, sugerindo que os dados possam ser explicados pela método de análise adotada. Outros valores de ajuste podem ser avaliados no anexo.

Com o incremento da sétima variável, analisamos a nota final da prova oral como se a nota do entrevistador fosse um dos parâmetros, desconsiderando o atual peso de 50% na composição da nota final da prova oral. Ao recalcularmos os valores para chegar às variáveis que mais explicam o fator ou construto da proficiência oral, temos a nota do entrevistador como a variável mais importante, e os demais valores bastante semelhantes à análise anterior em que apenas os seis parâmetros entraram na análise.

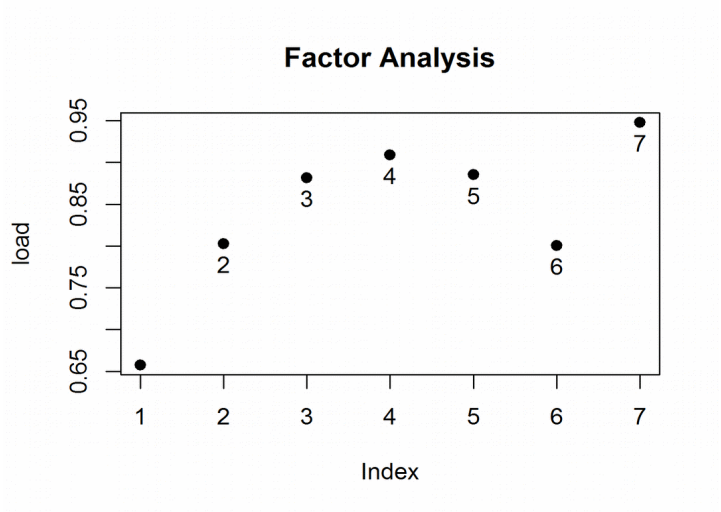
Tabela 6 - Características e pesos a serem atribuídos para cada variável analítica e nota entrevistador

	carga	peso	comunalidade
compreensão	0.6572906	0.0455407	0.43
competência Interacional	0.8028341	0.0950375	0.64
fluência	0.8816178	0.1836518	0.78
adequação lexical	0.9092421	0.1946057	0.83
adequação gramatical	0.8852773	0.1350449	0.78
pronúncia	0.8004653	0.0644487	0.64
nota entrevistador	0.9481050	0.3644149	0.90

Com o incremento da nota do entrevistador na análise, podemos novamente afirmar que trata-se de variáveis de um mesmo fator, ou seja, a medida é unidimensional. Os parâmetros que têm o maior carga fatorial na nota final da prova oral, do maior para o menor, são: *nota do entrevistador* (0.94), *adequação lexical* (0,90), *fluência* (0,88), *adequação gramatical* (0,88), *competência interacional* (0,80), *pronúncia* (0,80) e *compreensão* (0,65). Por meio da representação gráfica da carga fatorial de cada um dos parâmetros no gráfico a seguir (Gráfico 3), é possível afirmar que *compreensão* continua explicando pouco a nota final da prova oral. Após o incremento da nota do entrevistador aos demais parâmetros na composição da nota

final da prova oral, a nota do entrevistador é o parâmetro que mais explica a nota da prova oral quando comparamos separadamente esta variável com as demais. Isso pode ser inferido pois a variável 7 aparece no topo do gráfico abaixo sobre a análise fatorial das sete variáveis. Assim como na análise anterior, *competência interacional* e *pronúncia* são variáveis que contribuem de modo aproximadamente igual para o fator proficiência oral, assim como *fluência* e *adequação gramatical*, por apresentarem valores aproximados de carga. Se sobressaem com o valor de 0,94 para o critério *nota do entrevistador* e 0,90 para o critério *adequação lexical*. Os valores do peso representam o quanto cada aspecto contribui para a composição da nota final da avaliação oral, considerando a nota do entrevistador como um sétimo parâmetro.

Gráfico 3 - Análise fatorial das sete variáveis



Assim como no gráfico anterior de cargas de fatores, o gráfico acima organiza os valores da carga fatorial ou *loadings*, explicitados na tabela 6. O eixo x representa os valores das cargas que estão em função de sete parâmetros, sendo o 1 *compreensão*, o 2 *competência interacional,* o 3 *fluência*, o 4 *adequação lexical*, o 5 *adequação gramatical*, o 6 pronúncia*,* o 7 *nota do interlocutor*. A disposição dos pontos no gráfico demonstram os valores das cargas, sendo os pontos mais baixos os que menos explicam a nota da prova oral e os mais altos, o que mais contribuem para a avaliação oral.

Nos três gráficos a seguir, comparamos os pesos que vigoram atualmente na composição da nota com os valores estimados a partir das análises que apresentamos acima. No gráfico 5, organizamos os valores referentes à proposta de recálculo a partir

Gráfico 4 – Composição atual da prova oral

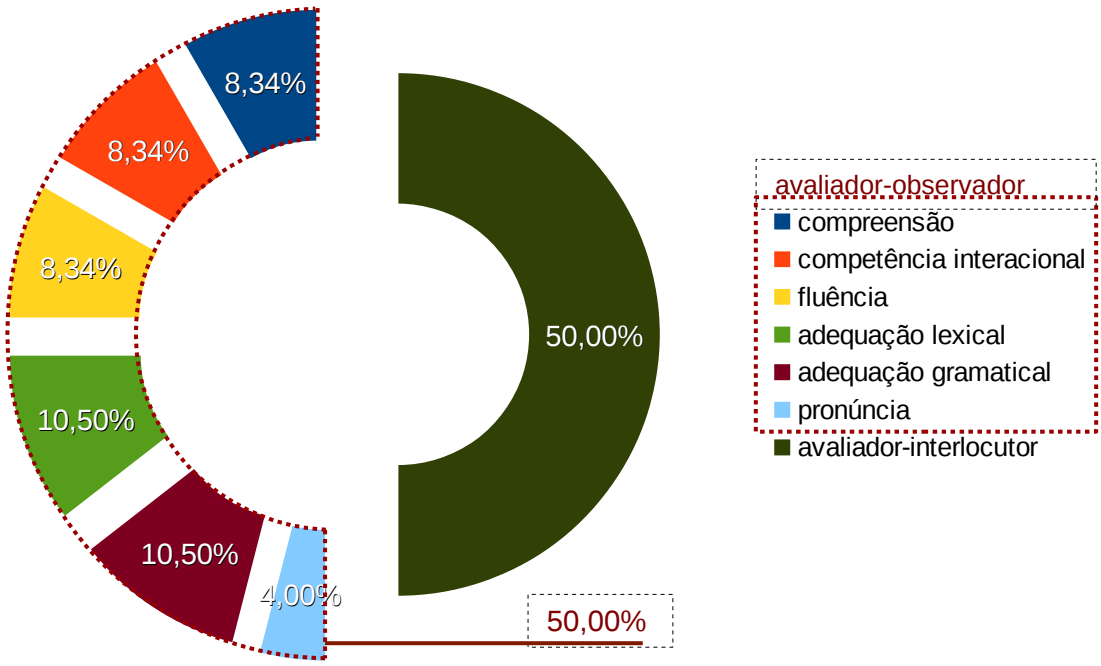
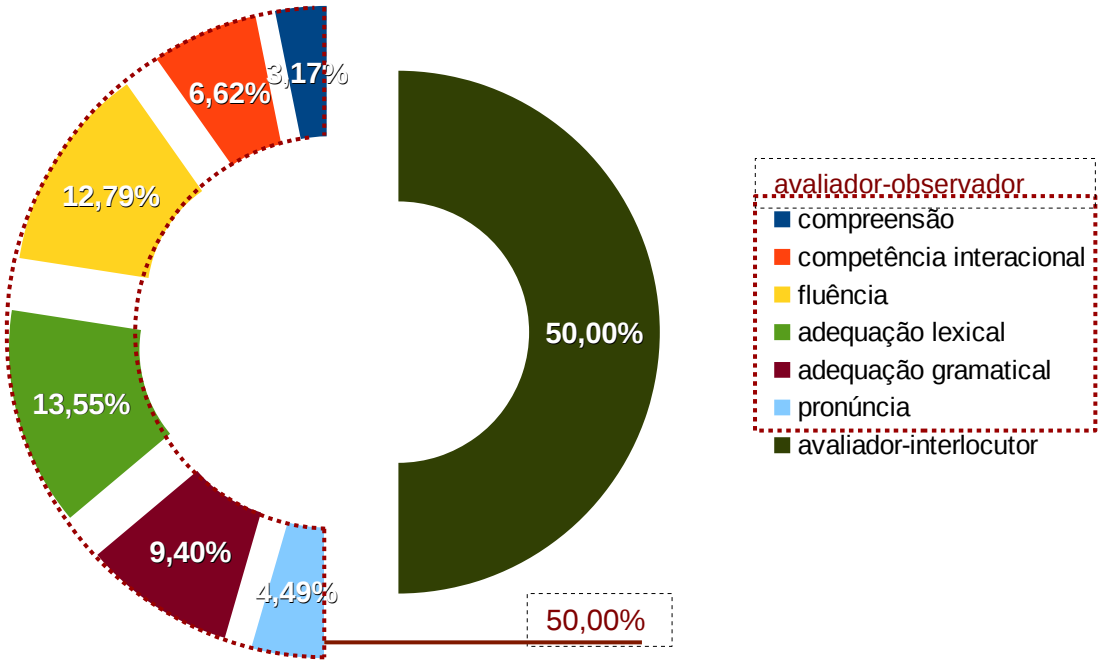
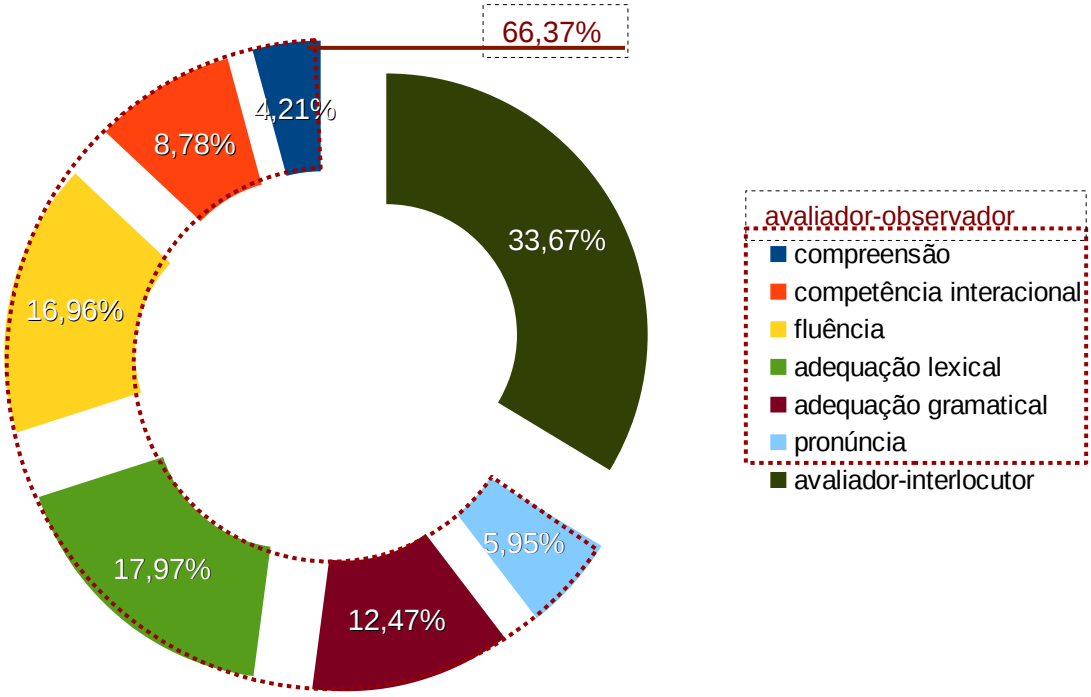


Gráfico 5 – Composição estimada da nota dos parâmetros da grade analítica, mantendo o peso atual da nota do avaliador-interlocutor



dos parâmetros analíticos na composição da nota do observador, considerando que a composição dos parâmetros analíticos teria 50% de peso na composição da nota final da prova oral. No gráfico 6, desconsidera-se o peso de 50% da nota do avaliador-interlocutor e os valores dos seis parâmetros da grade analítica e da nota do avaliador-interlocutor formam sete variáveis na composição da nota final na prova oral, ou seja, considera-se a nota do avaliador-interlocutor como uma variável. Vale ressaltar que o parâmetro que teve seu peso mais diminuído proporcionalmente foi, novamente, o da *compreensão.* Ou seja, com o incremento de um novo parâmetro (*nota do interlocutor), compreensão* perdeu ainda mais peso na composição da nota. Os outros parâmetros transferiram, em geral, um terço do seu peso para *nota do interlocutor*. Os valores relacionados ao intervalo de confiança dos valores de peso estimados para os sete itens podem ser verificados no anexo quatro.

Gráfico 6 – Composição estimada da nota da prova oral

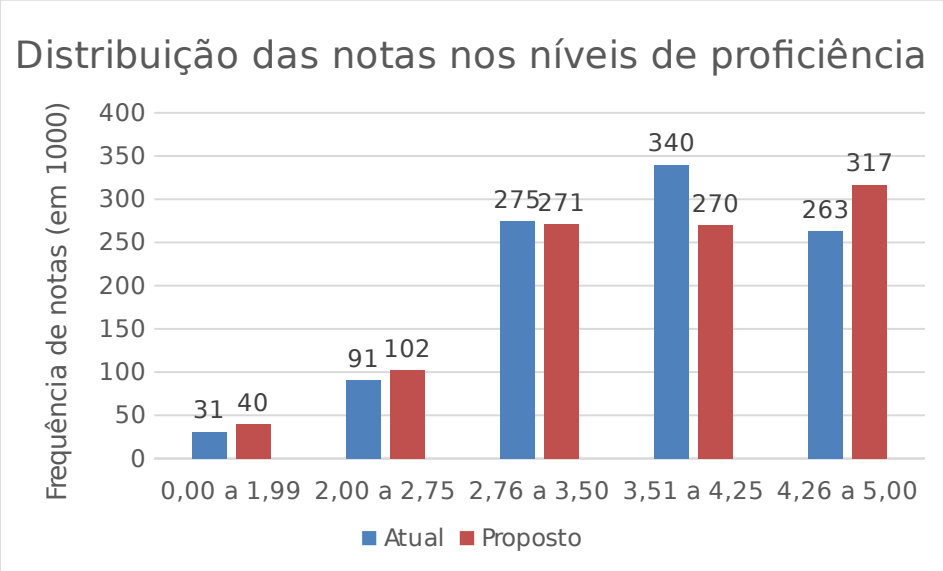


Com relação ao peso do conjunto de parâmetros que compõem a nota do *avaliador-observador* e da nota única do *avaliador-interlocutor*, embora a nota do *avaliador-interlocutor* seja o parâmetro que mais explica a nota da prova oral com um valor de 33.67%, a nota do *avaliador-observador*, ou seja, a composição entre as outras seis notas é a que explica mais a nota oral final. Ao somarmos o peso dos seis

parâmetros que compõem a nota analítica, temos 66.34% da nota final da prova oral explicada pela soma dos pesos de *compreensão*, *competência lexical*, *fluência, adequação lexical, adequação gramatical* e *pronúncia*. Dizendo de outra forma, a nota do observador é mais importante do que a nota do interlocutor, porque ao somarmos os pesos dos parâmetros analíticos na composição da nota final temos mais de 50% da composição da nota final explicada pela nota atribuída pelo *avaliador-observador.*

No gráfico 7, referente à comparação da distribuição das notas em faixas de proficiência, apresentamos nas colunas em azul a classificação da proficiência oral a partir da composição da nota que vigora atualmente. Os intervalos de faixas são definidos da seguinte forma: examinandos com notas entre 0,00 a 1,99 são classificados em *sem certificação*; entre 2,00 a 2,75 os examinandos são classificados como *intermediário*; entre 2,76 a 3,50 são *intermediário superior*; de 3,51 a 4,25 são *avançado;*e entre 4,26 a 5 os examinandos são classificados como *avançado superior*. O recálculo e reorganização dos examinandos nas faixas foi feito a partir do conjunto de dados que correspondem às notas de 1.000 examinandos. A classificação por faixas de proficiência a partir da nota final da prova oral foi comparada, considerando-se a composição da nota final a partir dos pesos atuais, nas faixas azuis, e a partir dos novos pesos propostos, nas faixas vermelhas do gráfico 7.

Gráfico 7 - Comparação da distribuição das notas em faixas de proficiência

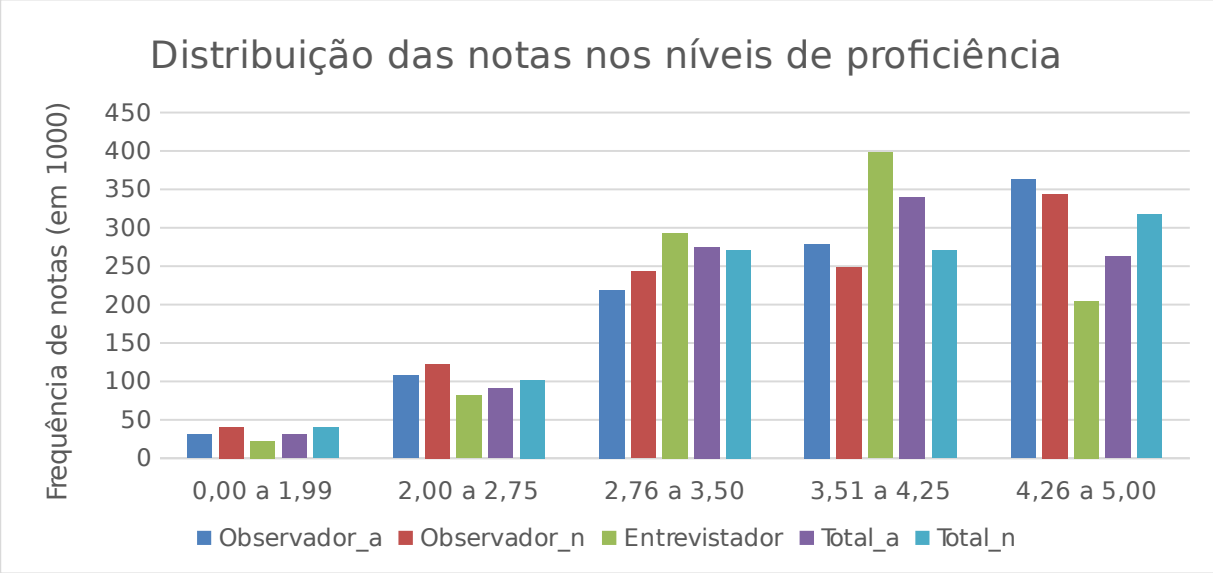


De maneira geral, os examinandos da faixa *avançado* espalharam-se por outras faixas de classificação, concentrando-se especialmente em *avançado superior*. Ao compor a nota final com os novos pesos, as faixas *sem certificação* e *intermediário*

aumentaram em comparação com a classificação feita a partir da composição da nota com os pesos que agora vigoram.

Para entender melhor o porquê do aumento dos examinandos classificados na faixa *avançado superior*, organizo por faixas de classificação da proficiência do exame as notas calculadas com os pesos atuais e propostos referente ao observador, ao interlocutor e à nota final no próximo gráfico.

Gráfico 8 - Comparação da distribuição das notas (observador, interlocutor e total) em faixas de proficiência



Em azul escuro (Observador_a) está representado a quantidade de participantes classificados nas faixas, considerando-se que apenas a nota do avaliador-observador fossem calculadas a partir dos pesos em vigor atualmente. As faixas vermelhas referem-se à nota do avaliador-observador calculada a partir dos pesos novos (Observador_n). A nota do avaliador-interlocutor, por ser uma nota única, não altera de valor, e por isso está denominada apenas como nota do avaliador-interlocutor em verde (Entrevistador). Em roxo está a nota total final, composta em 50% pela nota do *avaliador-observador* e os 50% restantes pela nota do avaliador-interlocutor, sem considerar os novos pesos (Total_a). A faixa em azul claro (Total_n) representa a nota final, que foi calculada a partir dos novos pesos para se chegar à nota analítica que é composta em 66.34% pela nota do avaliador-observador e em 33.67% pela nota do avaliador-interlocutor.

Houve uma tendência de aumento de examinandos classificados nas faixas *básico, intermediário* e *intermediário superior* quando comparamos a nota que foi calculada com os novos pesos com a forma atualmente em vigor. Após estes níveis a tendência se inverte, pois os examinandos nestas faixas diminuem. Ou seja, compondo a nota do observador com os pesos propostos, é provável que diminuam os examinandos classificados nas faixas *avançado* e *avançado superior* e aumentem os classificados nas faixas *básico, intermediário* e *intermediário superior.* O novo cálculo da nota do avaliador-observador pode reorganizar as classificações de forma a tender diminuir a nota de classificação do participante, uma vez que a proposta de novos pesos implica em colocar mais peso em parâmetros linguísticos nos quais os participantes tiram geralmente menores notas, e menos peso em parâmetros como o da *compreensão,* no qual 89% dos examinandos da amostra tiraram nota total. Embora haja esta tendência com relação a nota do observador, na soma da nota final não se verificou um aumento de examinandos classificados em *básico, intermediário* e *intermediário superior,* nem uma diminuição de examinandos em *avançado* e *avançado superior*. Isso se explica pelo fato da diminuição da nota do entrevistador na composição da nota final. A nota do entrevistador e a nota do observador, quando se trata da classificação nas faixas *avançado* e *avançado superior,* influenciam de formas distintas a composição da nota final. Na faixa *avançado* a nota final atual acompanha a nota do entrevistador. A nota final nova diminuiu porque foi composta mais pela nota do observador, responsável por 66.34%, do que pela nota do entrevistador, com 33.67% na composição da nota. É possível explicar o aumento do número de examinandos classificados na faixa *avançado superior* da mesma forma, uma vez que nesta faixa roxa, que corresponde à nota final atual, é menor que a faixa azul clara, que corresponde à nota final nova, composta com os novos pesos que aumentam a participação da nota do observador e diminuem a nota do entrevistador no cálculo final da nota da prova oral.

Embora os pesos dos parâmetros analíticos tenham sido propostos de forma a colocar mais ênfase nos aspectos linguísticos, provavelmente aumentando a dificuldade ou a probabilidade dos examinandos tirarem notas altas, pela análise, a nota do entrevistador parece estar descrita de forma que seja difícil os examinandos serem classificados na faixa *avançado superior. S*egundo os descritores das faixas na grade do entrevistador em anexo (Anexo 1, p. 144), o que diferencia a nota 4 da 5 é que a 5 “apresenta fluência e variedade ampla de vocabulário e de estruturas, com raras inadequações. Sua pronúncia é adequada” e em 4 “apresenta fluência e variedade ampla de vocabulário e de estruturas, com inadequações ocasionais na comunicação. Sua pronúncia pode apresentar algumas inadequações” enquanto que

em relação à autonomia, desenvoltura e compreensão os descritores são os mesmos. De acordo com a análise do gráfico 8, parece haver uma tendência do avaliador-entrevistador optar pelo nível *avançado* entre as faixas *avançado* e *avançado superior*, e, por isso, ao diminuir o peso da nota do entrevistador e aumentar a do observador a quantidade de examinandos classificados na faixa *avançado superior* aumentou. Ou seja, aumentar o peso dos parâmetros analíticos que se referem à aspectos linguísticos e aumentar o peso da nota do observador não significa necessariamente uma diminuição da nota final do examinando, porque o julgamento do entrevistador parece tender a concentrar a classificação dos examinandos na faixa *avançado,* quando em dúvida quanto à classificação entre *avançado e avançado-superior* em comparação com a nota atual ou nova do observador. Dessa forma, ao diminuir o peso da nota do entrevistador na composição da nota final nova, os examinandos foram reorganizados de forma a aumentar o número de classificados na faixa *avançado superior.*

5.3. ANÁLISE RASCH

Para analisar o quão discriminantes são os parâmetros de avaliação da escala analítica e a grade do entrevistador, adotaremos o Rasch básico, conforme já assinalado anteriormente. Doravante, cada um dos parâmetros da escala analítica e a nota do entrevistador serão chamados de itens. O Rasch básico foi o modelo mais apropriado para estudar o conjunto de dados coletados.

De maneira geral, o modelo especifica a dificuldade de cada um dos itens que compõem o teste e do próprio teste como um todo. Por meio da análise da dificuldade, é possível estudar a discriminação entre cada faixa de classificação de cada um dos itens do teste e da totalidade do mesmo.

Conforme já discutido anteriormente, os dados coletados para a análise tratam de mais de duas categorias de resposta – isto é, são politômicos –, uma vez que o examinando pode ser classificado quanto à sua proficiência em uma escala de zero a cinco pontos para cada um dos parâmetros de avaliação. Segundo Eckes (2015), os modelos politômicos são uma extensão do modelo Rasch e levam em conta itens politômicos, ou seja, que permitem mais de duas categorias de resposta. O autor explica que dados politômicos são resultado do uso de escalas para avaliação, como as grades da prova oral do exame Celpe-Bras. A grade de avaliação gradua ou divide a resposta ao item em mais possibilidades. Há dois tipos de modelos politômicos: o *Rasch Scale Model* e o *Partial Credit Model*. No primeiro modelo, calcula-se, dentre

outras variáveis, a dificuldade de transição de uma faixa de proficiência para outra. O *Partial Credit Model* tem uma estrutura matemática semelhante à do primeiro modelo; porém, além de possibilitar o cálculo da dificuldade de transição de uma faixa de proficiência para outra, também estima diferentes valores de *threshold* para itens distintos. Eckes (2015) sugere que o *Partial Credit Model* seja usado quando os itens apresentam um número diferente de categorias de resposta ou quando a dificuldade entre as categorias varia de item para item, porque o modelo permite investigar a estrutura individual de cada item de avaliação. Eckes (2015) afirma que o Partial Credit Model permite calcular a dificuldade específica de cada uma das categorias de resposta para cada um dos itens analisados, por isso será possível estudar a discriminação de cada um dos itens da escala da prova oral do Celpe-Bras.

Categorias de respostas se referem às faixas de classificação ou notas. No nosso caso estariam relacionadas às notas de zero a cinco e os respectivos níveis de proficiência, indo do nível *sem certificação* ao *avançado superior*. Os itens se referem a cada um dos parâmetros analíticos da grade usada pelo observador (*compreensão, competência interacional*, etc.) e à nota única do entrevistador, que estamos considerando como um parâmetro de avaliação. Ou seja, a nota do entrevistador é entendida aqui como um item, assim como a nota de *compreensão* e as demais notas analíticas. A dificuldade de pular de uma faixa de proficiência na escala pode variar de um item para outro; a diferença da dificuldade de mudança de nível entre *intermediário superior* e *avançado* e entre *avançado* e *avançado superior* são distintas, a depender do item ou parâmetro que está sendo avaliado. Por exemplo, a partir da nota atribuída ao item *compreensão*, examinandos com proficiência mediana tendem a ser mais facilmente classificados nas faixas *avançado* e *avançado superior* quando comparamos as notas atribuídas ao item *adequação lexical*. Por pressupor que os itens possam discriminar de forma diferente os examinandos ao longo das faixas de proficiência, utilizo na análise a extensão *Partial Credit Model* do Rasch (MAIR *et al.,* 2018) versão 0.16-0.

Na tabela 7, que apresento a seguir, ao avaliar a quantidade de notas por cada um dos parâmetros de avaliação da grade analítica e da nota do entrevistador, percebe-se uma carência de dados relacionados com as notas zero, um e dois. Ou seja, temos poucos examinandos com notas baixas no corpus que foi coletado. Cabe ressaltar que o conjunto de dados corresponde à 16,07% de examinandos inscritos no processo de certificação do Celpe-Bras no primeiro semestre de 2016.

A ausência de notas zero em *compreensão* fez com que os dados fossem organizados de forma que as notas 0 e 1 fossem agrupadas em uma categoria de

resposta que chamaremos de *categoria 0*. Esta ausência pode ser explicada a partir dos descritores. A nota zero em compreensão diz respeito a examinandos que apresentam “problema sério na compreensão do fluxo natural da fala. Necessidade de constante repetição e/ou reestruturação, mesmo em situação de fala muito simplificada e muito pausada.”, ou seja, pode descrever situações de avaliação em que não tenha havido interlocução. De certa forma, para haver interlocução é preciso que o examinando compreenda minimamente a fala do entrevistador e aparentemente este perfil de examinando é raro entre os inscritos no exame. Por este motivo, reagrupo as notas nas categorias que apresento na tabela 8. Trata-se da convenção que seguirei na exposição das análises.

Tabela 7 - Proporção de notas por cada dimensão do construto

	0	1	2	3	4	5
NT.Compreensão	0.000	0.010	0.023	0.077	0.190	0.700
NT.Competência interacional	0.004	0.018	0.066	0.195	0.295	0.422
NT.Fluência	0.001	0.030	0.084	0.236	0.283	0.366
NT.Adeq.Lexical	0.004	0.037	0.163	0.303	0.320	0.173
NT.Adeq.Gramatical	0.007	0.040	0.168	0.310	0.303	0.172
NT.Pronúncia	0.003	0.027	0.125	0.258	0.379	0.208
NT.Entrevistador	0.001	0.021	0.082	0.293	0.398	0.205

Tabela 8 - Convenção

correspondência entre significado das notas e faixa de certificação	notas do exame	convenção utilizada na análise
sem certificação ou básico	notas 0-1	categoria 0
intermediário	nota 2	categoria 1
intermediário superior	nota 3	categoria 2
avançado	nota 4	categoria 3
avançado superior	nota 5	categoria 4

5.3.1. AJUSTE DE MODELO E *ITEM FIT* E *OUTFIT STATISTICS*

Como já tratado anteriormente, a extensão *Partial Credit Model* do modelo Rasch (MAIR *et al.,* 2018) foi a que mais acomodou os dados, no entanto, cabe problematizar o quanto o teste como um todo e o quanto cada um dos itens se ajustaram ao modelo.

Para investigar o ajuste global, apresento o resultado do *Martin Lof Test.* Verguts e Boeck (2000) sugerem o *Martin Lof Test* para avaliar a unidimensionalidade de uma

escala, ou seja, itens politômicos. *Martin Lof Test* consiste em dividir o corpus ao meio e testar se há diferença entre eles. O critério escolhido para fazer o presente teste foi a mediada, por ser o critério padrão. O teste gera variados valores, dentre eles o *p-value.*

```
## LR-value: 501.595
## Chi-square df: 191
## p-value: 0
```

O valor de *p-value* é um medida de ajuste global ao modelo, espera-se que valor varie de 0 a 1, sendo o valor 0 que mais se relaciona à falta de ajuste e o valor 1 a um ajuste perfeito (VERGUTS; BOECK, 2000). A partir desse e dos outros valores acima, podemos concluir que há problemas no ajuste da escala, ou seja, a escala da prova oral do exame Celpe-Bras não se ajusta ao modelo Rasch, porque os escores são ou muito previsíveis ou imprevisíveis, segundo os parâmetros do modelo Rasch. Cabe a ressalta feita por DeMars (2010) de que o item ou um conjunto de itens deveria se acomodar ao modelo, senão o item ou a escala não deveria existir porque ela fere os pressupostos previstos pelo modelo.

Ao avaliar os valores de ajuste da média quadrada de INFIT MSQ (*Infit mean-square*) e OUTFIT MSQ (O*utfit mean-square*) de cada um dos itens é possível investigar como cada um deles se ajustaram ao modelo. Para avaliar o quão consistentes são os escores reais com relação às expectativas do modelo é preciso analisar os valores dos parâmetros de ajuste de item na tabela 9. Segundo Smith (1996), o propósito de analisar tais valores é o de gerar insumo para o debate sobre o controle de qualidade da medida, identificando aspectos dos dados que se encaixaram ou não nas especificações do modelo. Smith (1996) ressalta que a finalidade não é retirar ou não itens que não se ajustam, mas o de identificar e examinar o porquê do não ajuste, para então decidir aceitar, rejeitar ou modificar o item. No contexto da análise de escalas de itens politômicos, a modificação à qual se refere o autor pode envolver desde correções da maneira como a análise foi feita à retirada de alguma categoria da escala.

Na tabela 9, os valores de *p-value* de cada um dos itens sugerem que os itens *fluência, adequação gramatical, adequação lexical* e *nota do entrevistador* se ajustam ao modelo, os demais itens apresentam valores insatisfatórios, de zero a próximos de zero, ao seja, não correspondem às expectativas especificadas pela extensão *Partial Credit Model* do Rasch.

De maneira geral, os valores de *infit msq* e *outfit msq* são índices relevantes para calibrar exames, pois por meio deles os gestores de bancos de itens podem identificar itens problemáticos para exclui-los ou modificá-los de forma a tornar o teste mais consistente. Eckes (2015) afirma que valores altos para o resultado da diferença entre a média dos escores e as medidas esperadas pelo modelo resultam em valores altos de *outfit msq,* que não deveria exceder 2.0. Na tabela, apenas o item *compreensão* tem valor maior de 2. Para Smith (1996) itens com este padrão de ajuste ao modelo deveriam ser omitidos da escala. Linacre (2015 *apud* Aryadoust; Goh, 2009) também afirma que itens com valores de *outfit msq* acima de 2 são potencialmente problemáticos. Quanto aos valores de *infit msq,* tanto Eckes (2015) quanto Smith (1996) apontam que valores perto de 1 tanto para *infit msq* quanto para *outfit msq* sugerem um bom ajuste ao modelo, por isso boa qualidade da medida. Linacre (2015 *apud* Aryadoust; Goh, 2009) afirma que os valores deveriam variar entre 0.5 e 1.5, valores abaixo de 0.5 deveriam ser investigados. Destacam-se pelos valores de *infit msq* satisfatórios, os itens *fluência, adequação gramatical, adequação lexical*, *competência interacional* e *pronúncia*. Um pouco mais distante de 1 encontram-se os itens *compreensão* e, abaixo de 0.5, a *nota do entrevistado*r.

Tabela 9 – Parâmetros de ajuste de item

item	chisq	df	p-value	outfit msq	infit msg	outfit T	infit T
compreensão	2113.120	898	0.000	2.351	1.386	4.22	5.61
competência interacional	1172.627	898	0.000	1.304	1.114	3.67	2.20
fluência	652.512	898	1.000	0.726	0.737	4.72	-5.88
ad. gramatical	672.765	898	1.000	0.748	0.744	5.65	-5.96
ad. lexical	561.217	898	1.000	0.624	0.605	9.00	-9.80
pronúncia	1041.879	898	0.001	1.159	1.132	3.16	2.71
nota do entrevistador	378.713	898	1.000	0.421	0.431	15.85	-15.74

É possível investigar não só como cada um dos itens se ajustam ao modelo, mas também como cada um dos examinandos atendem ou não às expectativas da resposta estimada. Por isso, coloco em anexo (ANEXO 5) os valores de *person fit* para

cada um dos mil examinandos. Houve alguns casos de ajuste de pessoas em que os valores de *p-value* foram menores que 0.05, indicando falta de ajuste. Poderia ser possível investigar o não-ajuste se me fosse acessível as informações quanto à língua nativa do examinando, à quantidade de tempo que estou português ou até mesmo se fizeram ou não a prova em um mesmo posto de aplicação. Como uma das diretrizes de acesso aos dados do INEP é a desidentificação das informações, infelizmente, este tipo de análise não me foi possível.

A seguir, apresento as curvas de caraterísticas dos itens da escala da prova oral.

5.3.2. CURVA DE CARACTERÍSTICA DO ITEM *COMPREENSÃO*

Retomando as potencialidades de análise dos itens politômicos pelo Rasch, Eckes (2015) afirma que a qualidade de uma escala pode ser também avaliada a partir dos valores de *thresholds.* O autor explica que uma boa escala distingue ou discrimina com eficiência todos os níveis de proficiência dos examinandos. Isso significa que quanto maior a proficiência mais provável será do examinando ficar nos níveis mais altos da escala a partir da nota em todos os itens. Por exemplo, no nosso caso, seria o mesmo que esperar que um examinando que tivesse tirado uma nota 3 em *adequação lexical* também tirasse uma nota 3 ou próxima a esta em *compreensão* e nos demais parâmetros ou itens. Ou seja, é esperado de uma escala que os itens discriminem não exatamente da mesma forma, mas com a mesma eficiência. É recomendável que um examinando *intermediário superior* tire notas próximas de 3 em todos os itens da escala, mesmo havendo itens em que a nota tenderá a ser ligeiramente menor e outros em que tenderá a ser um pouco maior.

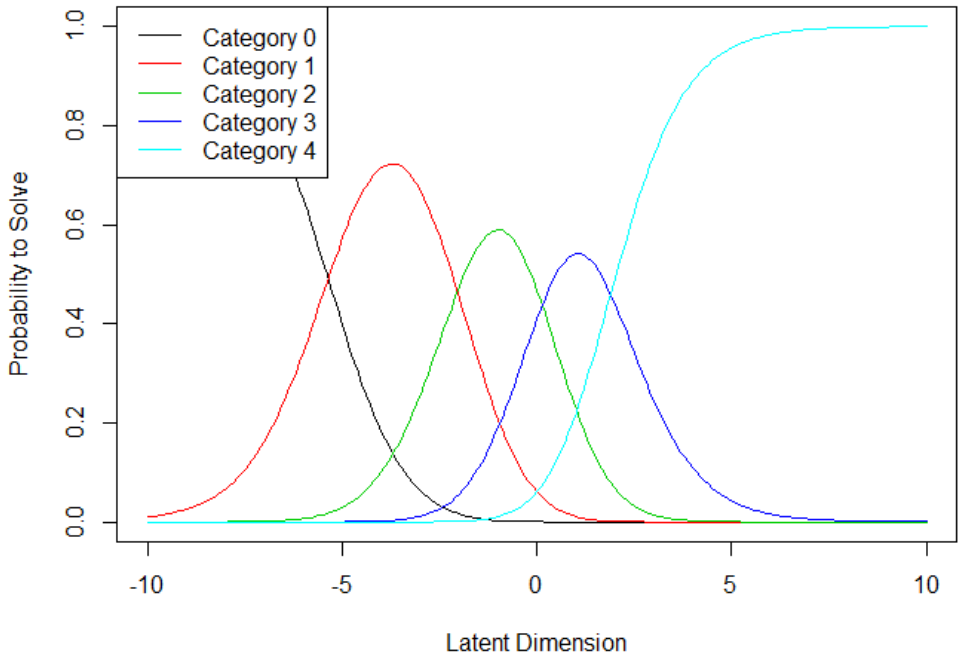
As diferenças entre notas e sua implicação na eficiência de discriminação dos examinandos ao longo das faixas de classificação das escalas – ou seja, a progressão da escala de cada item ou parâmetro de avaliação de acordo com a proficiência – tem como ser verificada a partir da análise dos valores que o Rasch denomina de *category threshold values*. *Category threshold values* é um intervalo estabelecido entre dois valores na métrica da habilidade estimada empiricamente pelo modelo. Retomando a discussão anterior, o valor da habilidade ou dimensão latente é estimado a partir da contabilização de acertos e erros dos examinandos. É importante ressaltar novamente que não há uma relação direta entre os valores que correspondem às faixas de proficiência atestada pelo Celpe-Bras e os valores da métrica da habilidade ou dimensão latente *(Latent Dimension)* no eixo horizontal dos gráficos 9 a 17 apresentados a seguir.

A partir dos valores de *thresholds* e sua relação com os valores da escala de perfis de habilidade dos examinandos, que foram estimados empiricamente, é possível organizar os gráficos da curva de característica de cada um dos itens. Nas curvas de característica de item politômico, representados nos gráficos 9 a 17, no eixo horizontal está organizada a escala de habilidade ou dimensão latente (*Latent dimension*), e no eixo vertical a probabilidade de acerto, que significa a probabilidade de um examinando estar ou não em determinadas categorias de resposta. Cada uma das curvas representam uma categoria de resposta.

Eckes (2015) sugere, como referência para uma boa escala, que os valores de *thresholds* entre uma faixa de certificação ou categoria de resposta e outra, devam ter um intervalo de 1.4 pontos e não mais que 5 pontos na escala de proficiência ou *Latent Dimension*, como está representado nos gráficos 9 a 17, apresentados nas próximas sessões. O autor explica também que quando os *thresholds* estão muito próximos na escala da proficiência, as faixas de proficiência são menos discriminantes do que deveriam ser. O intervalo da categoria de resposta informa o perfil de examinandos que a categoria abarca. Dizendo de outra forma, os valores na escala de habilidades correspondentes a uma categoria representam o perfil de examinandos que estão sendo classificados naquela categoria. Por isso, Eckes (2015) parte de estudos empíricos sobre escalas de itens politômicos para propor que este intervalo não possa ser nem muito pequeno, porque estaria abarcando um perfil reduzido de examinandos, nem muito grande, porque estaria contemplando um perfil amplo e, isso significaria que o item ou a escala do item não está discriminando entre diferentes perfis.

No gráfico 9, apresento a curva de característica do item *compreensão.* O eixo horizontal refere-se à escala de proficiência do examinando. O eixo vertical se refere aos valores de probabilidade dos examinandos que se encontram em determinados valores na métrica da habilidade serem classificados nas categorias analisadas, sendo a *categoria 0* correspondendo ao nível *sem certificação* ou *básico*, *categoria 1*, *intermediário*, etc. Ou seja, as curvas representam as categorias ou notas atribuídas a partir das escalas de certificação do exame Celpe-Bras. As curvas, que correspondem às faixas de certificação do exame Celpe-Bras, estão organizadas de maneira a informar a probabilidade para diferentes perfis de habilidade dos examinandos, representados nos eixos horizontais dos gráficos 12 em diante, de estar em uma ou outra categoria – ou seja, de ter tirado uma determinada nota de proficiência na prova oral do Celpe-Bras.

Gráfico 9 - Curva de característica do item *Compreensão*



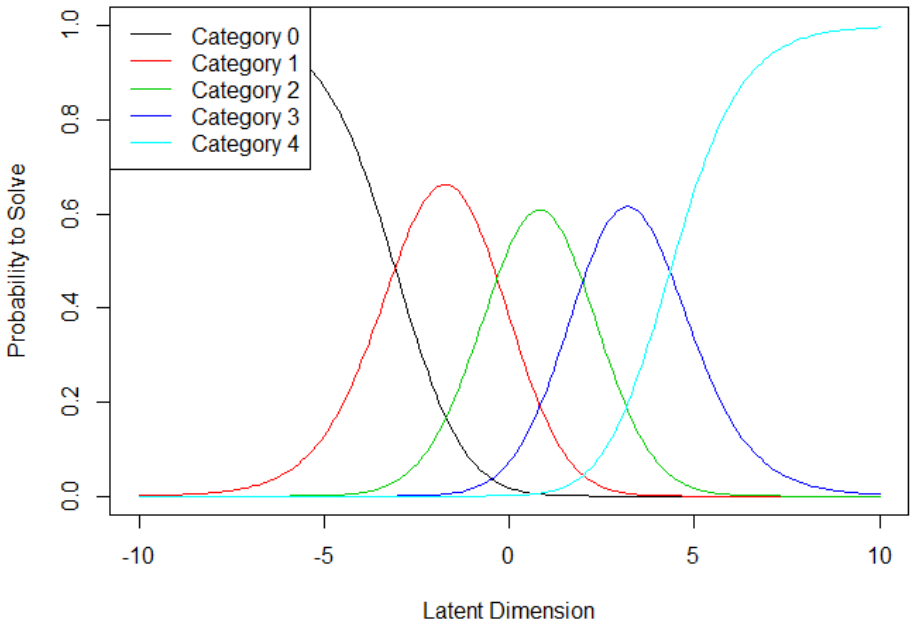
As curvas referentes às categorias, notas ou faixas de proficiência do item *compreensão* estão organizadas mais à esquerda na métrica da habilidade ou *Latent Dimension*. Isso significa que, de maneira geral, o item *compreensão* discrimina melhor examinandos com baixa proficiência ou, dizendo de outra forma, que o parâmetro *compreensão* é pouco eficiente para discriminar examinandos entre as faixas *intermediário superior, avançado e avançado superior*. Empiricamente, posso fazer esta afirmativa ao traçar uma reta paralela ao eixo horizontal, que se refere aos perfis de habilidade, na altura onde as curvas se encontram, ou seja, entre os pontos de *thresholds*. A partir da análise dos pontos determinados pelos valores de *thresholds*, que indicam onde os examinandos com um determinado valor de proficiência podem estar em uma categoria de resposta ou outra superior, é possível avaliar o quanto cada categoria é eficiente do ponto de vista da discriminação. A distância da interseção entre as categorias 0 e 1 é de 3 pontos na escala da proficiência. A distância da interseção entre as categorias 1 e 2 e 2 e 4 é de 2 pontos. Poder-se-ia dizer que o item discrimina melhor os examinandos com proficiência baixa, entre -6 e valores próximos de 1 no eixo horizontal que corresponde à escala da habilidade. A partir do valor 2, na escala dos valores de proficiência, os examinandos tendem a tirar nota máxima em *compreensão,* o que significa que, para examinandos que apresentam uma proficiência mediana ou máxima, este parâmetro de avaliação ou item não gera informações novas. Explicando de outra maneira, *compreensão* não distingue a proficiência oral de examinandos medianos dos que apresentam alta proficiência. Um examinando com valor de habilidade 1.5, no eixo horizontal, tem mais

probabilidade de estar na categoria 5 do que na 3, por exemplo. Outra observação a se fazer é que o item *compreensão* pouco discrimina as categorias 2 e 3 que representam as faixas *intermediário superior* e *avançado*, respectivamente, quando comparamos os valores que correspondem à discriminação entre as categorias 0 e 1, correspondentes à *sem certificação ou básico* e *intermediário*. O resultado corrobora a conclusão da análise de Schoffen (2003) que avalia o parâmetro *compreensão* como pouco importante para a definição das faixas avançada e intermediária[29].

5.3.3. CURVA DE CARACTERÍSTICA DO ITEM *COMPETÊNCIA INTERACIONAL*

No gráfico 10, apresento a curva de característica do item *competência interacional*. Assim como no gráfico 9, o eixo horizontal refere-se à escala de habilidade do examinando estimada empiricamente a partir do conjunto de notas, e o

Gráfico 10 - Curva de característica do item *Competência interacional*



eixo vertical se refere à escala de valores de probabilidade dos examinandos com determinados valores de proficiência poderem estar nas categorias analisadas, a depender da localização das curvas, que representam as categorias de resposta ao item *competência interacional*. As curvas representam as categorias ou faixas de

[29] Em 2002, o exame tinha apenas três faixas, a saber: sem certificação, intermediário e avançado.

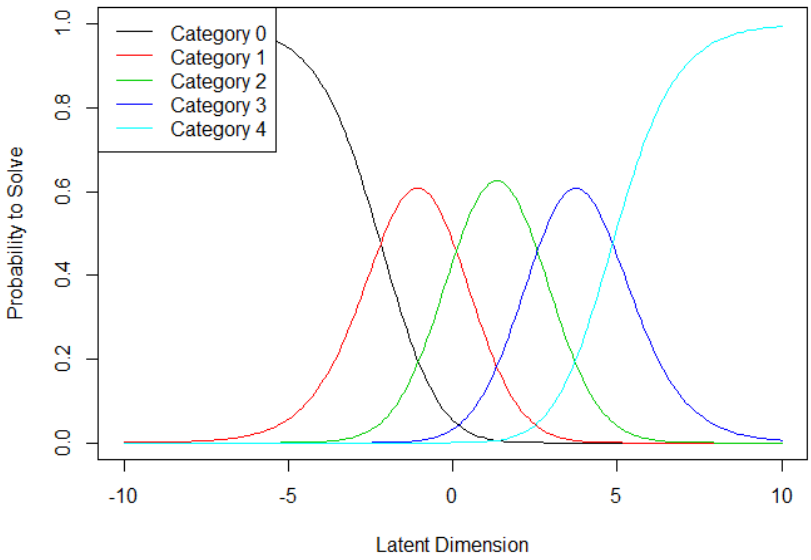
certificação. Nota-se que a faixa que representa *avançado superior* do item *competência interacional* alcança valores maiores no eixo horizontal, da dimensão latente, quando comparada à curva do item *compreensão*. Podemos inferir que a nota 5, correspondente a *avançado superior*, em *competência interacional* representa examinandos mais proficientes, quando comparamos com a nota 5, *avançado superior*, de *compreensão*. Ao traçar uma reta paralela ao eixo da habilidade na altura onde as curvas se encontram, ou seja, nos pontos determinados pelos valores de *thresholds*, que indicam onde os examinandos com um determinado valor de proficiência podem estar em uma faixa ou outra superior, é possível avaliar o quanto cada categoria é eficiente do ponto de vista da discriminação. A distância da interseção entre as categorias 0 e 1 é de 4 pontos aproximadamente na escala da proficiência. A distância da interseção entre as categorias 1 e 2, e 2 e 3 é de aproximadamente 2,5 pontos. Isto quer dizer que a distância entre os pontos de interseção, *thresholds*, das categorias medianas (1, 2 e 3) parecem ser mais semelhantes, quando comparadas às distâncias do gráfico da *compreensão*. O item *competência interacional* discrimina melhor os examinandos com proficiência mediana, entre -3 e valores próximos de 7 na métrica do eixo horizontal. Dizendo de outra forma, a escala que se refere ao item *competência interacional* discrimina melhor as categorias medianas, que representam as faixas *intermediário, intermediário superior* e *avançado,* do que o item *compreensão*. A partir do valor 7, na escala dos valores de proficiência, os examinandos tem 100% de chance de tirar nota máxima em *competência interacional*, ou seja, para examinandos que apresentam uma proficiência acima de 7, na escala da proficiência representada no eixo horizontal, o item *competência interacional* não gera informações novas sobre a proficiência do examinando.

5.3.4. CURVA DE CARACTERÍSTICA DO ITEM *FLUÊNCIA*

A curva de característica do item *fluência* assemelha-se às características do item *competência interacional*. Assim como o item *competência interacional*, *fluência* discrimina melhor os examinandos com valores medianos de habilidade, entre -3 e valores próximos de 7 na métrica do eixo da dimensão latente. Tanto o item *fluência* quanto o item *competência interacional* discriminam melhor as categorias medianas que representam as faixas *intermediário, intermediário superior* e *avançado* quando comparados ao item *compreensão*. Uma observação a se fazer é com relação à distância entre os pontos de interseção das categorias 2 e 3. A distância entre os valores de *threshold* dessas duas categorias apontam para uma capacidade um pouco

melhor de discriminação entre as notas 3 e 4, faixas *intermediário superior* e *avançado*, quando comparadas ao item *competência interacional*. Ou seja, a nota de *fluência* diferencia mais examinandos entre as faixas *intermediário superior* e *avançado* do que a nota de *competência interacional.*

Gráfico 11 - Curva de característica do item *Fluência*



5.3.5. CURVA DE CARACTERÍSTICA DOS ITENS *ADEQUAÇÃO LEXICAL E ADEQUAÇÃO GRAMATICAL*

O gráfico da curva de característica dos itens de *adequação lexical* e *adequação gramatical* são semelhantes. Ao analisarmos a distância entre os pontos de interseção, verifica-se que ambos os itens discriminam examinandos de habilidade oral elevada. O valor de *threshold* que corresponde à categoria 4, *avançado superior*, é de aproximadamente 7.25 pontos na escala de habilidade para ambos os itens. Isso significa que *adequação lexical* e *adequação gramatical* discriminam da mesma forma os examinandos com habilidade alta, representados pelos valores entre 7.3 e 9 pontos na escala da dimensão latente.

Retomando a definição de *Category threshold values,* tais valores referem-se ao intervalo estabelecido entre dois valores na métrica da habilidade estimada empiricamente pelo modelo. No gráfico da curva de característica de itens, o c*ategory threshold values* se refere à distância entre dois pontos na métrica da habilidade,

Gráfico 12 - Curva de característica do item *Adequação lexical*

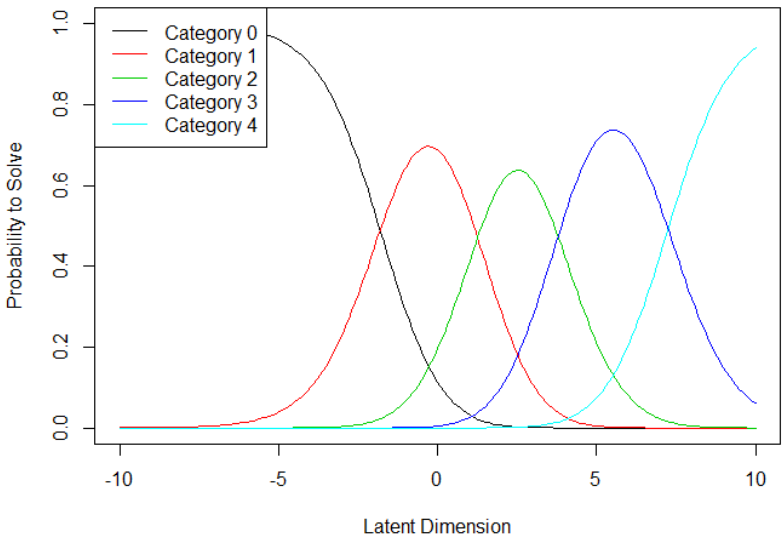
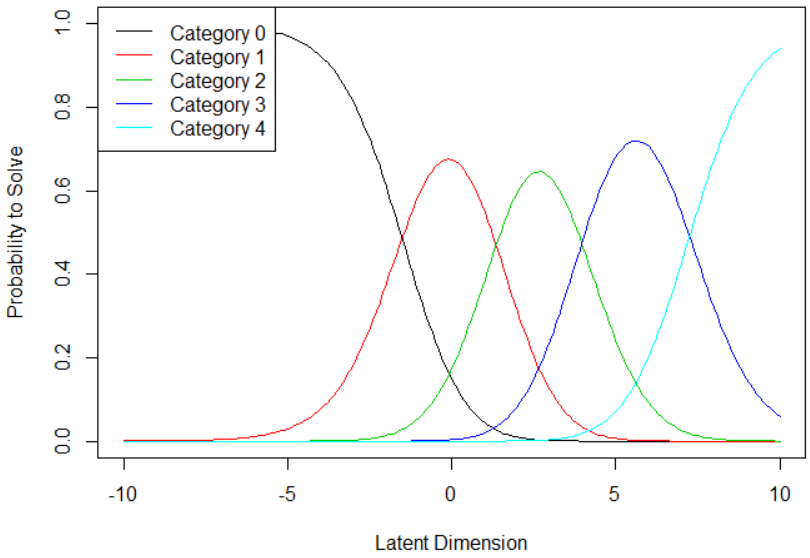


Gráfico 13 - Curva de característica do item *Adequação gramatical*



configurada no eixo horizontal, e representa o encontro entre duas categorias. Tais pontos se referem à probabilidade de um examinando estar em uma das duas categorias cujas curvas se cruzaram. Quando mais largo for o intervalo, mais perfis de examinandos poderão ser classificados no intervalo que corresponde a uma determinada categoria; quanto mais estreito menos perfis serão classificados. O intervalo de *threshold* das faixas *avançado* e *avançado superior* é de 3.3 pontos na escala da *adequação gramatical* e 3.5 pontos na escala da *adequação lexical,* respectivamente. Tomando como referência a sugestão de Eckes (2015) de que o intervalo não deve ser menor que 1.4 nem maior que 5, o intervalo dos itens podem ser considerados adequados e, mais do que isso, contribuem para diferenciar

examinandos com perfil de habilidade alta. A *categoria 3*, que representa a nota 4 ou a faixa *avançado* da escala oral do exame*,* é mais larga em ambos os itens quando comparada aos outros itens do exame. Por isso, podemos afirmar que a *adequação gramatical* e a *adequação lexical* são os itens que melhor discriminam entre os examinandos *avançado* e *avançado superior.*

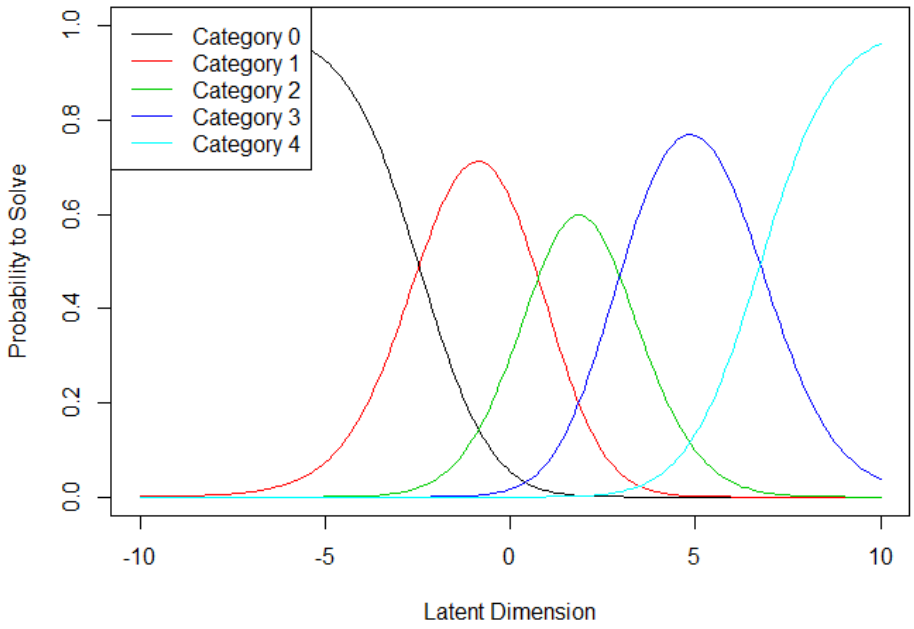
Além disso, o valor de *threshold* que corresponde à categoria 4 – *avançado* – para o item *compreensão* é de 1.89594; para o item *competência interacional*, 4.33582; *fluência*, 4.85844; *adequação lexical*, 7.25356; e para o item *adequação gramatical,* 7.25199. Quanto menor este valor, maior a probabilidade do item classificar examinandos de baixa habilidade, representados na métrica horizontal da dimensão latente na faixa *avançado-superior,* ou receber nota 5 no parâmetro de avaliação observado. Examinandos classificados no perfil próximos ao valor 7.25 apresentam alta probabilidade de serem classificados com nota 5 nos itens *adequação lexical* e *adequação gramatical*, ao passo que examinandos localizados próximos ao valor 1.89594 tirariam nota máxima no item *compreensão.*

A diferença entre *adequação lexical* e *adequação gramatical* está nos valores relacionados às categorias 1 e 2, que correspondem às faixas *intermediário* e *intermediário superior*. A distância entre os pontos de interseção da categoria 2 é menor para o item *adequação lexical* quando comparado ao da *adequação gramatical*. No item *adequação lexical*, o intervalo de respostas das categorias 1 e 3 são maiores que a categoria 2, assim, a distância entre as interseções da categoria 2 com outras categorias é mais estreita. Isso quer dizer que o item *adequação lexical* discrimina ainda melhor os examinandos cuja proficiência é baixa, uma vez que as categorias que representam essas faixas são menores, e cabe um espectro menor de perfis de examinandos. O fato do intervalo de respostas das categorias 1, *intermediário*, e 3, *avançado*, serem maiores que a categoria 2, *intermediário superior,* sugere que os descritores da faixa *intermediário superior* possam descrever um uso de estruturas lexicais que, na prática ou no contexto do exame, não é muito frequente. Segundo a grade analítica, a descrição da adequação lexical para *intermediário superior* se refere ao uso de “vocabulário adequado para 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”. Os descritores do nível *avançado* se referem ao uso de “vocabulário amplo e adequado para discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Poucas interferências de outras línguas”. Parece provável que os avaliadores-observadores estejam tendo dificuldades de diferenciar um uso de vocabulário *amplo e adequado* do uso *adequado* e com *algumas interferências* ou com *poucas*

interferências que caracterizam a diferença entre as faixas *intermediário superior* e *avançado,* a partir das descrições.

5.3.6. CURVA DE CARACTERÍSTICA DO ITEM *PRONÚNCIA*

Gráfico 14 - Curva de característica do item *Pronúncia*



O item *pronúncia* apresenta uma curva de característica próxima dos itens *adequação lexical* e *adequação gramatical*. A *pronúncia* é um item que discrimina melhor os examinandos cujo valor da proficiência é acima de 2 pontos na escala, aproximadamente. Assim como a categoria 2 do item *adequação lexical*, a categoria 2 do item *pronúncia* apresenta a menor distância entre os pontos de interseção, pois a categoria também está espremida entre outras categorias. A faixa *intermediário* está pouco definida para os itens *adequação lexical* e *pronúncia*. Ou, dizendo de outra forma, o intervalo da categoria de resposta é de 2.2 pontos aproximadamente e, ainda que esteja próximo à sugestão de Eckes (2015) de que o intervalo tenha que ser menor que 5 e maior que 1.4 pontos, em comparação com o padrão de intervalos entre os pontos de interseção das curvas a categoria 2 ou nota 3 (*intermediário superior)* parece estar espremida entre as outras curvas. DeMars (2010) afirma que uma hipótese possível para explicar o fato de alguma categoria estar espremida entre duas outras é a de que os descritores se refiram a algo que não é muito frequente na prática. No caso da pronúncia, parece provável que os avaliadores não façam muita distinção entre uma pronúncia “com interferência frequentes da língua materna”, “com

inadequações" e "com algumas inadequações", descritores das notas 2, 3 e 4 respectivamente.

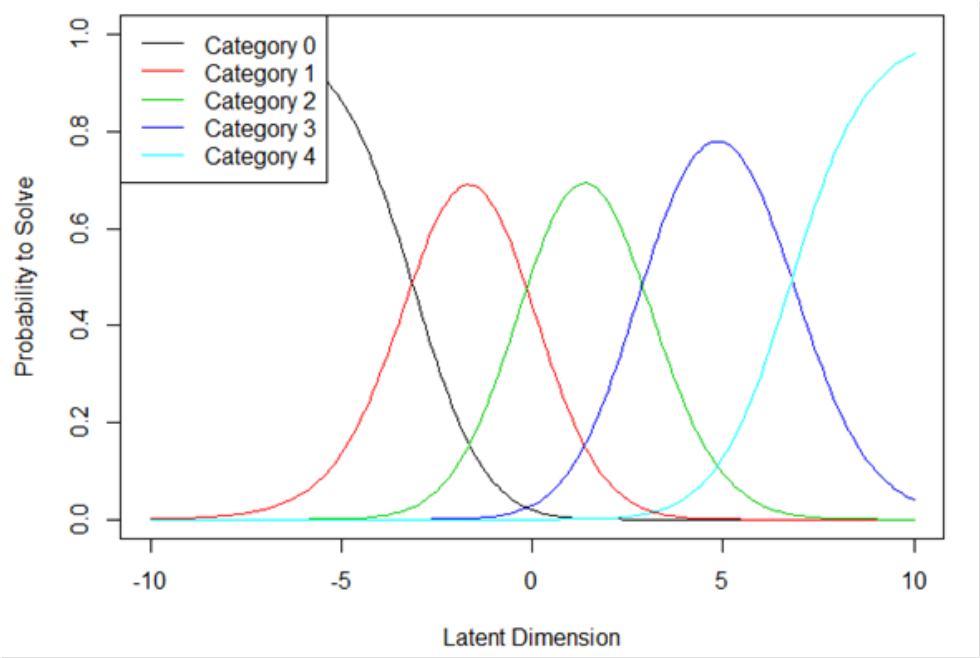
5.3.7. CURVA DE CARACTERÍSTICA DO ITEM *NOTA DO ENTREVISTADOR*

O aspecto da curva de característica do item *nota do entrevistador*, no que diz respeito à sua relação com o eixo dos perfis de habilidade estimado pelo modelo, representados no eixo horizontal, destoa de todos os itens analíticos. Cabe ressaltar, mais uma vez, que estou analisando a *nota do entrevistador* como se ela fosse um item e estou comparando-a com as demais notas ou itens avaliados pelo observador. Nesse sentido, não farei uma comparação da avaliação que o observador faz como um todo com a avaliação que o entrevistador também faz. O objetivo é comparar a nota atribuída a partir da escala do entrevistador, que é única, como em cada uma das escalas que correspondem a cada um dos itens ou parâmetros de avaliação da escala analítica.

O aspecto geral do gráfico 15 é mais homogêneo ou equilibrado porque a distância entre os pontos de interseção entre as categorias do meio da escala são mais uniformes, quando comparados ao padrão da maioria dos itens analíticos. O padrão mais ou menos regular de espaçamento entre as interseções das curvas representa uma capacidade de discriminação mais eficiente em um número maior de perfis de examinandos, representados no eixo horizontal, quando comparado ao padrão de espaçamento dos itens analíticos. O valor do intervalo de categoria de resposta do *intermediário* para *intermediário superior* é de 3 pontos; do *intermediário superior* para o *avançado* é de 2.9 pontos e do *avançado* para *avançado superior* é de 4 pontos. Os valores estão adequados, segundo as diretrizes de Eckes (2015), e apontam para uma maior capacidade de diferenciação entre as faixas *avançado* e *avançado superior,* assim como nos itens *adequação lexical, adequação gramatical* e *pronúncia.* A nota do entrevistador também discrimina muito bem os examinandos de proficiência alta, acima de 6.8 pontos da escala.

De acordo com Eckes (2015) a escala que distingue diferentes níveis de proficiência apresentam intervalos semelhantes que separam uma categoria e outra; por este motivo, o item *nota do entrevistador* é a escala mais eficiente em termos de discriminação de diferentes perfis de habilidade quando comparados a outros itens analíticos isolados.

Gráfico 15 - Curva de característica do item *Nota do entrevistador*



5.3.8. FUNÇÃO DA INFORMAÇÃO

A função da informação permite avaliar como cada item contribui para a qualidade do teste como um todo. A informação é função dos perfis de habilidade estimados empiricamente, e dos itens, que são representado pelas linhas no gráfico. A função da informação nos permite comparar os itens e também mostrar como cada item pode ajudar a discriminar os examinandos em diversas faixas de proficiência.

No gráfico 16, exposto a seguir, está representada a quantidade de informação que cada item agrega ao teste. No eixo horizontal está a escala de habilidades e no eixo vertical estão os valores da informação, sendo que, quanto mais alto o valor da informação mais confiáveis são as informações que o item fornece em função de determinados perfis de habilidade ou dimensão latente. DeMars (2010) afirma que na perspectiva da TRI, a confiabilidade está relacionada à eficiência de discriminação da escala. Ou seja, quanto mais informação tem um item em diferentes perfis de examinandos, mais confiável é a escala. A partir da análise é possível inferir que a nota 5 do item *compreensão* é menos confiável que a nota 5 atribuída a partir da escala do entrevistador ou da *adequação lexical*, por exemplo.

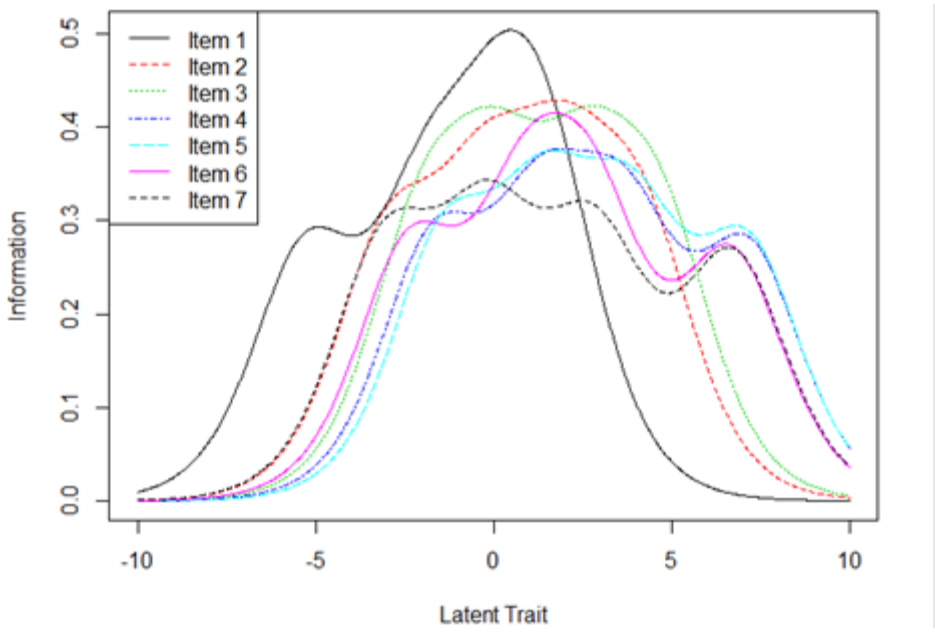
No gráfico 16, o item 1 se refere à *compreensão*, o 2 à *competência interacional*, o 3 à *fluência*, o 4 à *adequação lexical,* o 5 à *adequação gramatical*, o 6 à *pronúncia* e o 7 à *nota do entrevistador.*

O item *compreensão* perde muita informação após o valor próximo de 1 na escala de proficiência, significando que, após este valor, as informações que o item fornece podem não ser tão confiáveis quando comparadas àquelas que o mesmo item fornece quando se refere à valores de proficiência entre -2 e 0. Os itens *competência interacional* e *fluência* têm valores semelhantes. Por volta dos valores -2 e 5 é quando a quantidade de informação é mais elevada, ou seja, são itens que fornecem informações confiáveis ou que discriminam melhor examinandos entre esses valores de perfis de habilidade dos examinandos. As curvas que correspondem aos itens a*dequação lexical* e *adequação gramatical* quase se sobrepõem, ou seja, geram informações muito próximas ou contribuem de forma semelhante para o teste como um todo.

O item *pronúncia* gera mais informação ao teste quando se refere aos examinandos próximos ao ponto 2 da escala de proficiência. O item discrimina também, ainda que com menor eficiência, examinandos que correspondam à -2 pontos na escala de proficiência.

A *nota do entrevistador* apresenta valores altos de informação em vários valores na escala de proficiência; isso significa que a escala é mais confiável em diferentes faixas de proficiência, ou seja, o item diferencia variados níveis de proficiência.

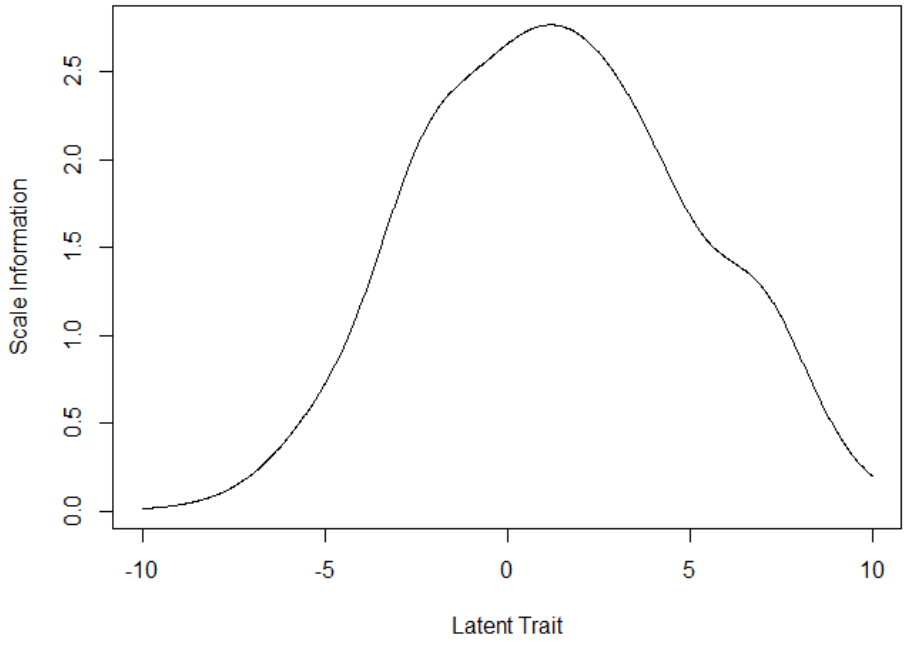
Gráfico 16 - Curva de informação dos itens



Os itens a*dequação lexical* e *adequação gramatical, pronúncia* e *nota do entrevistador* têm um comportamento semelhante quanto ao aspecto da informação e estão relacionados à discriminação de examinandos. Os itens apresentam valores similares e altos de informação por volta do valor 7 na escala de habilidade, ou seja, *adequação lexical* e *gramatical* são os parâmetros que mais geram informações confiáveis sobre examinandos cuja proficiência oral é alta.

No gráfico 17 que apresento a seguir, é apresentada a função de informação do teste como um todo. A partir do aspecto da função, nota-se que a totalidade dos itens concentra mais informações no meio da escala entre os valores -2 e 2 na escala de habilidade ou variável latente (*Latent trait*). O teste perde muita informação a partir do valor 2 na escala horizontal. Dizendo de outra forma, o teste diferencia bem examinandos que se concentram nos valores -2 e 2 da escala da habilidade. De forma geral, tais valores podem estar relacionados às faixas *intermediário* e *intermediário superior*. Se a confiabilidade de uma escala de parâmetros de avaliação está associada a sua capacidade de discriminação ao longo das faixas, é possível concluir por meio da análise, que a nota da prova oral diferencia bem ou é confiável para as faixas de certificação medianas como as de *intermediário e intermediário superior,* pois é a faixa de valores no eixo horizontal na qual os itens geram mais informações.

Gráfico 17 - Curva de informação do teste



Após o pico da curva de informação, observa-se uma diminuição quase uniformemente íngreme; só não o é porque próximo ao valor 6 na escala da habilidade há um leve degrau. Este aspecto da curva, da descida após o pico, refere-se à contribuição dos itens a*dequação lexical, adequação gramatical, pronúncia* e *nota do entrevistador* ao diferenciar examinandos com perfis altos de habilidade das faixas *avançado* e *avançado superior.* No entanto, ao somarmos todos os esforços para a diferenciação dessas faixas, ou seja, na sumarização de como o conjunto de itens diferenciam variados perfis de examinandos, as informações mais confiáveis que o teste gera está relacionada a examinandos de proficiência mediana, deixando a desejar ou não gerando a mesma quantidade de informação a respeito de examinandos de níveis mais altos. Como o corpus desta pesquisa é composto por notas elevadas, é provável que os examinandos do Celpe-Bras como um todo tenham a mesma característica. Essa constatação sugere como encaminhamento de mudança que se invista em mais itens que tenham um comportamento de discriminação que diferencie examinandos entre as faixas *avançado* e *avançado superior* como fazem os parâmetros da a*dequação lexical, adequação gramatical, pronúncia* e *nota do entrevistador.* Outra maneira de calibrar o exame seria reestruturar as descrições da escala dos demais itens ou revisar a situação ou tarefa da prova oral de forma que a atribuição de nota 5 em *compreensão* possa ser comparável ou signifique algo semelhante a uma nota 5 em *adequação lexical*, por exemplo.

5.3.9. MAPA

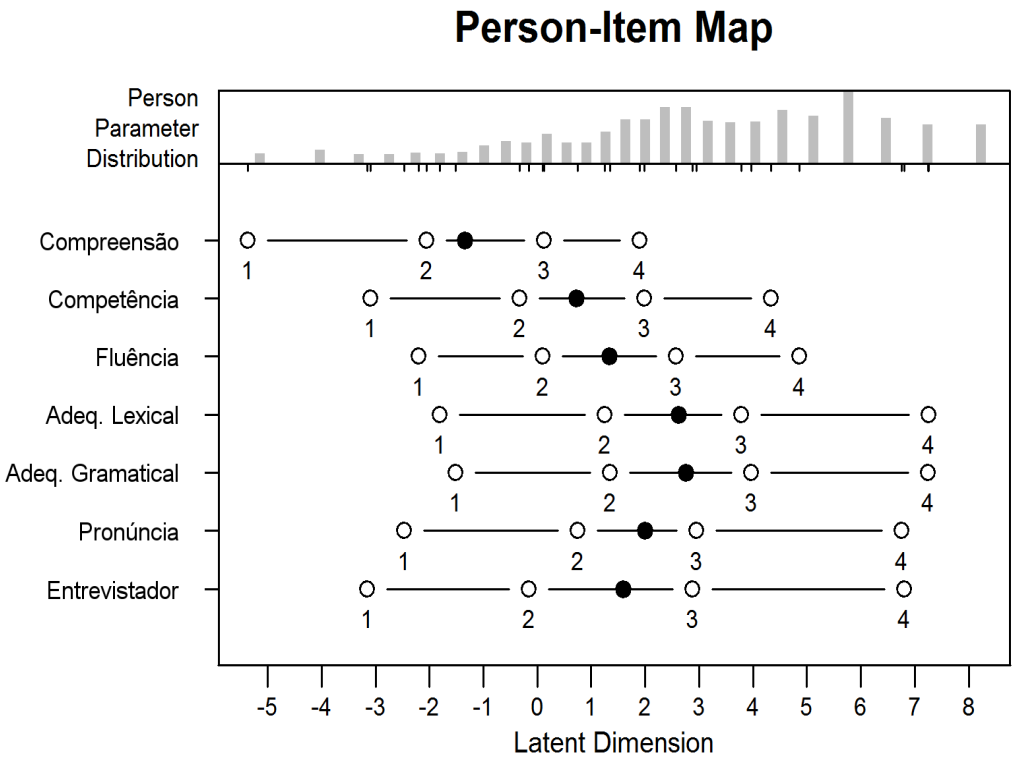
Uma outra maneira de analisar como os itens contribuem com o teste como um todo é analisando o mapa que sumariza todas as informações das curvas de característica dos itens. O mapa organiza os perfis de habilidades dos examinando e as características de cada um dos itens. Portanto, é possível observar que a distribuição dos examinandos na métrica da habilidade, ou variável latente, concentra-se entre os valores de 0 a 5, no topo do mapa. No eixo y, os números abaixo das bolinhas brancas representam a categoria; as bolinhas brancas representam o valor de *threshold* e é a partir das distâncias entre estes valores que se pode avaliar a capacidade de discriminação das categorias em função dos valores de proficiência.

O gráfico 18 apresenta um resumo de todas as análises do Rash, com a relação entre proficiência dos examinandos em cada item, de maneira que é possível visualmente comparar todas as informações até agora apresentadas em gráficos separadamente. Na linha horizontal, no topo do mapa, está representada a distribuição

dos examinandos a partir do cálculo da habilidade que é estimado pelo modelo. No eixo y estão organizados todos os sete parâmetros de avaliação e cada uma das linhas se refere à informação de cada item que, por sua vez, está em função da escala de habilidade ou a métrica da variável latente, representada no eixo x. Nas linhas horizontais no interior do mapa, que se referem aos parâmetros de avaliação, os números 1, 2, 3 e 4 representam as categorias, ou seja, as faixas de certificação ou notas das escala da prova oral do exame, e a distância entre elas representa o intervalo entre os pontos de *threshold,* que nos gráficos de curva de característica do item se referem à interseção entre as curvas.

Comparando os aspectos de cada item no mapa, novamente visualiza-se que o item *compreensão* é o mais fácil do exame, uma vez que é provável que os examinandos obtenham uma nota 5 neste item enquanto obteriam nota 3 em outros itens.

Gráfico 18 - Mapa do Rasch



Competência interacional e *fluência* são itens mais difíceis para os examinandos se comparados à *compreensão,* e apresentam um comportamento semelhante. De forma geral, os itens cujas linhas estão representadas no mapa mais à direita são os mais difíceis ou os que discriminam melhor os examinandos com proficiência mais

alta. Novamente é possível concluir que os itens *adequação lexical, adequação gramatical* e *pronúncia* conseguem distinguir melhor as categorias 3 e 4, *avançado* e *avançado superior*, respectivamente. É importante ressaltar que esses três itens têm um comportamento semelhante. Um pouco menos eficiente para distinguir as categorias 2 e 3 é o item *pronúncia.* A curta distância entre as bolinhas brancas em 2 e 3 reforçam esta interpretação, que já havia apresentado na análise da curva de característica do item *pronúncia.* Mais uma vez, nota-se o aspecto do item *nota do entrevistador* como o mais bem distribuído por toda escala, pois a escala contempla a faixa mais ampla de proficiência dos examinandos e a distância entre as bolinhas brancas, ou seja, intervalos ou valores de *threshold* são semelhantes.

5.4. DISCUSSÃO DOS RESULTADOS

Retomando a discussão sobre a finalidade da análise fatorial, é consenso entre os estatísticos que o cálculo fatorial esteja relacionado a perguntas que têm como objetivo investigar o construto a ser medido por algum instrumento. Neste sentido, a análise fatorial aqui apresentada foi eficiente ao informar o que o teste mede e o quanto cada parâmetro de avaliação está correspondendo à medida, ou seja, o quanto cada item contribui para composição da nota final.

Com base na análise fatorial, tanto a grade analítica quanto a do entrevistador medem apenas um construto, que é o da proficiência oral. Isso quer dizer que o instrumento é unidimensional, que mede apenas uma coisa – a proficiência oral (e não outras proficiências). Ou seja, a grade é válida do ponto de vista do construto que ela pretende avaliar. Não estou afirmando que a prova oral não esteja avaliando outros aspectos, que provavelmente interagem com os parâmetros da escala como a simpatia, cordialidade, comportamento do entrevistador, etc. A afirmação que faço, a partir do resultado da análise fatorial, é a de que o construto que se sobressai na avaliação por meio da atribuição das notas aos parâmetros das grades da prova oral é o da proficiência oral.

Quanto à relação entre os parâmetros de avaliação, a análise fatorial mostrou evidência empírica de forte inter-relação entre as variáveis observáveis, que são as notas atribuídas a partir da grade analítica e a nota do entrevistador.

Retomando a discussão sobre a contribuição deste tipo de análise para a compreensão dos parâmetros de avaliação da proficiência oral, no contexto da análise da avaliação do Celpe-Bras, fiz uma análise considerando a composição da nota

atribuída pelo observador a partir das seis notas dos parâmetros analíticos, e uma análise acrescentando a nota do entrevistador aos seis critérios analíticos para avaliar a composição da nota final da prova oral. Ao acrescentar a nota do entrevistador, ignorei o peso de 50% que cada um dos avaliadores têm na composição atual da nota final.

Na primeira análise, cujo foco é o peso de cada parâmetro analítico na composição da nota do observador, quanto ao recálculo do peso dos parâmetros na composição da nota do observador, temos que ela foi 70% composta pelos parâmetros da *adequação lexical, fluência e adequação gramatical.*

Na segunda análise, ao acrescentar o item *nota do entrevistador* e compará-lo aos itens analíticos, a análise sugere que a nota do entrevistador deva ter um peso de aproximadamente 33% na composição da nota final da prova oral, e não 50%, como hoje vigora. Empiricamente a nota do observador é mais importante do que a nota do entrevistador, porque ao somarmos os pesos dos parâmetros analíticos na composição da nota final, temos aproximadamente 66%, ou seja, mais de 50% da composição da nota final sendo explicada a partir da nota atribuída pelo avaliador-observador.

Ao aplicar os novos pesos ao conjunto de dados, inicialmente supus que a quantidade de examinandos classificados nas faixas *avançado superior* aumentaria, porque mais pesos estariam sendo atribuídos aos parâmetros que se referiam à *adequação lexical, adequação gramatical* e *fluência* na composição da nota do observador. No entanto, ao diminuir o peso da *nota do entrevistador* na composição da nota final da prova oral, muitos examinandos migraram da faixa *avançado* para *avançado superior.* Ou seja, com a nova proposta de peso, as notas entre os examinandos das faixas avançado provavelmente aumentariam, agravando ainda mais o problema detectado na análise TRI que é o fato do modelo de atribuição de notas do exame não ser eficiente ao diferenciar os dois níveis mais altos de certificação do exame.

O resultado da aplicação dos novos pesos demonstrou que a *nota do entrevistador* foi provavelmente eficiente para discriminar entre as faixas *avançado* e *avançado superior*. Tal interpretação sobre a *nota do entrevistador* foi confirmada na análise da curva de resposta ao item, que demonstrou eficiência ao discriminar examinandos de diferentes perfis de habilidade. Por isso, diminuir o peso da *nota do entrevistador* na composição da nota final não seria uma boa estratégia, porque a avaliação de discriminação do exame como um todo aponta que o teste carece de itens que

diferenciem examinandos classificados nas faixas *intermediário superior*, *avançado* e *avançado superior.*

Além disso, a análise Rasch sobre a eficiência de discriminação dos itens da grade do observador indica a necessidade de investimento na diferenciação entre as faixas *avançado* e *avançado superior,* aspecto que está melhor resolvido na grade do entrevistador. Dessa forma, diminuir o peso da nota do avaliador-interlocutor, que é mais eficiênte quanto à diferenciação deste perfil de examinandos, e dar mais peso na nota do observador, que necessita de revisão justamente para diferenciar melhor os examinandos de níveis mais altos, não parece ser uma boa solução. A proposta de aumento de peso da nota do observador seria possível se antes houvesse uma revisão dos parâmetros, das suas relações com a tarefa e com os descritores. Por isso, sem calibrar a nota do observador, os pesos propostos pela análise fatorial com relação à porcentagem de contribuição de cada uma das duas notas dos avaliadores na composição da nota final da prova oral é desaconselhável; no entanto, os pesos propostos para a composição da nota do observador é uma proposta a ser considerada.

A nota relacionada ao parâmetro *compreensão* demonstrou indícios que subsidiam a necessidade de repensar tanto o valor do peso na composição da nota quanto a maneira como a escala está discriminando os examinandos. Os resultados da análise fatorial sugerem que a *compreensão* é o parâmetro que menos explica a nota final na prova oral, em comparação com os outros. Embora a *compreensão* esteja fortemente relacionada com o construto da proficiência oral, conforme descrito na matriz de correlação, a análise fatorial mostrou um baixo valor de peso referente à nota de *compreensão* na composição tanto da nota analítica quanto da nota final da prova oral. Uma possível explicação para que o parâmetro tenha tido esse valor tanto pode ser pautada pela situação de avaliação e da tarefa quanto pela maneira como o parâmetro está sendo descrito na grade. Por se tratar de uma situação de entrevista oral em que a relação é assimétrica, ou seja, a interação é guiada ou controlada, e que as tarefas do exame provavelmente envolvam responder algumas perguntas sobre o Elemento Provocador na maior parte do tempo, pode ser que a avaliação oral proposta na situação do exame ofereça poucas oportunidades para que a compreensão oral seja avaliada com o mesmo detalhamento e grau de complexidade que o uso de estruturas linguísticas, por exemplo. O entrevistador interage com o examinando ao fazer algumas perguntas e ao incentivá-lo a falar; o examinando não tem que necessariamente compreender uma exposição oral mais complexa por parte do entrevistador. No contexto do exame, o que se espera de um bom entrevistador é que ele dê o máximo de oportunidades para que o examinando fale, desenvolva ideias

oralmente e não necessariamente que o examinando compreenda as ideias e as opiniões do entrevistador. Se a tarefa do exame tivesse como propósito a compreensão de turnos de fala mais longos, talvez a avaliação da compreensão fosse mais complexa ou exigiria mais habilidade do examinando como o parâmetro *adequação lexical* exige, conforme demonstrei empiricamente, por exemplo. Neste contexto de avaliação, o papel do entrevistador estaria mais próximo a de um ouvinte. Por este motivo ter notas ou ser classificado nas notas 4 e 5 porque compreende o fluxo natural da fala, conforme descrito na grade, parece ser algo que acontece com muitos examinandos, uma vez que compreender algumas breves perguntas feitas pelo entrevistador é uma habilidade que iniciantes na língua ou até mesmo pessoas que falam línguas próximas como os falantes de espanhol, possam conseguir demonstrar nesta situação de prova, mesmo que não demonstrem, por exemplo, pronúncia adequada, uso variado e amplo de estruturas gramaticais e lexicais – ou seja, nota 4 ou 5 nos demais parâmetros avaliados.

A análise da curva de característica do item *compreensão* trouxe ainda mais evidências de que a maneira como a compreensão oral está sendo avaliada precisa ser revista. Na análise Rasch, o item *compreensão* se mostrou eficiente para fornecer informações ou discriminar os examinandos de baixa proficiência, que correspondem às faixas *sem certificação ou básico* e *intermediário*, e se mostrou ineficiente para discriminar as faixas acima de *intermediário*. Conforme já discutido, uma boa escala de item deve ser eficiente para discriminar todas as faixas previstas pelo exame. Retomando a discussão de Messick (1987) sobre validade de construto, o objetivo do trabalho foi também o de confrontar as informações sobre a estrutura da EPO com sua consistência na atribuição de notas para verificar se algum parâmetro de avaliação das grades apresenta propriedades empíricas insuficientes, que podem estar distorcendo a representatividade do construto a ser medido. Considero que a nota de *compreensão* está distorcendo a representatividade do construto e neste caso, como sugere Messick (1987), pode ser recomendável a sua reposição por um item que esteja em consonância com as especificações do teste.

Outro ponto que merece destaque na análise fatorial e que foi mais detalhadamente esclarecido na análise realizada com base na TRI é a alta correlação entre *adequação gramatical* e *adequação lexical,* sugerindo inicialmente que os dois parâmetros pudessem estar medindo a mesma dimensão do construto. No entanto, quando analisamos os valores da análise fatorial e as curvas de característica destes itens é possível verificar que o peso de *adequação lexical* é maior quando comparado à *adequação gramatical*, o que sugere, de alguma forma, que embora o conhecimento lexical e gramatical possam ser similares ou estar sobrepostos, do ponto de vista

teórico, ao operacionalizá-los na grade, os parâmetros podem não gerar necessariamente as mesmas informações, e por isso eles teriam pesos distintos. A partir da análise de discriminação dos itens, é possível afirmar que ambos discriminam faixas de proficiências altas previstas pelo exame, embora o façam de formas distintas. Retomando a discussão sobre outras grades de avaliação oral, Fulcher (2003) resume que as críticas da grade do exame do FSI se voltavam ao peso maior dado aos parâmetros lexicais e gramaticais. A análise empírica dos dados da prova do Celpe-Bras sugere uma explicação para o elevado peso dado a estes critérios: as adequações gramatical e lexical são dimensões da proficiência oral que, quando operacionalizadas em uma grade, podem ser eficientes para discriminar examinandos de alta proficiência, como demonstrado na análise da grade do Celpe-Bras. É importante destacar que em ambos os exames – Celpe-Bras e FSI – a metodologia adotada é a da EPO. Ao reafirmar a importância dos parâmetros lexical e gramatical no contexto da avaliação oral, não estou defendendo que estes sejam os aspectos mais importantes da avaliação oral, mas que são eficientes para discriminar examinandos de alta proficiência, assim como o parâmetro da fluência. Como a grade se propõe a avaliar e distinguir níveis como o *avançado superior* do *avançado*, é preciso investir na revisão da grade do avaliador-observador de forma a acrescentar parâmetros com comportamento de discriminação semelhantes ao das adequações lexical e gramatical e da fluência.

A análise fatorial de *fluência* também merece destaque. Segundo Fulcher (2003), a fluência está relacionada à fluidez, à automaticidade e ao impacto na compreensão do que está sendo dito. A análise aponta para a necessidade de aumento do peso deste parâmetro na composição da nota do observador. Metodologicamente, as tarefas da prova oral do Celpe-Bras fazem com que os turnos de fala se concentrem na fala no examinando. No entanto, a *compreensão* e a *fluência* têm o mesmo peso na nota final. Retomando Fulcher (2003), a grade e a tarefa são faces de uma mesma moeda, e por isso a grade deve ser pensada, levando em conta a tarefa. O autor ressalta ainda que os parâmetros que fazem parte do construto da proficiência oral devem ser operacionalmente avaliáveis. Os dados analisados do Celpe-Bras indicam que é preciso relacionar melhor o desenho da prova e o peso dos parâmetros da *compreensão* e da *fluência*, principalmente, uma vez que o desenho da prova gera mais insumos para a avaliação da fluência do que para a avaliação da compreensão, embora o peso atual na nota final do observador dos parâmetros seja o mesmo. Isso ocorre porque é esperado que o examinando compreenda os turnos de fala do entrevistador, ao passo que o parâmetro da fluência, por exemplo, pode ser

avaliado com base na maioria do tempo da interação, pois espera-se que o tempo de fala do examinando seja maior do que o do entrevistador.

Eckes (2015) afirma que, dentre outros aspectos, os parâmetros de avaliação e a maneira como seus descritores estão graduados pelas faixas de certificação podem interferir sistematicamente nas notas, ou seja, podem ser fonte de erro, comprometendo a confiabilidade da nota. Para o autor, os itens podem ser descritos de forma que seja pouco provável que os examinandos atinjam uma determinada nota ao ser avaliado a partir de um determinado parâmetro. Um exemplo que podemos citar do problema colocado por Eckes (2015), a partir dos dados empíricos, é a nota de *pronúncia* e *adequação lexical* na faixa *intermediário superior.* Na curva de característica destes itens, o intervalo de *threshold* se mostrou estreito para esta faixa. É provável que os descritores de *pronúncia* e *adequação lexical* estejam organizados de forma a diminuir a probabilidade do examinando ser classificado na faixa *intermediário superior*.

De maneira geral, a análise empírica das escalas de proficiência utilizadas para avaliação oral teve como pergunta de fundo a confiabilidade da nota que, no contexto deste trabalho, significa a capacidade de discriminação das faixas de proficiência. Ao avaliar separadamente cada item, a análise aponta que as notas de *compreensão* são mais confiáveis ou eficientes quando se trata de examinandos com perfis baixo de certificação entre as faixas *sem certificação, básico* e *intermediário*. As notas medianas entre as faixas *intermediário e intermediário superior* são eficientes quando atribuídas aos itens *competência interacional* e *fluência*. Notas altas ou entre as faixas avançado e avançado superior são confiáveis quando se trata da avaliação dos aspectos da *adequação lexical, adequação gramatical* e *pronúncia.* A *nota do entrevistador* se mostrou eficiente ou confiável para todas as faixas de certificação previstas no exame.

A partir da análise de discriminação dos itens, os resultados empíricos sugerem que, de forma geral, a nota da prova oral seja mais confiável quando diz respeito às categorias medianas na escala (*intermediário e intermediário superior)* do que às demais categorias, porque os itens se mostraram mais eficientes para discriminar estas faixas.

6. CONSIDERAÇÕES FINAIS

O presente trabalho teve como objeto de pesquisa a proposta de atribuição de notas dadas ao desempenho oral de examinandos, no contexto do exame do Celpe-Bras. Por meio da discussão sobre as características do exame e as origens deste tipo de proposta de avaliação nos contextos do ensino e da avaliação de línguas adicionais, argumentei que a interação face a face da prova oral é uma tarefa cuja situação é a entrevista de proficiência oral. Quanto aos debates em torno da validade da situação de entrevista de proficiência oral, considerei que são potencialmente válidos os resultados de exames de proficiência oral que ofereçam oportunidade para o examinando demonstrar sua capacidade de interagir oralmente em Língua Portuguesa, como previsto na tarefa do Celpe-Bras. No entanto, ponderei que, para afirmar algo sobre a validade das inferências a serem feitas a partir das notas, seria preciso analisar os resultados do teste.

A pesquisa, então, teve como objetivo a análise da nota para responder a duas perguntas, sendo a primeira sobre o quanto cada uma das notas representa cada um dos construtos operacionalizados pelos parâmetros de avaliação. A segunda questão de pesquisa buscou avaliar como cada nota em cada uma das faixas de proficiência contribui para gerar informações confiáveis sobre a certificação do examinando, por meio da análise dos intervalos entre as faixas que visa avaliar a capacidade de discriminação de cada item.

Neste trabalho, fundamentei-me na proposta de Messick (1987), para o qual a validade é um conceito único com diferentes aspectos. O autor afirma que a validade de construto catalisa a discussão sobre os demais aspectos da validade. A validade de construto consiste na investigação da relação nota-construto. Para o autor, a validação implica dar ênfase na nota e não no instrumento, porque é por meio da análise dos resultados que é possível verificar propriedades de validade e confiabilidade da medida.

Para analisar as qualidades psicométricas dos instrumentos de avaliação da prova oral do Celpe-Bras, apresentei uma análise fatorial e uma análise de discriminação de itens, por meio do modelo matemático Rasch. A análise empírica das escalas de proficiência utilizadas para avaliação oral teve como pergunta de fundo a validade de construto do exame.

Assim sendo, apresentei algumas evidências sobre o significado da nota. Na análise de correlação entre as notas, os resultados apontam para uma correlação linear entre as variáveis *nota do avaliador-interlocutor*, nota de *compreensão*, nota de *fluência* etc. Isso significa que as notas estão avaliando um mesmo construto. Por meio da análise fatorial, demonstrou-se empiricamente que o significado da nota oral está diretamente relacionado ao construto da proficiência oral; ou seja, a nota total da prova é válida do ponto de vista do construto que se pretende avaliar.

Quanto aos pesos de cada uma das seis notas que compõem a nota final analítica atribuída pelo *avaliador-observador*, os resultados da análise fatorial indicam a necessidade de revisão de pesos para sua composição. De maneira geral, na nova proposta, o parâmetro da *compreensão* deve ter seu peso diminuído e o parâmetro da *fluência,* da *adequação lexical* e da *adequação gramatical,* aumentados. Na análise do peso de cada parâmetro analítico na composição da nota do observador, 70% da nota analítica é composta pelos parâmetros da *adequação lexical, fluência* e *adequação gramatical*. Os outros 30% da nota analítica é composto pelos parâmetros *compreensão, competência interacional* e *pronúncia*. Houve uma diminuição de 16% para 6% na contribuição para a nota analítica que se refere ao parâmetro *compreensão.*

Ao avaliar o quanto cada avaliador contribui para a composição da nota final da prova oral, a análise fatorial indica que a nota do avaliador-observador é a que deveria ter mais peso, 66.34% com relação a nota única do avaliador-interlocutor que deveria ter peso de 33.67% na composição da nota da prova oral. No entanto, ao aplicar os novos pesos nos deparamos com outras questões. Como a grade do *avaliador-observador* carece de revisão para que os examinandos dos níveis *avançado superior*

e *avançado* sejam discriminados com eficiência, com o aumento do peso da nota analítica e diminuição da nota holística, a questão se torna ainda mais problemática porque a nota do avaliador-interlocutor apresentou-se como mais eficiente para discriminar examinandos entre as faixas *avançado* e *avançado-superior.* Por este motivo, antes de se efetivar a mudança dos pesos dos parâmetros, a mudança de peso da contribuição das notas dos dois avaliadores na composição da nota final proposta pela análise empírica é desaconselhável. O item *nota do interlocutor* se mostrou o mais eficiente para discriminar examinandos de diferentes perfis de habilidade, quando comparado aos outros seis que compõem a nota analítica, por isso diminuir seu peso na composição da nota poderia comprometer ainda mais a confiabilidade do exame. É preciso, paralelamente, rever os descritores da escala do avaliador-observador, com especial atenção para o parâmetro *compreensão* que, hoje, encontra-se intimamente relacionado ao papel do interlocutor.

Conforme discutido anteriormente, poucas são as oportunidades de se avaliar a compreensão oral do examinando na situação da entrevista oral da forma como está sendo estruturada atualmente, porque espera-se que o examinando fale a maior parte do tempo, sem necessariamente demonstrar uma compreensão oral de turnos de fala extensos do avaliador-interlocutor. A partir dos resultados das análises, argumentei que a situação de prova desafia mais o examinando quanto ao seu vocabulário do que quanto à compreensão oral. Recentemente, o nome *avaliador-entrevistador* no exame Celpe-Bras foi substituído por *avaliador-interlocutor.* No entanto, parece que há uma necessidade de repensar não só o nome, mas o papel do *avaliador-interlocutor* na interação e sua relação com a avaliação da *compreensão oral,* caso se decida pela manutenção da operacionalização do construto da compreensão oral na grade analítica.

Messick (1987) afirma que a validade de construto pode ter a finalidade de confrontar as informações sobre a estrutura da prova com sua consistência na atribuição de notas para poder avaliar se algum parâmetro de avaliação das grades apresenta propriedades empíricas insuficientes, que podem distorcer a representatividade do construto medido. Considero que *compreensão* tenha apresentado propriedades empíricas pouco representativas do construto operacionalizado na situação de prova oral do Celpe-Bras. Tais propriedades empíricas apresentadas na presente tese podem ser constadas principalmente pelos valores de comunalidade da análise fatorial e do *outfit* na análise Rasch que se referem ao item *compreensão*. Messick (1987) sugere, neste caso, a sua substituição por outro em consonância com as especificações do teste e, por isso, aponto para a

necessidade de se revisar este aspecto da avaliação oral, conectando-o com a situação de prova e com o papel do *avaliador-interlocutor*.

Com relação ao potencial discriminatório de cada parâmetro de avaliação da prova oral, o pressuposto é o de que a confiabilidade de uma escala esteja associada a sua capacidade de discriminação ao longo das faixas de certificação. Historicamente, os descritores de grades orais foram sendo elaborados de forma que pouco discriminam um nível de outro (FULCHER, 2003). É importante lembrar que o desafio de descrever verbalmente desempenhos orais e organizá-los em uma escala é extremamente complexo e vai além da descrição verbal do desempenho, e por isso é necessário avaliar empiricamente os níveis de descrição pelo estudo da nota. Messick (1987) afirma que o modelo composto – quando se tem uma nota em número que corresponde a uma faixa de proficiência – é desafiador para o processo de validação de construto porque é preciso investigar não só o significado da nota em si, no caso da nota final da prova oral, mas também a validade da nota de corte de cada uma das faixas de certificação.

A partir da análise de discriminação dos itens, os resultados empíricos sugerem que a nota da prova oral seja mais confiável quando diz respeito às faixas medianas na escala, como as *intermediário* e *intermediário superior*, do que das demais faixas, pois são as faixas de proficiência nas quais os itens geram mais informações, ou dito de outra forma, são os itens que discriminam melhor os examinandos.

Por meio da análise TRI, a nota 5 de *compreensão*, por exemplo, não está relacionada com a nota *avançado superior* dos demais parâmetros. Além disso, a nota 3 – que corresponde à faixa *intermediário superior* – de *pronúncia* e *adequação lexical* demonstra empiricamente a necessidade de mais investimento na discriminação das faixas com as quais fazem fronteira. A partir da análise de discriminação de itens, é possível afirmar que a grade do Celpe-Bras está organizada de forma que nem todos os parâmetros de avaliação refletem a mesma faixa de proficiência em todos os descritores. Dizendo de outra forma, ao considerar intervalos entre os valores de *threshold* dos parâmetros de avaliação, posso afirmar que uma mesma faixa de nota em cada um dos parâmetros pode não discriminar da mesma forma o mesmo perfil de examinandos.

A pesquisa se limitou à analisar empiricamente o significado das notas e apontar aspectos em que a relação nota-construto carecem de revisão e análise. Nesse sentido, a partir do resultado do presente trabalho, outras pesquisas podem dar sequência ao aprimoramento da grade oral do Celpe-Bras ao analisar, por exemplo, o que faz um examinando nas faixas *intermediário superior, avançado e avançado*

superior em todos os parâmetros de avaliação de forma a coletar elementos que subsidiem a reescrita ou revisão dos descritores. É também recomendável a realização de outra pesquisa que avalie empiricamente a relação nota-construto da nova proposta de redação de descritores, utilizando a análise fatorial e a TRI para que os resultados sejam comparados a esta pesquisa. Outra maneira de avaliar a qualidade das evidências apresentadas neste trabalho seria a de conduzir as mesmas análises com um conjunto de dados diferente, de modo que estejam controladas as notas por faixa de certificação. Como o corpus deste trabalho concentrou notas mais altas, poderia ser interessante replicar a mesma análise feita com outros dados, de maneira que uma mesma quantidade de notas para cada faixa de certificação forme o conjunto de notas.

Retomando a discussão de Messick (1987), o construto e a sua operacionalização estão entrelaçados, e por isso as evidências que fundamentam as notas embasam também o desenho do instrumento. Porém, se encontramos evidências de que as notas não fundamentam ou distorcem o construto, é preciso decidir se é o construto e/ou a medida que deverão ser revistas. Por meio das análises empíricas, proponho a mudança de peso na composição da nota analítica, bem como o investimento na revisão dos descritores, da tarefa e da situação de entrevista de proficiência oral, de forma que os parâmetros possam distinguir níveis como o *avançado superior* do *avançado* na grade do avaliador-observador.

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

REFERÊNCIAS

ARYADOUST, V.S.; GOH, C. A Rasch analysis of an international English language testing system listening sample test. IN: 3 REDESIGNING PEDAGOGY INTERNATIONAL CONFERENCE, Anais…. Singapore. 2009

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. Standards for educational and psychological testing. Nova Iorque: AERA, 2014.

BACHMAN, Lyle F. *Fundamental considerations in language testing*. Oxford: Oxford University Press. 1990.

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

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. *Certificado de Proficiência em Língua Portuguesa para Estrangeiros: Manual do Examinando*. Brasília, 2010.

_____. Ministério da Educação. Site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Apresenta informações sobre aplicação do exame Celpe-Bras*. Disponível em: <http://www.inep.gov.br/celpebras/estrutura_exame.asp>, acesso: 21 jan. 2017.

_____. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Portaria n.334, de 02 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.* Disponível em: <http://www.inep.gov.br/celpebras/, acesso: 15 jul. 2013.

_____. Ministério da Educação. Secretaria de Ensino Superior. *Certificado de Proficiência em Língua Portuguesa para Estrangeiros: Manual do Examinando*. Brasília, 2015.

_____. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital n. 1, de 28 de janeiro de 2016 - de*

abertura de inscrições do exame Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras/2016.1), 2016a. Disponível em: <http://download.inep.gov.br/outras_acoes/celpe_bras/legislacao/2016/edital_n1_de28012016_celpe_Bras_2016.1.pdf>, acesso em 04 set. 2017.

_____. Ministério da Educação. Secretaria de Ensino Superior. *Certificado de Proficiência em Língua Portuguesa para Estrangeiros: Elementos Provocadores*. Brasília, 2016b.

_____. Ministério da Educação. Secretaria de Ensino Superior. *Certificado de Proficiência em Língua Portuguesa para Estrangeiros: Roteiro de Interação*. Brasília, 2016c.

BROWN, H. Douglas; ABEYWICKRAMA, Priyanvada. Language Assessment: principles and classroom practice. Nova Iorque: Longman Pearson, 2010.

BROWN, Annie. *Interviewer variability in oral proficiency interviews*. Frankfurt: PeterLang. 2005.

BROWN, Thimothy A. *Confirmatory factor analysis for applied research*. Nova Iorque: Guilford Press, 1960/2015.

BYGATE, Martin. Teaching and testing speaking. In: LONG, Michael H.; DOUGHTY, Catherine J. *The handbook of language teaching*. Chichester: Wiley-Blackwell, 2011. 411-440 p.

CANTY, Angelo; RIPLEY, Brian. *Boot:* Bootstrap R (S-Plus) Functions. R package, versão 1.3-20. 2017.

CEBRASPE. Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos. Curso de formação de aplicadores do exame Celpe-Bras: fórum de discussão. Brasília, 2017.

COURA-SOBRINHO, Jerônimo**.** *O sistema de avaliação Celpe-Bras: o processo de correção e a certificação.* In: CONGRESSO INTERNACIONAL DE POLÍTICA LINGUÍSTICA NA AMÉRICA DO SUL, João Pessoa. Anais... João Pessoa. 2006.

COURA-SOBRINHO, Jerônimo e DELL'ISOLA. Regina Lúcia Péret. O contrato de comunicação na avaliação de proficiência em língua estrangeira. IN: JÚDICE, Norimar & DELL'ISOLA. Regina L. Péret. *Português-Língua Estrangeira:* novos diálogos. Niterói: Intertexto, 2009.

CHRISTÓFOLO, J.E.; GONÇALVES, A.M. Internacionalização do ensino superior na Colômbia. IN: MINISTÉRIO DAS RELAÇÕES EXTERIORES DO BRASIL. *Mundo afora*: políticas de internacionalização de universidades. Brasília: Ministério das Relações Exteriores, 2012. 88-97p.

DAVISON, A. C.; HINKLEY, D. V. *Bootstrap Methods and Their Applications.* Cambridge: Cambridge University Press, 1997.

DEMARS, Christine. *Item response theory:* understanding statistics measurement. Oxford: Oxford University press, 2010.

DINIZ, L.R.A. Política linguística do Estado brasileiro para a divulgação do português em países de língua oficial espanhola. *Trabalhos em Linguística Aplicada,* Campinas, vol.51, nº 2, jul/Dec. 2012a. Disponível em: <http://www.scielo.br/scielo.php?pid=S0103-18132012000200009&script=sci_arttext >, acesso em 06 ago. de 2013.

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

EUROPA. Quadro europeu comum de referência para as línguas: aprendizagem, ensino, avaliação. Porto: Asa Edições, 2001.

FERREIRA, Laura Márcia Luiza. *Habilidades de leitura na proposta de interação do exame Celpe-Bras.* Dissertação de Mestrado. UFMG, Belo Horizonte: UFMG, 2012.

FILHO, Dalson Brito Figueredo; JÚNIOR, José Alexandre da Silva. Visão além do alcance: uma introdução à análise fatorial. *Opinião pública*, v.16, n.1, jun.2010, 160-185p.

FULCHER, Glenn. *Testing second language speaking.* Londres: Routledge, 2003.

FULCHER, Gleenn; DAVIDSON, Fred. *Language testing and assessment:* an advanced resource book. Routledge: Nova Iorque, 2007, 91-114p.

FURTOSO, Viviane Aparecida Bagio. *Desempenho oral em português para falantes de outras línguas:* da avaliação à aprendizagem de línguas estrangeiras em contexto online. Tese de Doutorado. São José do Rio Preto: Unesp, 2011.

HAMBLETON, Ronald K.; SWAMINATHAN, H.; ROGERS, H. Jane. *Fundamentals of item response theory*. Califórnia: Sage Publications Inc. 1991.

HUGHES, Arthur. *Testing for language teachers*. Cambridge: Cambridge University Press, 1989.

JORDAO, Clarissa Menezes. EAL - ELF - EFL - EGL: quem dá conta? *Revista Brasileira de Linguística Aplicada (online)*, vol.14, n.1, pp.13-40. 2014 Disponível em: <http://dx.doi.org/10.1590/S1984-63982014000100002>, acesso em 10 de ago. 2017.

JOHNSON, Marysia*. The art of non-conversation*: a reexamination of the validity of the oral proficiency interview. Yale Haven & London: Yale University Press, 2001.

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

KUNNAN, A.J. An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing,* 9, 1992, 30-49p..

LONG, M.H. Methodological principles for language teaching. In: LONG, M.H.; DOUGHTY, C.J. *The handbook of language teaching*. Oxford: Wiley-Blackwell, 2011.

LUCKESI, Cipriano Carlos. *Avaliação da aprendizagem escolar: estudos e proposições*. 22ed. São Paulo: Editora Cortez, 2011.

MAIR, P.; HATZINGER, R.; MAIER M. J. *eRm*: Extended Rasch Modeling. Versão 0.16-0. 2018. Disponível em: <http://r-forge.r-project.org/projects/erm/>, acesso em 01 de mai. 2018.

McNAMARA, Tim. *Language Testing.* Oxford: Oxford University Press, 2000.

McNAMARA, Tim. Language Testing. In: DAVIES, Alan; ELDER, Catherine. *The handbook of applied linguistics.* 2004

McNAMARA, Tim; KNOCH, Ute. The Rasch wars: the emergence of Rasch

measurement in language learning. *Language Testing*. v. 29, 2012, 555–576p.

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

NIEDERAUER, Marcia. Competência interacional: critério para avaliação da produção oral em língua adicional. *Trab. linguist. apl.* [online]. 2014, vol.53, n.2, 2014, pp.403-424. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S010318132014000200008&lng=pt&nrm=iso>, acesso em 30 de mai. 2016.

NORRIS, John M. Task-based Teaching and Testing. In: LONG, Michael H.; DOUGHTY, Catherine J. *The handbook of language teaching.* Chichester: Wiley-Blackwell, 2011, 577-594 p.

Revelle, W. Psych: Procedures for Personality and Psychological Research, version 1.8.4. Evanston: Northwestern University, 2018. Disponível em: <https://CRAN.R-project.org/>acesso em mai. 2018.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena. 2018 Disponível em: <http://www.R-project.org/> acesso em mai. 2018.

SAKAMORI, L. *A atuação do entrevistador na interação face a face no exame Celpe-Bras*. Campinas: Unicamp, 2006. Dissertação de Mestrado, PPG em Linguística Aplicada, IEL, Unicamp, 2006.

SCARAMUCCI, Matilde V. Ricardi. O Projeto Celpe-Bras no Âmbito do Mercosul: contribuições para uma definição de proficiência comunicativa. In: ALMEIDA FILHO, J.C.P (Org.) *Português para Estrangeiros: interface com o espanhol*. 2.ed. Campinas: Pontes, 2001, 77-90 p.

_____. O exame Celpe-Bras em contexto hispanofalante: percepções de professores e candidatos. In: WIEDEMANN, L.; SCARAMUCCI, V.R. (Orgs.) *Português para Falantes de Espanhol.* 1.ed. Campinas: Pontes, 2008, 175-190 p.

SCARAMUCCI, Matilde V. Ricardi ; SCHLATTER, Margarete ; GARCEZ, P. M. O papel da interação na pesquisa sobre aquisição e uso de língua estrangeira: implicações para o ensino e para a avaliação. *Letras de Hoje*, PUCRS - Porto Alegre, v. 39, n.3, 2004. 345-378p.

SCARINO, Angela. Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, v. 30, n. 3, p. 309-317, 2013

SCHOFFEN, Juliana Roquele. *Avaliação de proficiência oral em língua estrangeira*: descrição dos níveis de candidatos falantes de espanhol no exame Celpe-Bras. Dissertação de Mestrado. Porto Alegre: UFRGS, 2003.

_____. *Gêneros do discurso e parâmetros de avaliação de proficiência em português como língua estrangeira no exame Celpe-Bras*. Tese de Doutorado. Porto Alegre: UFRGS, 2009.

SCHOFFEN, Juliana Roquele. *Introdução ao exame Celpe-Bras*. Comunicação apresentada durante o Evento de Correção de Instrumentos de Aplicação do Celpe-Bras, 08 a 13 de julho. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Brasília: 2013.

SHOHAMY, Elana. Assessment. IN: CELCE-MURCIA, Marianne; OLSHTAIN, Elite. *Discourse and context in language teaching:* a guide for language teachers. Cambridge: Cambridge University Press, 2000. 201-215 p.

SMITH, R.M., Polytomous mean-square fit statistics, *Rasch Measurement Transactions.* v.10, n. 3, p. 516-517, 1996. Disponível em <https://rasch.org/rmt/rmt103a.htm> acesso em mai. 2018.

SWALES, John M. *Genre Analysis: English in academic and research settings.* Cambridge: Cambridge University Press, 1990.

TAYLOR, Lynda. Communicating the theory, practice and principles of language testing to test stakeholders: some reflections. *Language Testing*, v. 30, n. 3, p.403-412. 2013

VERGUTS, T.; BOECK, P.D ; A note on the Martin-Lof test for unidimensionality. *Methods of Psicological Reseach – ONLINE*, v.5, n.1, 2000. Disponível em: <https://www.dgps.de/fachgruppen/methoden/mpr-online/issue9/art4/Verguts.html> acesso mai. 2018.

UNDERHILL, Nic. *Testing spoken language:* a handbook of oral testing techniques. Cambridge: Cambridge University Press, 1987.

Celpe-Bras 2015/2

GRADE DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE
OBSERVADOR

INEP Ministério da Educação GOVERNO FEDERAL BRASIL PÁTRIA EDUCADORA

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

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

cespe

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

ANEXO 2: grade holística ou avaliação do interlocutor

Celpe Bras

FICHA DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE | ENTREVISTADOR

INEP Ministério da Educação GOVERNO FEDERAL BRASIL PÁTRIA EDUCADORA

Nome do examinando: Xxxxxxxxxxxxxxxxxxxxxx xx Xxxxxxxxxxxxxxxxxxxxxx xx Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Nacionalidade: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Número de Inscrição: 9999999999

Documento de identificação n.º: XXXXXXXXXXXXXXXXXXXXX

Posto aplicador XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

4360144035

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

Elementos provocadores utilizados

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 examinando

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

ENTREVISTADOR

POSTO APLICADOR

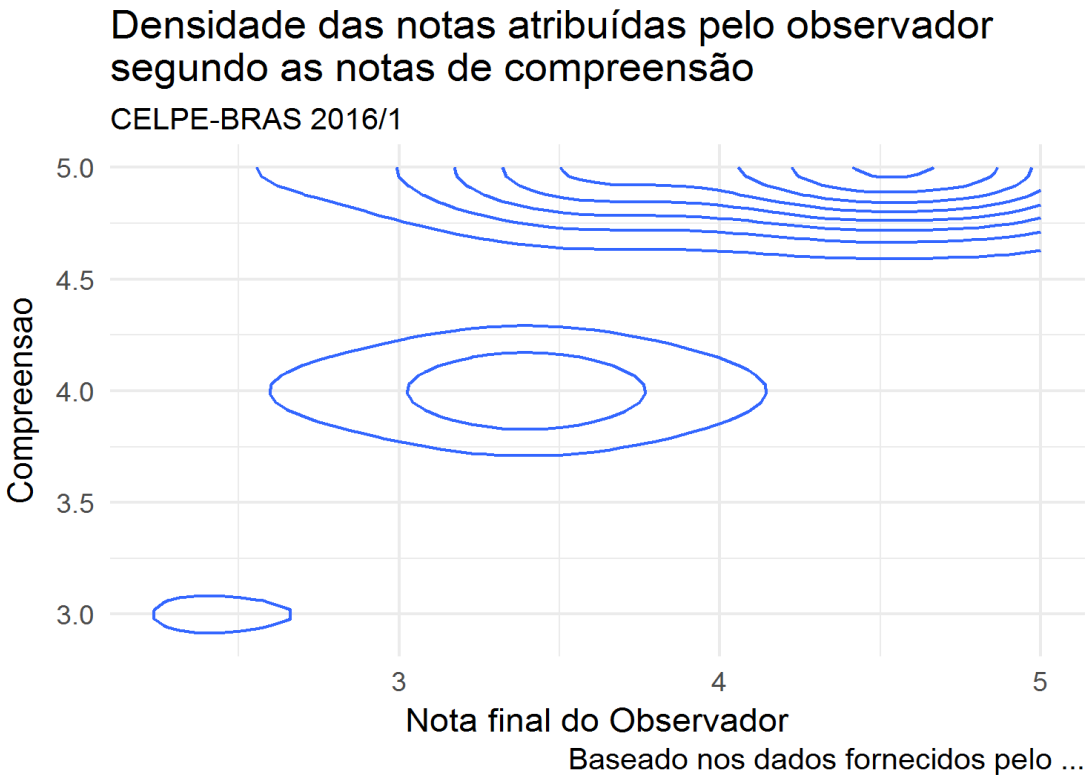
DATA ____________/____________/____________

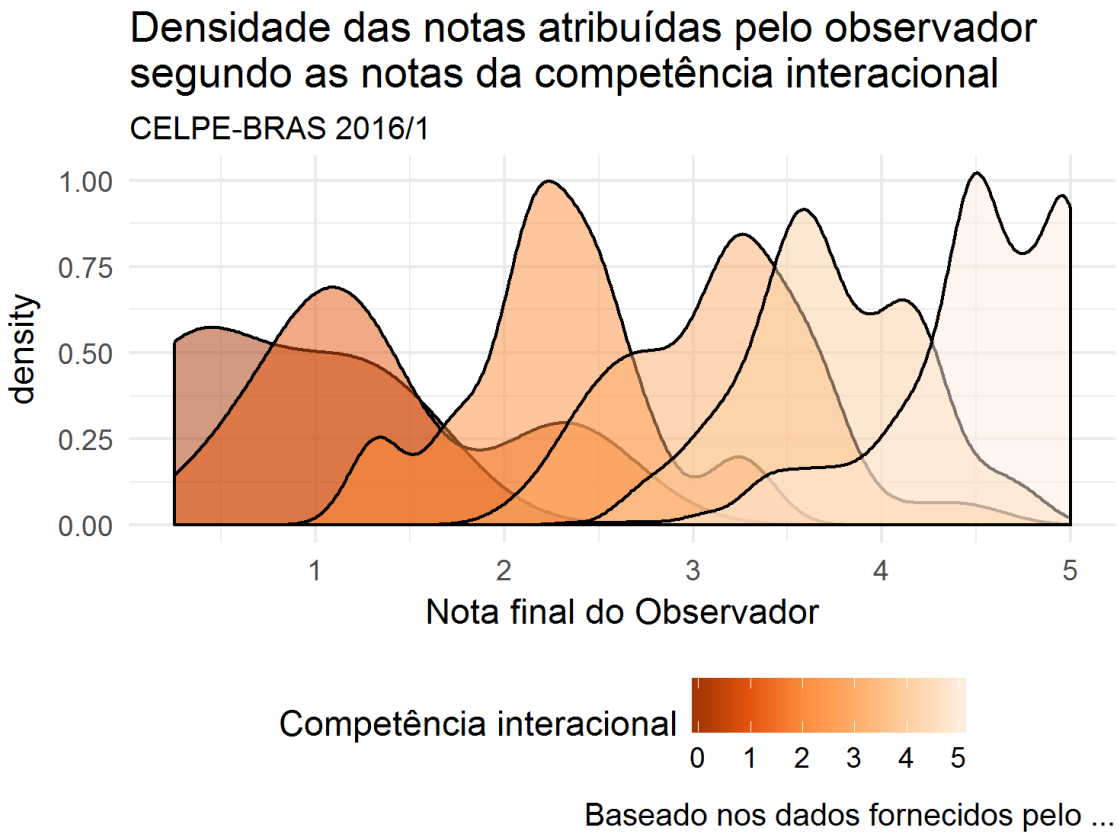
Rubrica

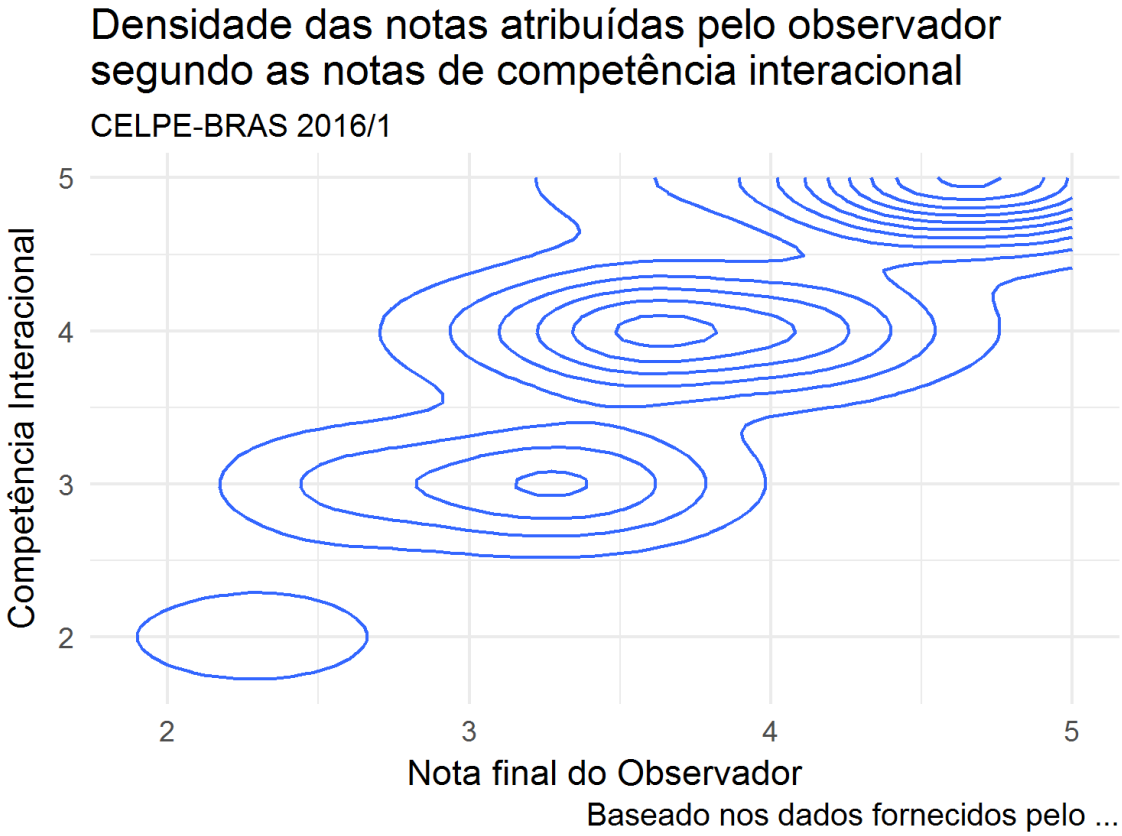
cespe

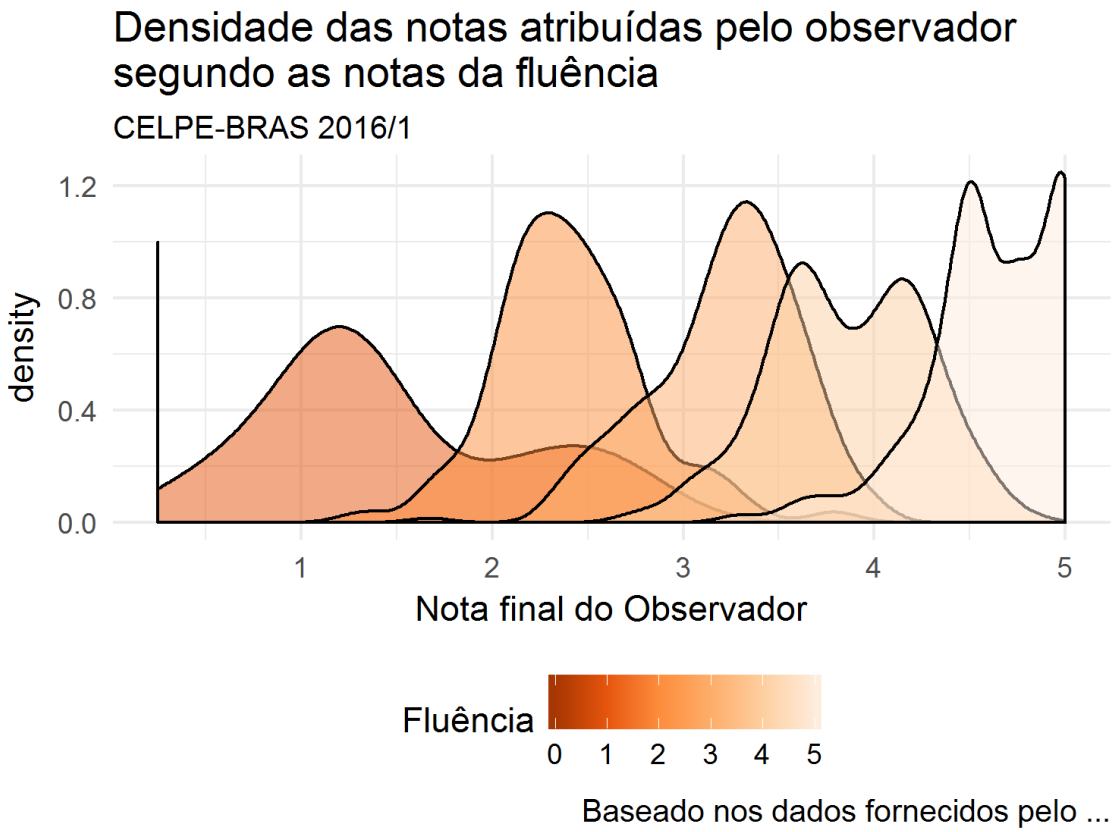
Cebraspe
Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

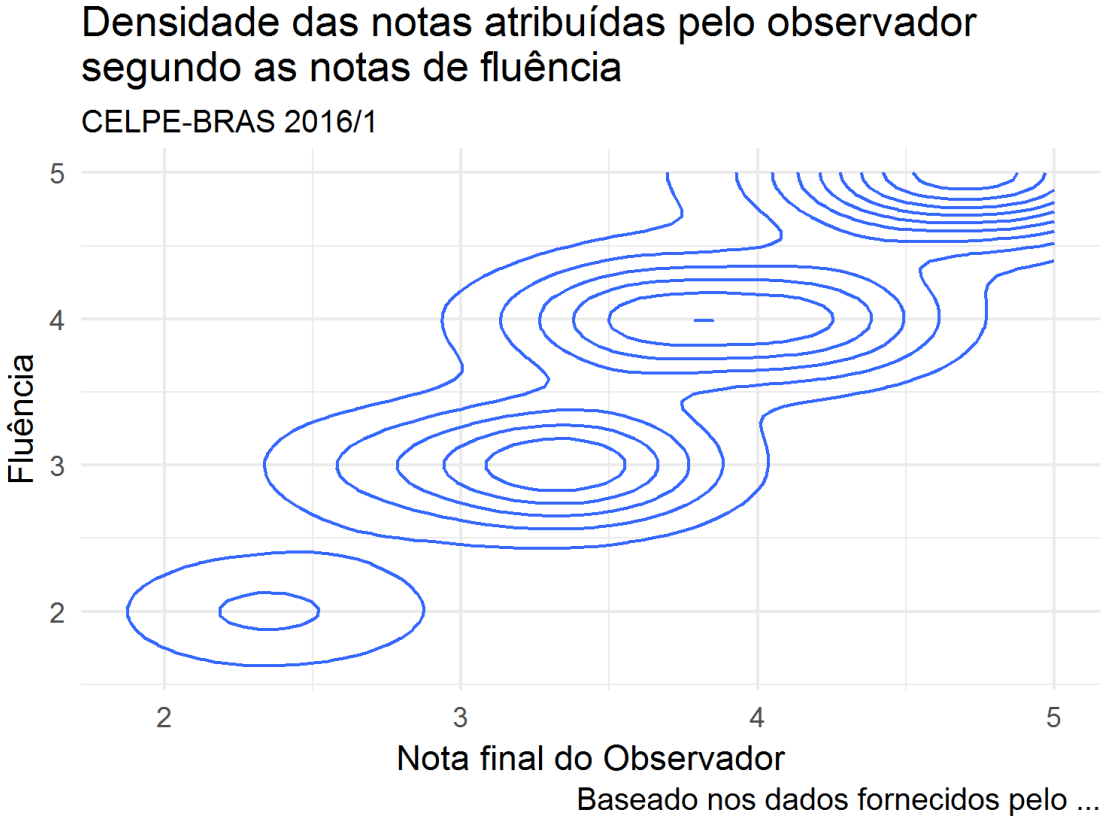
ANEXO 3: gráficos complementares à análise da composição do corpus

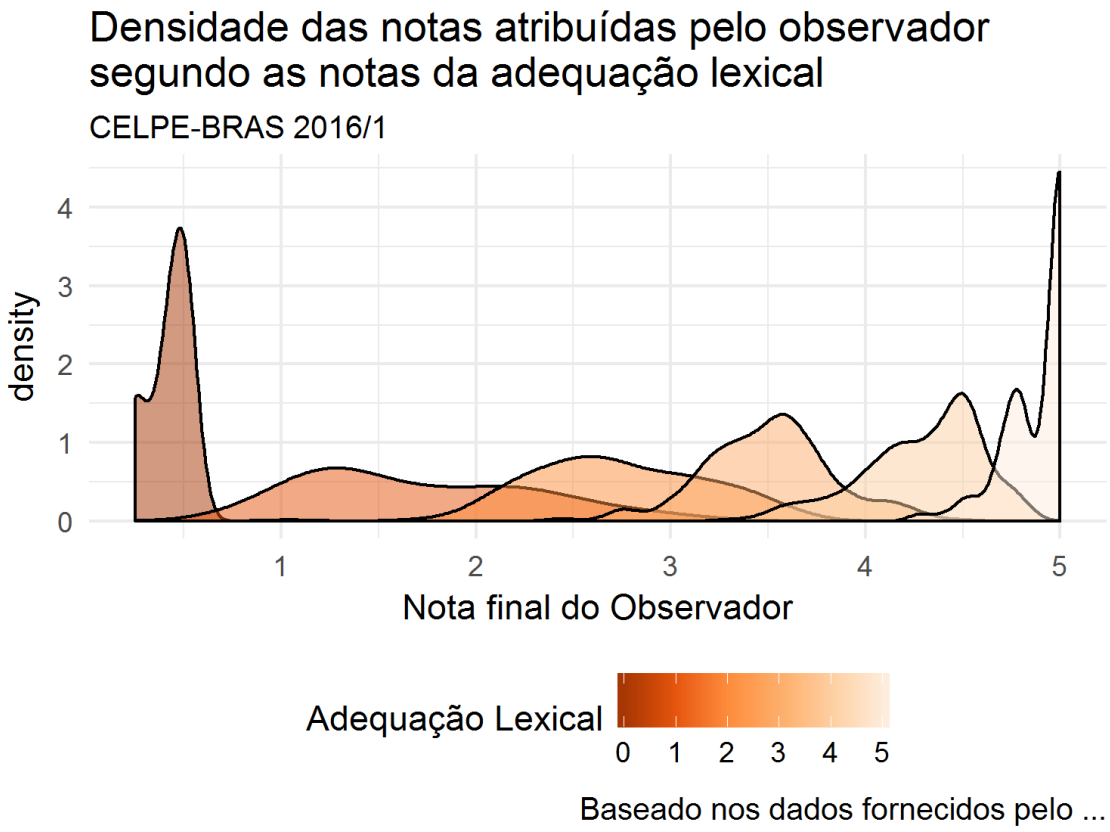


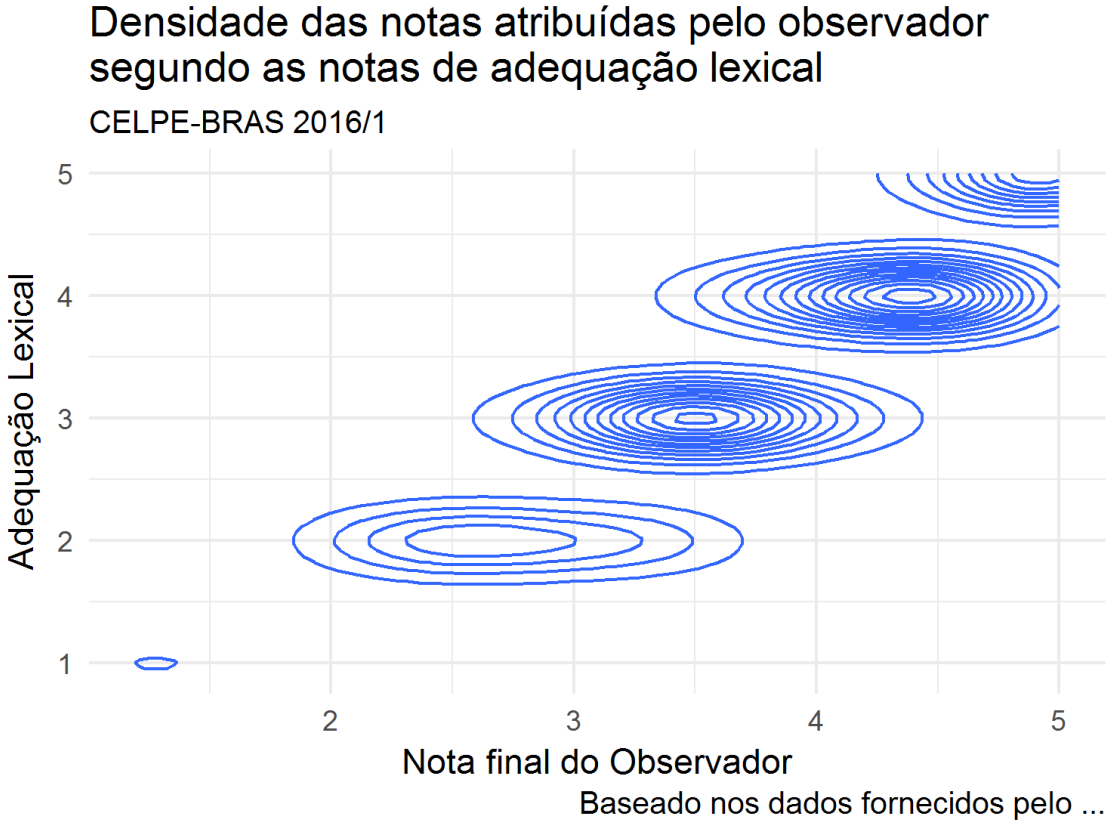

Densidade das notas atribuídas pelo observador segundo as notas da competência interacional
CELPE-BRAS 2016/1
density
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
Nota final do Observador
Competência interacional
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...

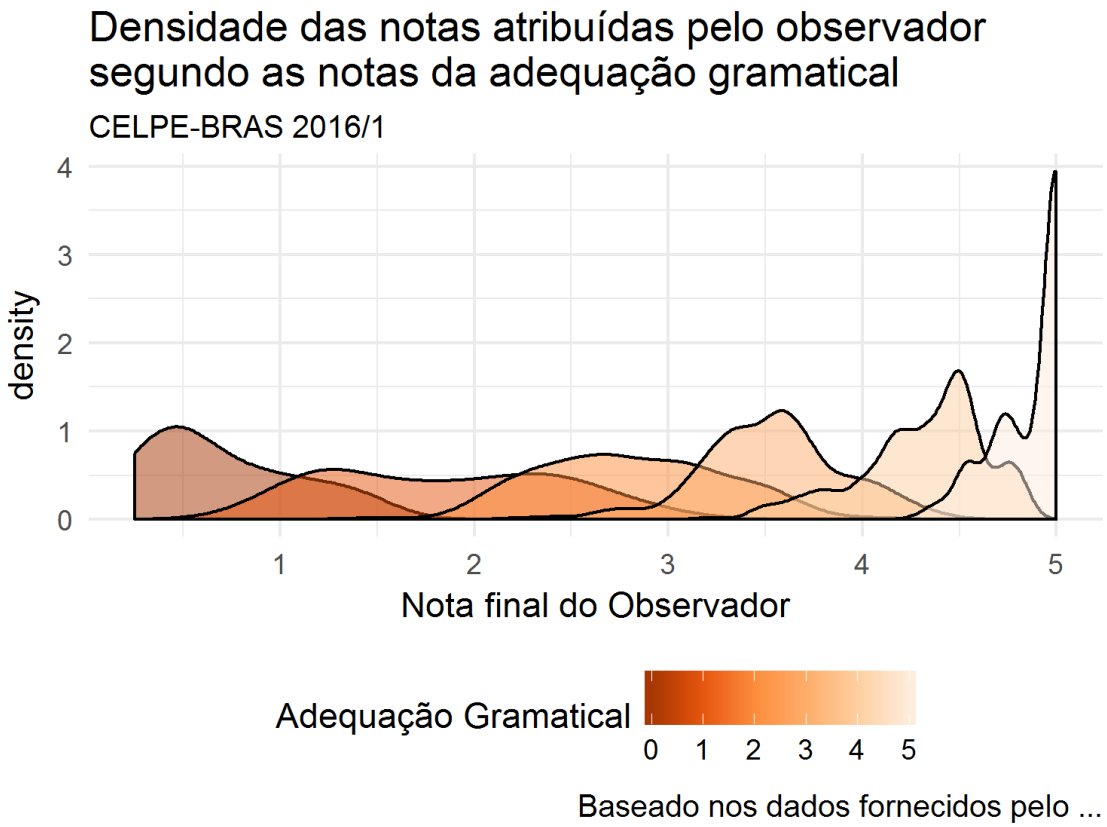

Densidade das notas atribuídas pelo observador segundo as notas de competência interacional
CELPE-BRAS 2016/1
Competência Interacional
5
4
3
2
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...

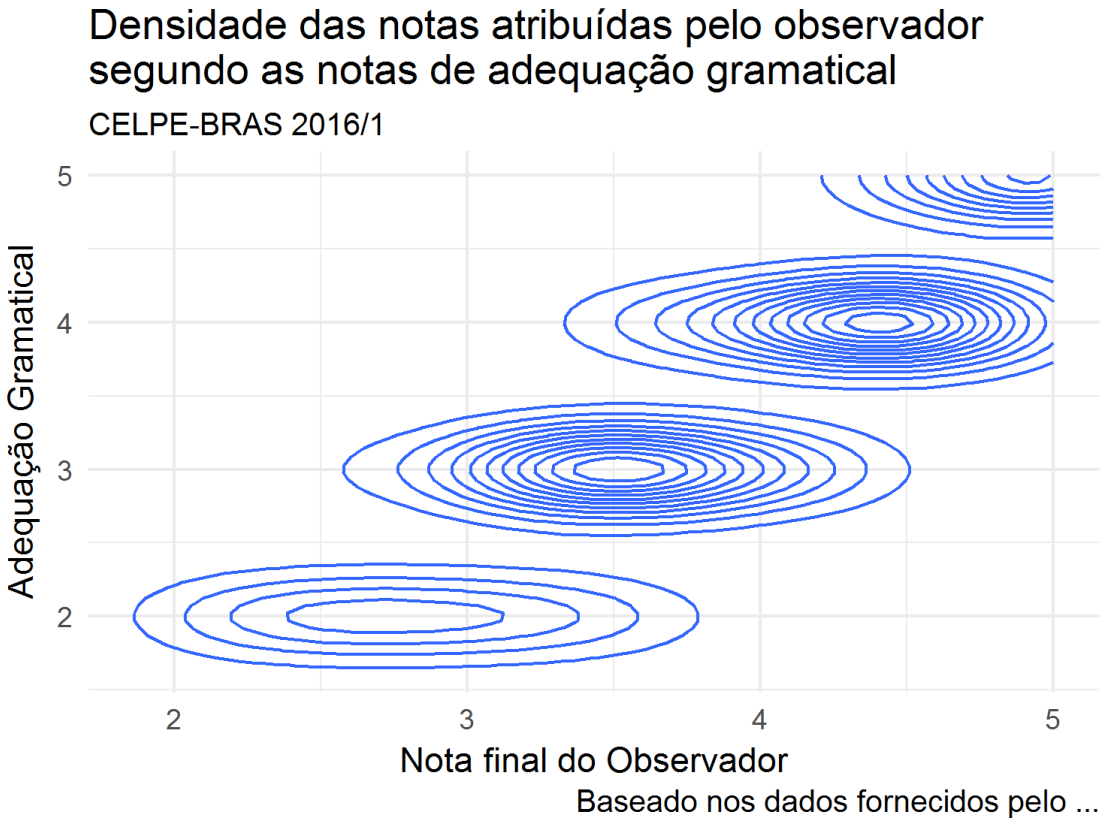

Densidade das notas atribuídas pelo observador segundo as notas da fluência
CELPE-BRAS 2016/1
density
1.2
0.8
0.4
0.0
1
2
3
4
5
Nota final do Observador
Fluência
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...

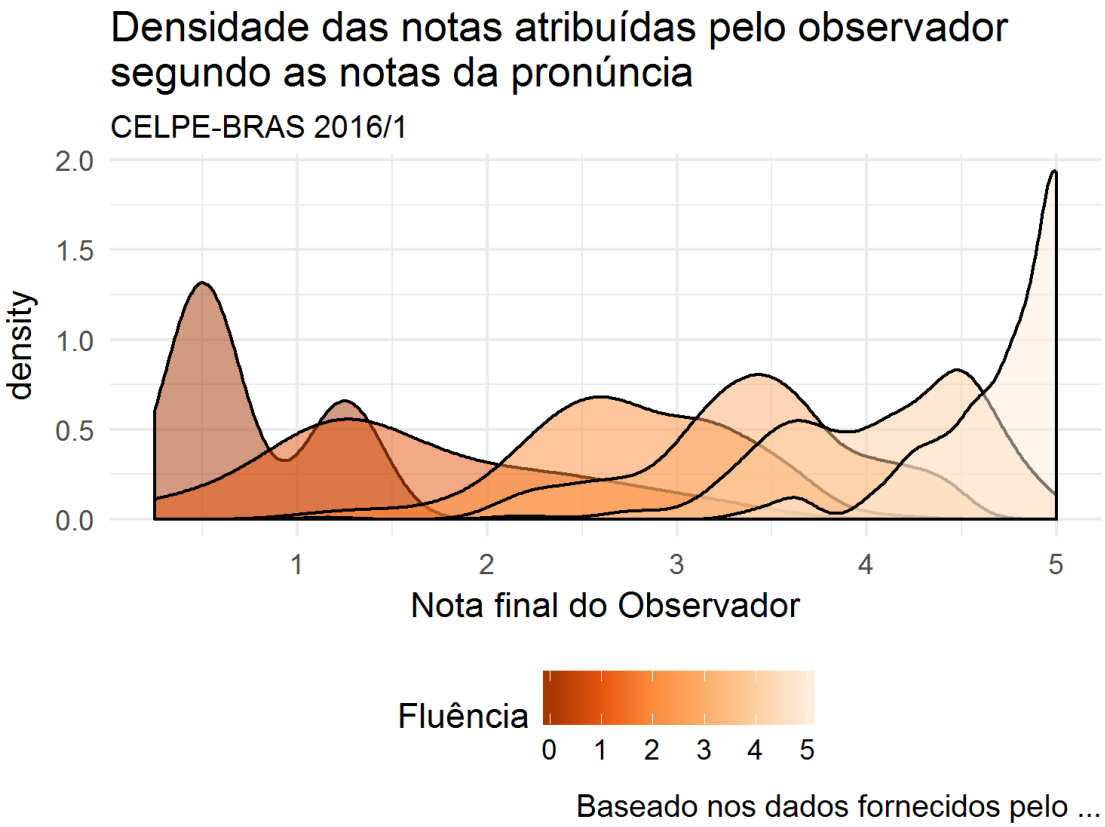

Densidade das notas atribuídas pelo observador segundo as notas de fluência
CELPE-BRAS 2016/1
Fluência
5
4
3
2
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...

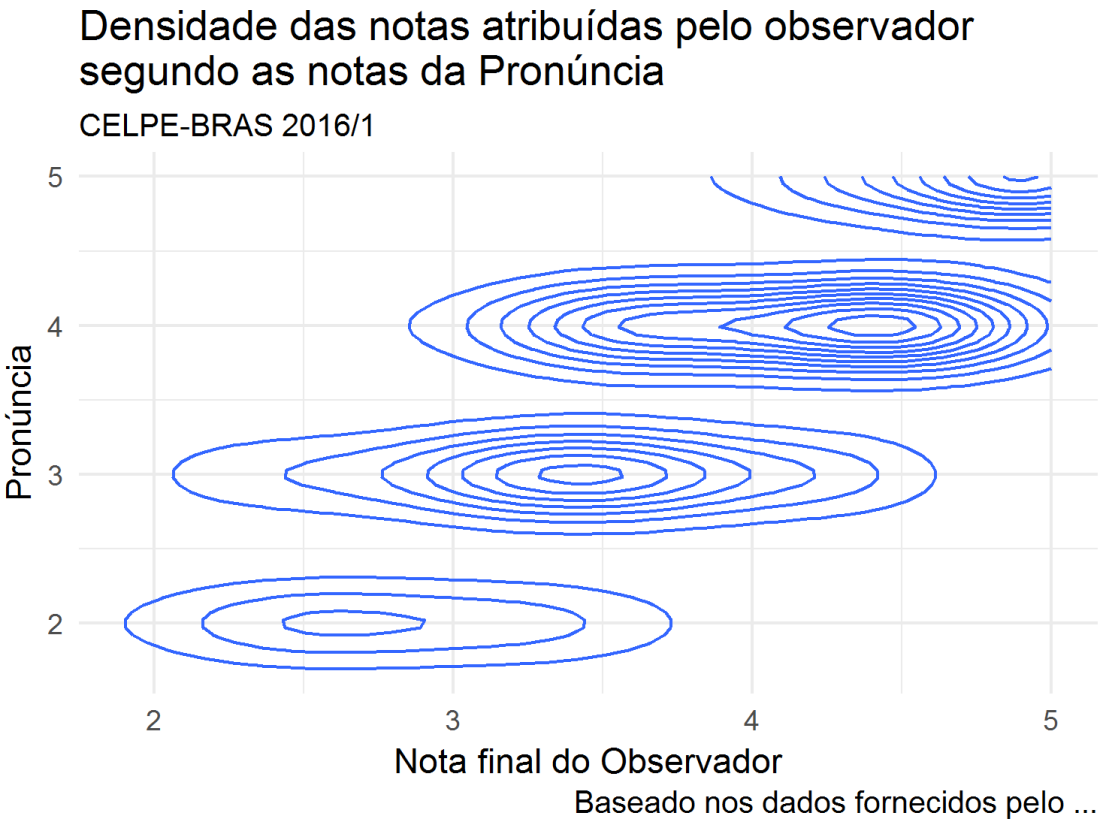

Densidade das notas atribuídas pelo observador segundo as notas da adequação lexical
CELPE-BRAS 2016/1
density
0
1
2
3
4
1
2
3
4
5
Nota final do Observador
Adequação Lexical
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...

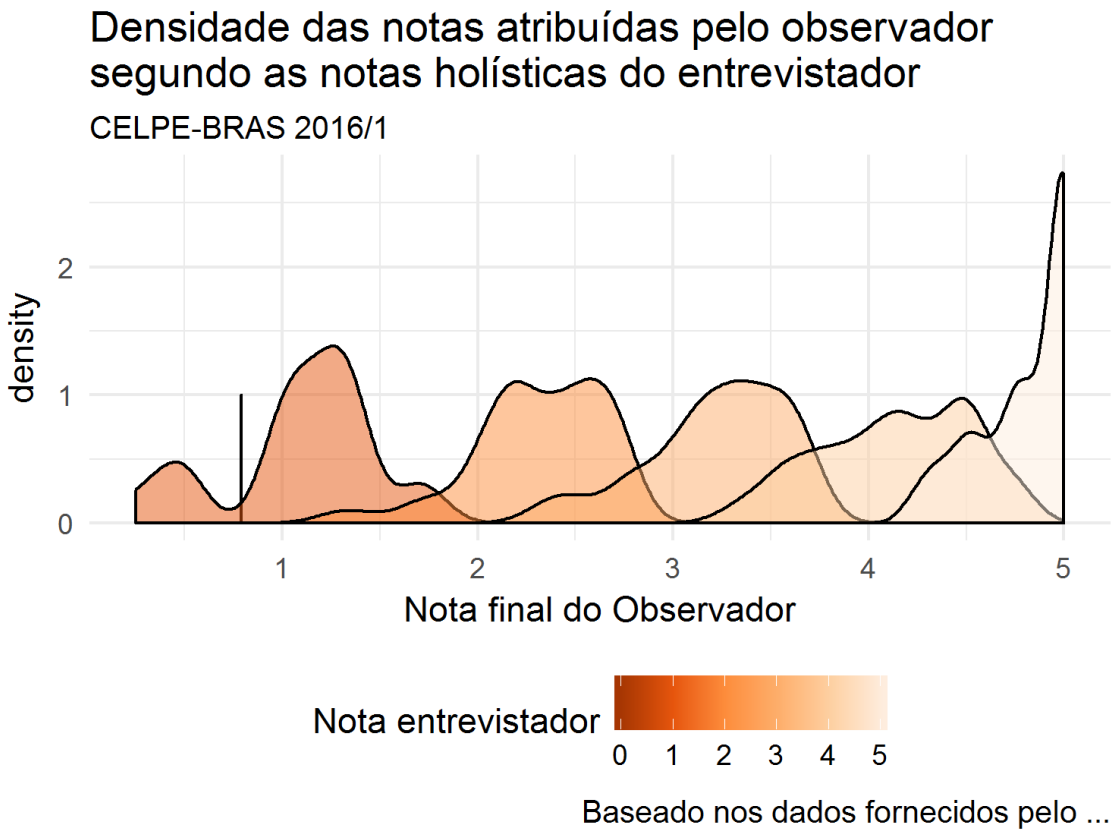

Densidade das notas atribuídas pelo observador segundo as notas de adequação lexical
CELPE-BRAS 2016/1
Adequação Lexical
1
2
3
4
5
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...

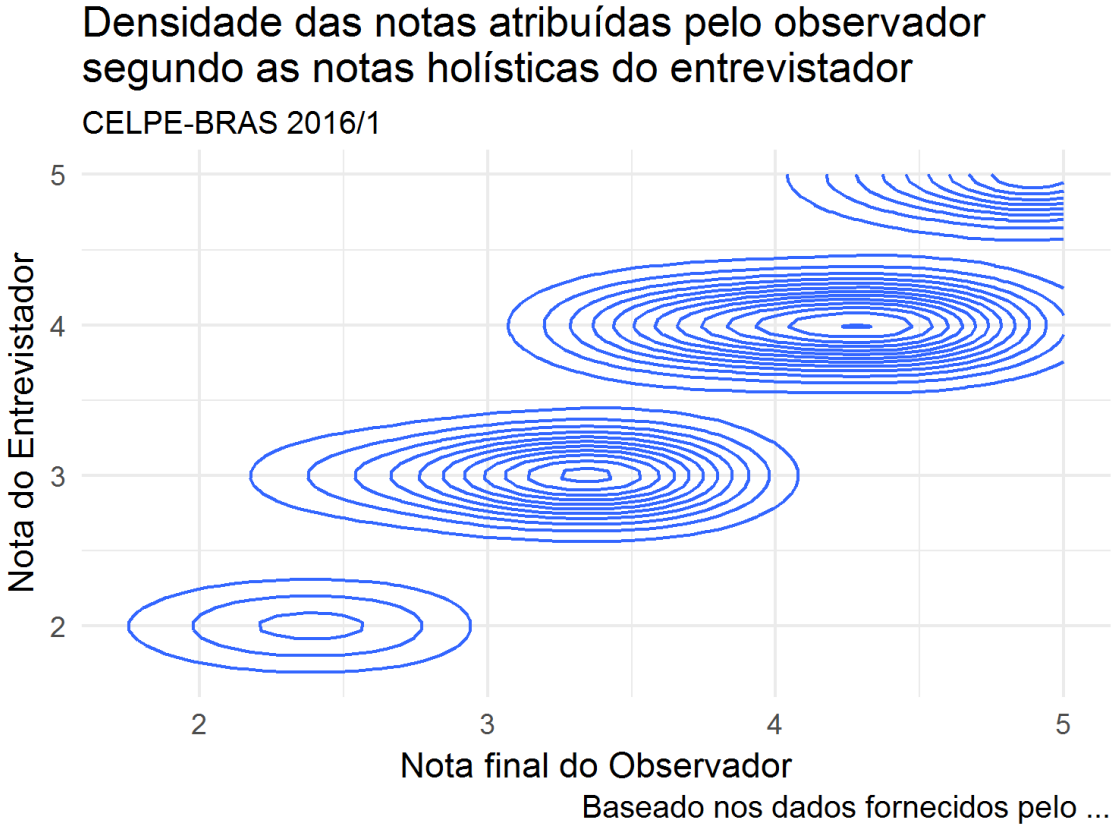

Densidade das notas atribuídas pelo observador segundo as notas da adequação gramatical
CELPE-BRAS 2016/1
density
0
1
2
3
4
1
2
3
4
5
Nota final do Observador
Adequação Gramatical
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...


Densidade das notas atribuídas pelo observador segundo as notas de adequação gramatical
CELPE-BRAS 2016/1
Adequação Gramatical
2
3
4
5
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...

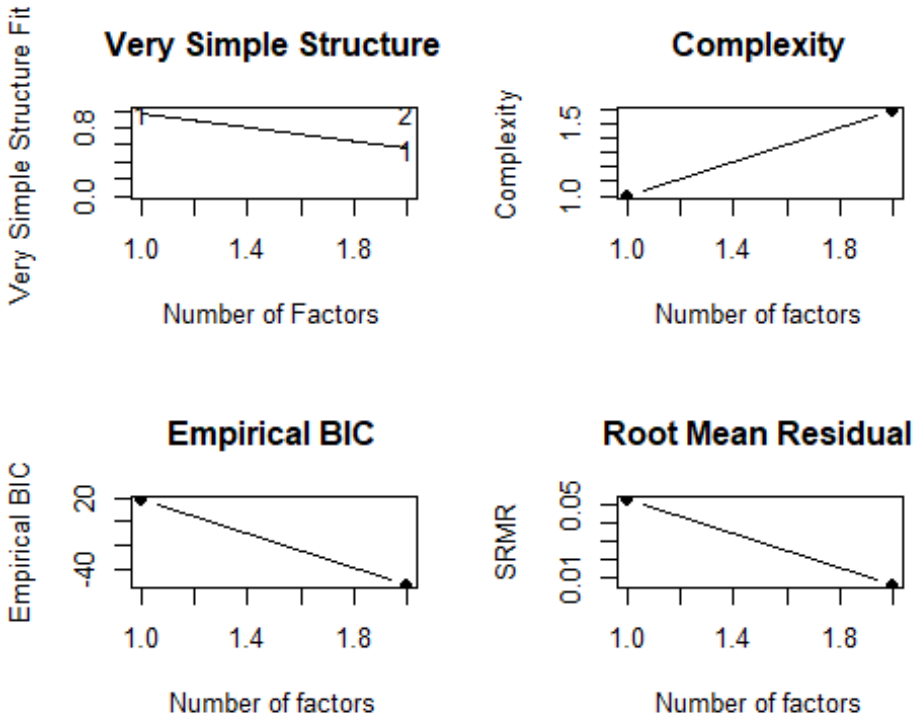

Densidade das notas atribuídas pelo observador segundo as notas da pronúncia
CELPE-BRAS 2016/1
density
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
Nota final do Observador
Fluência
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...


Densidade das notas atribuídas pelo observador segundo as notas da Pronúncia
CELPE-BRAS 2016/1
Pronúncia
5
4
3
2
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...


Densidade das notas atribuídas pelo observador
segundo as notas holísticas do entrevistador
CELPE-BRAS 2016/1
density
0
1
2
1
2
3
4
5
Nota final do Observador
Nota entrevistador
0 1 2 3 4 5
Baseado nos dados fornecidos pelo ...


Densidade das notas atribuídas pelo observador
segundo as notas holísticas do entrevistador
CELPE-BRAS 2016/1
Nota do Entrevistador
2
3
4
5
2
3
4
5
Nota final do Observador
Baseado nos dados fornecidos pelo ...

ANEXO 4: análise fatorial




```
  Notas.Oral.vss.fit

  ##
##                Number                of                factors
## Call: vss(x = x, n = n, rotate = rotate, diagonal = diagonal,
fm                   =                   fm,
##      n.obs = n.obs, plot = FALSE, title = title, use = use,
cor                            =                            cor)
## VSS complexity 1 achieves a maximimum of 0.97   with   1
factors
## VSS complexity 2 achieves a maximimum of 0.98   with   2
factors
## The Velicer MAP achieves a minimum of 0.07  with  1  factors
## Empirical BIC achieves a minimum of  -53.41  with  2  factors
## Sample Size adjusted BIC achieves a minimum of  -8.28  with
2                                                      factors
##
##      Statistics      by      number      of      factors
```

```
##    vss1 vss2    map dof chisq     prob sqresid   fit RMSEA BIC
SABIC                                                       complex
## 1 0.97 0.00 0.070   14    484 2.6e-94     0.85 0.97 0.184 387
431.5                                                         1.0
## 2 0.56 0.98 0.088    8     22 5.8e-03     0.45 0.98 0.041 -34
-8.3                                                          1.6
##                 eChisq                SRMR      eCRMS      eBIC
##     1            115.5      0.0524      0.064                 19
## 2    1.8 0.0066 0.011  -53
```

```
  Notas.Oral.fa.fit <-
  psych::fa(
    r=(Notas.Oral.fa),
    nfactors=1,
    n.iter = 1000L,
    fm="pa",
    alpha = 0.95,
    cor="cor")
```

```
Notas.Oral.fa.fit
```

```
  ## Factor Analysis with confidence intervals using method =
psych::fa(r = (Notas.Oral.fa), nfactors = 1, n.iter = 1000L,
##     fm = "pa", alpha = 0.95, cor = "cor")
## Factor Analysis using method =  pa
## Call: psych::fa(r = (Notas.Oral.fa), nfactors = 1, n.iter =
1000L,
##     fm = "pa", alpha = 0.95, cor = "cor")
## Standardized loadings (pattern matrix) based upon correlation
matrix
##                   PA1   h2    u2 com
## NT_COMPREENSAO   0.65 0.42 0.577   1
## NT_COMPETENCIA   0.80 0.64 0.359   1
## NT_FLUENCIA      0.88 0.78 0.223   1
## NT_AD_GRAMATICAL 0.88 0.78 0.225   1
## NT_AD_LEXICAL    0.91 0.82 0.177   1
## NT_PRONUNCIA     0.80 0.64 0.362   1
## NT_ENTREVISTADOR 0.95 0.90 0.099   1
```

```
##
##                 PA1
## SS loadings    4.98
## Proportion Var 0.71
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## The degrees of freedom for the null model are  21  and the
objective function was  6.87 with Chi Square of  6845.68
## The degrees of freedom for the model are 14  and the
objective function was  0.49
##
## The root mean square of the residuals (RMSR) is  0.05
## The df corrected root mean square of the residuals is  0.06
##
## The harmonic number of observations is  1000 with the
empirical chi square  115.51  with prob <  4.7e-18
## The total number of observations was  1000  with Likelihood
Chi Square =  483.71  with prob <  2.6e-94
##
## Tucker Lewis Index of factoring reliability =  0.897
## RMSEA index =  0.184  and the 5 % confidence intervals are
0.183 0.184
## BIC =  387
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1
## Correlation of (regression) scores with factors   0.98
## Multiple R square of scores with factors          0.96
## Minimum correlation of possible factor scores     0.92
##
##  Coefficients and bootstrapped confidence intervals
##                   low  PA1 upper
## NT_COMPREENSAO   0.60 0.65  0.69
## NT_COMPETENCIA   0.77 0.80  0.83
## NT_FLUENCIA      0.86 0.88  0.90
```

```
## NT_AD_GRAMATICAL 0.87 0.88  0.89
## NT_AD_LEXICAL    0.90 0.91  0.92
## NT_PRONUNCIA     0.77 0.80  0.82
## NT_ENTREVISTADOR 0.94 0.95  0.96
```

```
  Notas.Oral.fa.fit$fit
```

```
  ## [1] 0.9698597
```

```
  Notas.Oral.fa.fit$fit.off
```

```
  ## [1] 0.994554
```

```
  Notas.Oral.fa.fit$RMSEA
```

```
  ##      RMSEA      lower      upper confidence
##  0.1836240  0.1829210  0.1839888  0.0500000
```

```
  Notas.Oral.fa.fit$R2.scores
```

```
  ## [1] 0.9617451
```

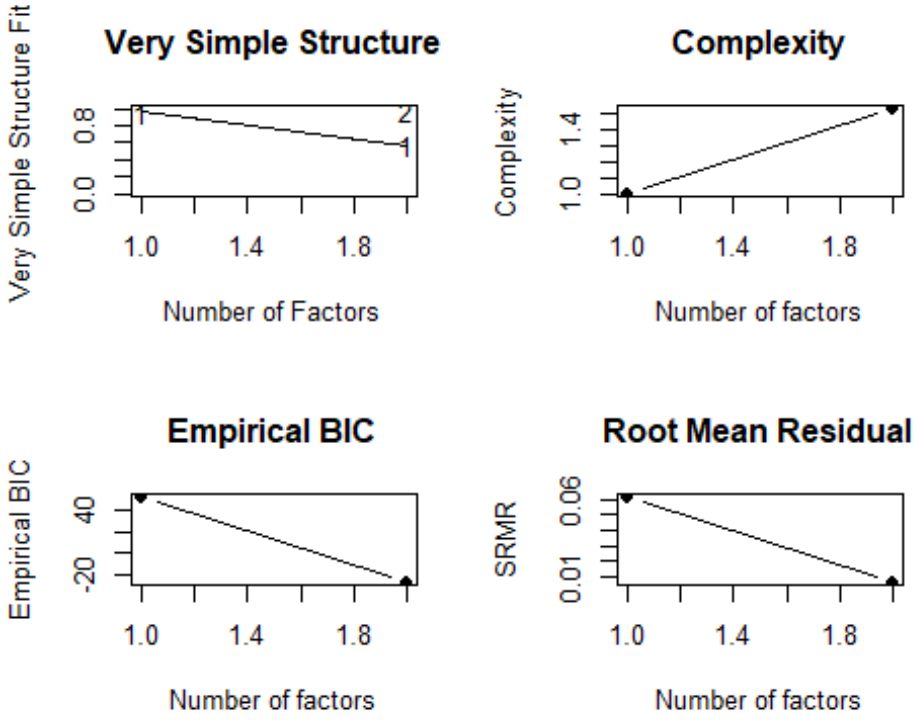
```
  Notas.Oral.fa.fit$R2.scores
```

```
  ## [1] 0.9617451
```

```
  Notas.Oral.fa.fit$TLI
```

```
  ## [1] 0.8966921
```

```
  Notas.Oral.vss.fit.6 <-
 nfactors(
   x=Notas.Oral.fa.6,
   n=2,
   fm="pa")
```




```
  Notas.Oral.vss.fit.6
## Number of factors
## Call: vss(x = x, n = n, rotate = rotate, diagonal = diagonal,
fm = fm,
##     n.obs = n.obs, plot = FALSE, title = title, use = use,
cor = cor)
## VSS complexity 1 achieves a maximimum of 0.96  with  1
factors
## VSS complexity 2 achieves a maximimum of 0.98  with  2
factors
## The Velicer MAP achieves a minimum of 0.1  with  1  factors
## Empirical BIC achieves a minimum of  -26.39  with  2  factors
## Sample Size adjusted BIC achieves a minimum of  -4.48  with
2  factors
##
## Statistics by number of factors
##   vss1 vss2  map dof chisq    prob sqresid  fit RMSEA BIC
SABIC complex
## 1 0.96 0.00 0.10   9   468 5.0e-95    0.84 0.96  0.23 405
434.0     1.0
## 2 0.57 0.98 0.13   4    10 3.3e-02    0.44 0.98  0.04 -17
```

```
-4.5     1.5
##   eChisq   SRMR eCRMS eBIC
## 1  114.7 0.0618 0.080   52
## 2    1.2 0.0064 0.012  -26
```

```
Notas.Oral.fa.fit.6 <-
  psych::fa(
    r=(Notas.Oral.fa.6),
    nfactors=1,
    n.iter = 1000L,
    fm="pa",
    alpha = 0.95,
    cor="cor")

Notas.Oral.fa.fit.6
```

```
   ## Factor Analysis with confidence intervals using method =
psych::fa(r = (Notas.Oral.fa.6), nfactors = 1, n.iter = 1000L,
##     fm = "pa", alpha = 0.95, cor = "cor")
## Factor Analysis using method =  pa
## Call: psych::fa(r = (Notas.Oral.fa.6), nfactors = 1, n.iter =
1000L,
##     fm = "pa", alpha = 0.95, cor = "cor")
## Standardized loadings (pattern matrix) based upon correlation
matrix
##                   PA1   h2   u2 com
## NT_COMPREENSAO   0.65 0.42 0.58   1
## NT_COMPETENCIA   0.80 0.64 0.36   1
## NT_FLUENCIA      0.88 0.78 0.22   1
## NT_AD_GRAMATICAL 0.88 0.78 0.22   1
## NT_AD_LEXICAL    0.91 0.82 0.18   1
## NT_PRONUNCIA     0.80 0.64 0.36   1
##
##                  PA1
## SS loadings    4.07
## Proportion Var 0.68
##
## Mean item complexity =  1
```

```
## Test of the hypothesis that 1 factor is sufficient.
##
## The degrees of freedom for the null model are  15  and the
objective function was  4.99 with Chi Square of  4971.94
## The degrees of freedom for the model are 9  and the objective
function was  0.47
##
## The root mean square of the residuals (RMSR) is  0.06
## The df corrected root mean square of the residuals is  0.08
##
## The harmonic number of observations is  1000 with the
empirical chi square  114.66  with prob <  1.7e-20
## The total number of observations was  1000  with Likelihood
Chi Square =  467.58  with prob <  5e-95
##
## Tucker Lewis Index of factoring reliability =  0.846
## RMSEA index =  0.226  and the 5 % confidence intervals are
0.225 0.227
## BIC =  405.41
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1
## Correlation of (regression) scores with factors   0.97
## Multiple R square of scores with factors          0.94
## Minimum correlation of possible factor scores     0.88
##
##  Coefficients and bootstrapped confidence intervals
##                    low  PA1 upper
## NT_COMPREENSAO   0.60 0.65  0.69
## NT_COMPETENCIA   0.77 0.80  0.83
## NT_FLUENCIA      0.87 0.88  0.90
## NT_AD_GRAMATICAL 0.87 0.88  0.90
## NT_AD_LEXICAL    0.89 0.91  0.92
## NT_PRONUNCIA     0.77 0.80  0.82
```

```
  Notas.Oral.fa.fit.6$fit

  ## [1] 0.9575081

  Notas.Oral.fa.fit.6$fit.off

  ## [1] 0.9917168

  Notas.Oral.fa.fit.6$RMSEA

  ##          RMSEA          lower          upper  confidence
##  0.2262465  0.2254232  0.2267516  0.0500000

  Notas.Oral.fa.fit.6$R2.scores

  ## [1] 0.9398724

  Notas.Oral.fa.fit.6$R2.scores

  ## [1] 0.9398724

  devtools::session_info()

                      ## - Session info
   -------------------------------------------------------------
    ##  setting  value
    ##  version  R version 3.5.0 Patched (2018-04-23 r74633)
    ##  os       Windows 10 x64
    ##  system   x86_64, mingw32
    ##  ui       RTerm
    ##  language (EN)
    ##  collate  Portuguese_Brazil.1252
    ##  tz       America/Sao_Paulo
    ##  date     2018-05-22
                                ##
                         ## - Packages
 -------------------------------------------------------------
        ##  package     * version     date       source

    ##  assertthat    0.2.0       2017-04-11 CRAN (R 3.5.0)

    ##  backports     1.1.2       2017-12-13 CRAN (R 3.5.0)
```

```
    ##  bindr         0.1.1.9000  2018-05-12 Github
              (krlmlr/bindr@b6e6fd6)
    ##  bindrcpp    * 0.2.2.9000  2018-05-12 Github
              (krlmlr/bindrcpp@bd5ae73)
    ##  broom         0.4.4       2018-05-12 Github
              (tidyverse/broom@570b25a)
##  callr         2.0.3       2018-04-11 CRAN (R 3.5.0)

##  cellranger    1.1.0       2016-07-27 CRAN (R 3.5.0)

##  cli           1.0.0       2017-11-05 CRAN (R 3.5.0)

##  clisymbols    1.2.0       2017-05-21 CRAN (R 3.5.0)

##  colorspace    1.3-2       2016-12-14 CRAN (R 3.5.0)

##  crayon        1.3.4       2017-09-16 CRAN (R 3.5.0)

##  debugme       1.1.0       2017-10-22 CRAN (R 3.5.0)

##  desc          1.2.0       2018-05-01 CRAN (R 3.5.0)

    ##  devtools      1.13.5.9000 2018-05-12 Github
              (hadley/devtools@13ee56b)
##  digest        0.6.15      2018-01-28 CRAN (R 3.5.0)

    ##  dplyr       * 0.7.5.9000  2018-05-12 Github
              (tidyverse/dplyr@09209ae)
    ##  evaluate      0.10.3      2018-05-12 Github
              (hadley/evaluate@06f8e24)
    ##  forcats     * 0.3.0.9000  2018-05-12 Github
              (tidyverse/forcats@f4a7fd1)
##  foreign       0.8-70      2017-11-28 CRAN (R 3.5.0)

    ##  ggplot2     * 2.2.1.9000  2018-05-12 Github
              (tidyverse/ggplot2@4463da6)
##  glue          1.2.0       2017-10-29 CRAN (R 3.5.0)
```

```
    ##  gtable        0.2.0.9000  2018-05-12 Github
              (hadley/gtable@0ed36a4)
    ##  haven         1.1.1.9000  2018-05-12 Github
              (tidyverse/haven@746eb3e)
    ##  hms           0.4.2       2018-05-12 Github
              (tidyverse/hms@c0cfc01)
##  htmltools     0.3.6       2017-04-28 CRAN (R 3.5.0)

  ##  httr          1.3.1       2018-05-12 Github (r-
               lib/httr@6b2dadc)
##  jsonlite      1.5         2017-06-01 CRAN (R 3.5.0)

    ##  knitr         1.20.3      2018-05-12 Github
              (yihui/knitr@dc028f4)
##  lattice       0.20-35     2017-03-25 CRAN (R 3.5.0)

    ##  lazyeval      0.2.1.9000  2018-05-12 Github
              (hadley/lazyeval@93c455c)
    ##  lubridate     1.7.4       2018-05-12 Github
              (tidyverse/lubridate@45395b4)
    ##  magrittr      1.5.0       2018-05-12 Github
              (tidyverse/magrittr@0a76de2)
##  memoise       1.1.0       2017-04-21 CRAN (R 3.5.0)

##  mnormt        1.5-5       2016-10-15 CRAN (R 3.5.0)

##  modelr        0.1.2       2018-05-11 CRAN (R 3.5.0)

##  munsell       0.4.3       2016-02-13 CRAN (R 3.5.0)

##  nlme          3.1-137     2018-04-07 CRAN (R 3.5.0)

##  pillar        1.2.2       2018-04-26 CRAN (R 3.5.0)

  ##  pkgbuild      1.0.0       2018-05-12 Github (r-
               lib/pkgbuild@0457039)
```

```
##  pkgconfig     2.0.1       2017-03-21 CRAN (R 3.5.0)

  ##  pkgload       1.0.0       2018-05-12 Github (r-
              lib/pkgload@35efedd)
##  plyr          1.8.4       2016-06-08 CRAN (R 3.5.0)

##  psych       * 1.8.4       2018-05-06 CRAN (R 3.5.0)

    ##  purrr       * 0.2.4.9000  2018-05-12 Github
              (tidyverse/purrr@fda4bbe)
  ##  R6            2.2.2.9000  2018-05-12 Github (r-
               lib/R6@a9eb0f1)
    ##  Rcpp          0.12.17     2018-05-12 Github
              (RcppCore/Rcpp@001db74)
    ##  readr       * 1.2.0       2018-05-12 Github
              (tidyverse/readr@d6d622b)
    ##  readxl        1.1.0.9000  2018-05-12 Github
              (tidyverse/readxl@b10a1a8)
    ##  reshape2      1.4.3       2018-05-12 Github
              (hadley/reshape@777638a)
    ##  rlang         0.2.0.9001  2018-05-12 Github
              (tidyverse/rlang@ccdbd8b)
    ##  rmarkdown     1.9.11      2018-05-12 Github
              (rstudio/rmarkdown@41b2bab)
##  rprojroot     1.3-2       2018-01-03 CRAN (R 3.5.0)

##  rstudioapi    0.7         2017-09-07 CRAN (R 3.5.0)

    ##  rvest         0.3.2.9000  2018-05-12 Github
              (hadley/rvest@9a51a5d)
    ##  scales        0.5.0.9000  2018-05-12 Github
              (hadley/scales@d767915)
##  sessioninfo   1.0.0       2017-06-21 CRAN (R 3.5.0)

##  stringi       1.2.2       2018-05-02 CRAN (R 3.5.0)

    ##  stringr     * 1.3.1       2018-05-12 Github
```

```
                (tidyverse/stringr@eff4e4d)
    ##  testthat      2.0.0       2017-12-13 CRAN (R 3.5.0)

    ##  tibble      * 1.4.2       2018-01-22 CRAN (R 3.5.0)

    ##  tidyr       * 0.8.0       2018-01-29 CRAN (R 3.5.0)

    ##  tidyselect    0.2.4       2018-02-26 CRAN (R 3.5.0)

        ##  tidyverse   * 1.2.1.9000  2018-05-12 Github
                (tidyverse/tidyverse@83f6ec3)
    ##  usethis       1.3.0       2018-02-24 CRAN (R 3.5.0)

        ##  withr         2.1.2       2018-05-12 Github
                (jimhester/withr@79d7b0d)
      ##  xml2          1.2.0.9000  2018-05-12 Github (r-
                 lib/xml2@ba3511f)
    ##  yaml          2.1.19      2018-05-01 CRAN (R 3.5.0)
```

```
  scores.fa.boot <- function(data,b, fm="pa", cor="cor")
{
  d = data[b,]
  weights <-
    psych::fa(
      r=d,
      nfactors=1,
      fm=fm,
      cor=cor
    )$weights
  return(weights)
}

loadings.fa.boot <- function(data,b, fm="pa", cor="cor")
{
  d = data[b,]
  loadings <-
    psych::fa(
      r=d,
```

```
      nfactors=1,
      fm=fm,
      cor=cor
    )$loadings
  return(loadings)
}

   Notas.Oral <-
  readr::read_delim(
    "Demanda_Laura_Notas da parte Oral_Celpe_Bras_2016.1.csv",
    ";", escape_double = FALSE,
    col_types =
      cols(
        `NT_AD_GRAMATICAL` = col_integer(),
        `NT_AD_LEXICAL` = col_integer(),
        `NT_COMPETENCIA` = col_integer(),
        `NT_COMPREENSAO` = col_integer(),
        `NT_FLUENCIA` = col_integer(),
        `NT_PRONUNCIA` = col_integer()
        ),
     trim_ws = TRUE, skip = 1)

   Notas.Oral %>%
  mutate(
    NT_COMPREENSAO = if_else(NT_COMPREENSAO == 0, 1L,
NT_COMPREENSAO) - 1,
    NT_COMPETENCIA = if_else(NT_COMPETENCIA == 0, 1L,
NT_COMPETENCIA) - 1,
    NT_FLUENCIA = if_else(NT_FLUENCIA == 0, 1L, NT_FLUENCIA) -
1,
    NT_AD_GRAMATICAL = if_else(NT_AD_GRAMATICAL == 0, 1L,
NT_AD_GRAMATICAL) - 1,
    NT_AD_LEXICAL = if_else(NT_AD_LEXICAL == 0, 1L,
NT_AD_LEXICAL) - 1,
    NT_PRONUNCIA = if_else(NT_PRONUNCIA == 0, 1L, NT_PRONUNCIA)
- 1,
    NT_ENTREVISTADOR = if_else(NT_ENTREVISTADOR == 0,  1,
NT_ENTREVISTADOR) - 1
```

```
  ) %>%
  select(
    NT_COMPREENSAO, NT_COMPETENCIA, NT_FLUENCIA,
    NT_AD_GRAMATICAL, NT_AD_LEXICAL, NT_PRONUNCIA,
    NT_ENTREVISTADOR
  ) -> Notas.Oral.fa

Notas.Oral.fa %>%
  select(-NT_ENTREVISTADOR) -> Notas.Oral.fa.6


   ptm  <-  proc.time()  #  have  a  look  at  the  time  it  takes

Celpe.boot                                                      <-
                                                          boot::boot(
                                                       data=Notas.Oral.fa,
                                                   statistic=scores.fa.boot,
                                                                R=10000L,
                                                    parallel="multicore",
                                                                 ncpus=3,
                                                fm="pa",          cor="cor"
                                                                       )

proc.time() - ptm # about 5 minutes on my end

   ##                      user           system     elapsed
##  276.10    3.95  349.96

   for(i in 1:7) {
 cat('\nitem', i,' - ', names(Celpe.boot$data)[i],'\n')

  print(
    boot::boot.ci(
      Celpe.boot,
      index=i,
      type=c("norm","basic", "perc", "bca")
    )
```

```
  )
}

  ##
## item 1  -  NT_COMPREENSAO
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0301,  0.0541 )   ( 0.0300,  0.0542 )
##
## Level     Percentile            BCa
## 95%   ( 0.0295,  0.0537 )   ( 0.0300,  0.0540 )
## Calculations and Intervals on Original Scale
##
## item 2  -  NT_COMPETENCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0714,  0.1132 )   ( 0.0711,  0.1129 )
##
## Level     Percentile            BCa
## 95%   ( 0.0716,  0.1135 )   ( 0.0714,  0.1131 )
## Calculations and Intervals on Original Scale
##
```

```
## item 3  -  NT_FLUENCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.1599,  0.2142 )   ( 0.1589,  0.2137 )
##
## Level     Percentile            BCa
## 95%   ( 0.1599,  0.2147 )   ( 0.1598,  0.2145 )
## Calculations and Intervals on Original Scale
##
## item 4  -  NT_AD_GRAMATICAL
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0987,  0.1560 )   ( 0.0981,  0.1552 )
##
## Level     Percentile            BCa
## 95%   ( 0.1003,  0.1574 )   ( 0.0997,  0.1567 )
## Calculations and Intervals on Original Scale
##
## item 5  -  NT_AD_LEXICAL
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
```

```
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.1545,  0.2270 )   ( 0.1543,  0.2275 )
##
## Level     Percentile            BCa
## 95%   ( 0.1551,  0.2283 )   ( 0.1537,  0.2267 )
## Calculations and Intervals on Original Scale
##
## item 6  -  NT_PRONUNCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0455,  0.0850 )   ( 0.0454,  0.0853 )
##
## Level     Percentile            BCa
## 95%   ( 0.0446,  0.0844 )   ( 0.0449,  0.0849 )
## Calculations and Intervals on Original Scale
##
## item 7  -  NT_ENTREVISTADOR
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
```

```
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.3356,  0.4197 )   ( 0.3352,  0.4201 )
##
## Level     Percentile            BCa
## 95%   ( 0.3343,  0.4192 )   ( 0.3355,  0.4201 )
## Calculations and Intervals on Original Scale
```

```
   ptm <- proc.time() # have a look at the time it takes

Celpe.boot <-
  boot::boot(
    data=Notas.Oral.fa,
    statistic=loadings.fa.boot,
    R=10000L,
    parallel="multicore",
    ncpus=3,
    fm="pa", cor="cor"
  )

proc.time() - ptm # about 5 minutes on my end
```

```
   ##    user  system elapsed
##  256.42    3.39  291.38
```

```
   for(i in 1:7) {
 cat('\nitem', i,' - ', names(Celpe.boot$data)[i],'\n')

  print(
    boot::boot.ci(
      Celpe.boot,
      index=i,
      type=c("norm","basic", "perc", "bca")
    )
  )
}
```

```
##
## item 1  -  NT_COMPREENSAO
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.6060,  0.6948 )   ( 0.6079,  0.6958 )
##
## Level     Percentile            BCa
## 95%   ( 0.6043,  0.6922 )   ( 0.6041,  0.6919 )
## Calculations and Intervals on Original Scale
##
## item 2  -  NT_COMPETENCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.7717,  0.8290 )   ( 0.7731,  0.8298 )
##
## Level     Percentile            BCa
## 95%   ( 0.7709,  0.8277 )   ( 0.7697,  0.8266 )
## Calculations and Intervals on Original Scale
##
## item 3  -  NT_FLUENCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
```

```
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.8643,  0.8991 )   ( 0.8651,  0.9000 )
##
## Level     Percentile            BCa
## 95%   ( 0.8632,  0.8981 )   ( 0.8628,  0.8977 )
## Calculations and Intervals on Original Scale
##
## item 4  -  NT_AD_GRAMATICAL
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.8655,  0.8951 )   ( 0.8661,  0.8957 )
##
## Level     Percentile            BCa
## 95%   ( 0.8651,  0.8947 )   ( 0.8639,  0.8939 )
## Calculations and Intervals on Original Scale
##
## item 5  -  NT_AD_LEXICAL
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
```

```
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.8950,  0.9194 )   ( 0.8952,  0.9197 )
##
## Level     Percentile            BCa
## 95%   ( 0.8947,  0.9193 )   ( 0.8943,  0.9189 )
## Calculations and Intervals on Original Scale
##
## item 6  -  NT_PRONUNCIA
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.7728,  0.8248 )   ( 0.7736,  0.8259 )
##
## Level     Percentile            BCa
## 95%   ( 0.7719,  0.8242 )   ( 0.7704,  0.8230 )
## Calculations and Intervals on Original Scale
##
## item 7  -  NT_ENTREVISTADOR
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = Celpe.boot, type = c("norm",
"basic",
##     "perc", "bca"), index = i)
```

```
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.9414,  0.9582 )   ( 0.9419,  0.9589 )
##
## Level     Percentile            BCa
## 95%   ( 0.9401,  0.9571 )   ( 0.9404,  0.9573 )
## Calculations and Intervals on Original Scale
```

ANEXO 5: análise TRI

```
  Notas.Oral <-
  readr::read_delim(
    "Demanda_Laura_Notas da parte Oral_Celpe_Bras_2016.1.csv",
    ";", escape_double = FALSE,
    col_types =
      cols(
        `NT_AD_GRAMATICAL` = col_integer(),
        `NT_AD_LEXICAL` = col_integer(),
        `NT_COMPETENCIA` = col_integer(),
        `NT_COMPREENSAO` = col_integer(),
        `NT_FLUENCIA` = col_integer(),
        `NT_PRONUNCIA` = col_integer()
        ),
     trim_ws = TRUE, skip = 1)

  Notas.Oral %>%
  mutate(
    NT_COMPREENSAO = if_else(NT_COMPREENSAO == 0, 1L,
NT_COMPREENSAO) - 1,
    NT_COMPETENCIA = if_else(NT_COMPETENCIA == 0, 1L,
NT_COMPETENCIA) - 1,
    NT_FLUENCIA = if_else(NT_FLUENCIA == 0, 1L, NT_FLUENCIA) -
1,
    NT_AD_GRAMATICAL = if_else(NT_AD_GRAMATICAL == 0, 1L,
NT_AD_GRAMATICAL) - 1,
    NT_AD_LEXICAL = if_else(NT_AD_LEXICAL == 0, 1L,
NT_AD_LEXICAL) - 1,
    NT_PRONUNCIA = if_else(NT_PRONUNCIA == 0, 1L, NT_PRONUNCIA)
- 1,
    NT_ENTREVISTADOR = if_else(NT_ENTREVISTADOR == 0,  1,
NT_ENTREVISTADOR) - 1
  ) %>%
  select(
    NT_COMPREENSAO, NT_COMPETENCIA, NT_FLUENCIA,
    NT_AD_GRAMATICAL, NT_AD_LEXICAL, NT_PRONUNCIA,
```

```
    NT_ENTREVISTADOR
  ) -> Notas.Oral.pcm
```

Partial Credit Model

```
   Celpe.pcm.fit <- eRm::PCM(X = Notas.Oral.pcm, se = TRUE)
Celpe.pcm.fit.Thresholds <- eRm::thresholds(Celpe.pcm.fit)
summary(Celpe.pcm.fit)

   ##
## Results of PCM estimation:
##
## Call:  eRm::PCM(X = Notas.Oral.pcm, se = TRUE)
##
## Conditional log-likelihood: -3318.609
## Number of iterations: 50
## Number of parameters: 27
##
## Item (Category) Difficulty Parameters (eta): with 0.95 CI:
##                      Estimate Std. Error lower CI upper CI
## NT_COMPREENSAO.c2      -7.415      0.748   -8.882   -5.949
## NT_COMPREENSAO.c3      -7.284      0.704   -8.664   -5.905
## NT_COMPREENSAO.c4      -5.388      0.657   -6.676   -4.101
## NT_COMPETENCIA.c1      -3.093      0.386   -3.849   -2.337
## NT_COMPETENCIA.c2      -3.426      0.384   -4.178   -2.674
## NT_COMPETENCIA.c3      -1.445      0.382   -2.194   -0.696
## NT_COMPETENCIA.c4       2.891      0.408    2.091    3.690
## NT_FLUENCIA.c1         -2.198      0.305   -2.795   -1.601
## NT_FLUENCIA.c2         -2.097      0.316   -2.716   -1.479
## NT_FLUENCIA.c3          0.473      0.342   -0.197    1.144
## NT_FLUENCIA.c4          5.332      0.399    4.551    6.113
## NT_AD_GRAMATICAL.c1    -1.518      0.224   -1.957   -1.080
## NT_AD_GRAMATICAL.c2    -0.167      0.258   -0.673    0.339
## NT_AD_GRAMATICAL.c3     3.795      0.321    3.166    4.423
## NT_AD_GRAMATICAL.c4    11.047      0.427   10.210   11.883
## NT_AD_LEXICAL.c1       -1.811      0.241   -2.284   -1.339
## NT_AD_LEXICAL.c2       -0.556      0.269   -1.083   -0.029
## NT_AD_LEXICAL.c3        3.232      0.324    2.597    3.866
```

```
## NT_AD_LEXICAL.c4      10.485      0.424    9.654   11.317
## NT_PRONUNCIA.c1       -2.471      0.297   -3.053   -1.888
## NT_PRONUNCIA.c2       -1.724      0.309   -2.329   -1.119
## NT_PRONUNCIA.c3        1.223      0.338    0.561    1.886
## NT_PRONUNCIA.c4        7.975      0.411    7.169    8.781
## NT_ENTREVISTADOR.c1   -3.152      0.377   -3.891   -2.412
## NT_ENTREVISTADOR.c2   -3.302      0.372   -4.031   -2.573
## NT_ENTREVISTADOR.c3   -0.419      0.374   -1.152    0.313
## NT_ENTREVISTADOR.c4    6.384      0.420    5.562    7.207
##
## Item Easiness Parameters (beta) with 0.95 CI:
##                           Estimate Std. Error lower CI upper
CI
## beta NT_COMPREENSAO.c1      5.369      0.744    3.911
6.826
## beta NT_COMPREENSAO.c2      7.415      0.748    5.949
8.882
## beta NT_COMPREENSAO.c3      7.284      0.704    5.905
8.664
## beta NT_COMPREENSAO.c4      5.388      0.657    4.101
6.676
## beta NT_COMPETENCIA.c1      3.093      0.386    2.337
3.849
## beta NT_COMPETENCIA.c2      3.426      0.384    2.674
4.178
## beta NT_COMPETENCIA.c3      1.445      0.382    0.696
2.194
## beta NT_COMPETENCIA.c4     -2.891      0.408   -3.690
-2.091
## beta NT_FLUENCIA.c1         2.198      0.305    1.601
2.795
## beta NT_FLUENCIA.c2         2.097      0.316    1.479
2.716
## beta NT_FLUENCIA.c3        -0.473      0.342   -1.144
0.197
## beta NT_FLUENCIA.c4        -5.332      0.399   -6.113
-4.551
```

```
## beta NT_AD_GRAMATICAL.c1    1.518      0.224    1.080
1.957
## beta NT_AD_GRAMATICAL.c2    0.167      0.258   -0.339
0.673
## beta NT_AD_GRAMATICAL.c3   -3.795      0.321   -4.423
-3.166
## beta NT_AD_GRAMATICAL.c4  -11.047      0.427  -11.883
-10.210
## beta NT_AD_LEXICAL.c1       1.811      0.241    1.339
2.284
## beta NT_AD_LEXICAL.c2       0.556      0.269    0.029
1.083
## beta NT_AD_LEXICAL.c3      -3.232      0.324   -3.866
-2.597
## beta NT_AD_LEXICAL.c4     -10.485      0.424  -11.317
-9.654
## beta NT_PRONUNCIA.c1        2.471      0.297    1.888
3.053
## beta NT_PRONUNCIA.c2        1.724      0.309    1.119
2.329
## beta NT_PRONUNCIA.c3       -1.223      0.338   -1.886
-0.561
## beta NT_PRONUNCIA.c4       -7.975      0.411   -8.781
-7.169
## beta NT_ENTREVISTADOR.c1    3.152      0.377    2.412
3.891
## beta NT_ENTREVISTADOR.c2    3.302      0.372    2.573
4.031
## beta NT_ENTREVISTADOR.c3    0.419      0.374   -0.313
1.152
## beta NT_ENTREVISTADOR.c4   -6.384      0.420   -7.207
-5.562
```

Thresholds

```
summary(Celpe.pcm.fit.Thresholds)
```

```
  ##
##                                 Estimate Std. Err.    2.5 %
97.5 %
## thresh beta NT_COMPREENSAO.c1   -5.36863   0.74373 -6.82632
-3.91094
## thresh beta NT_COMPREENSAO.c2   -2.04645   0.30933 -2.65273
-1.44017
## thresh beta NT_COMPREENSAO.c3    0.13065   0.17767 -0.21757
0.47888
## thresh beta NT_COMPREENSAO.c4    1.89594   0.14496  1.61182
2.18006
## thresh beta NT_COMPETENCIA.c1   -3.09324   0.38573 -3.84925
-2.33723
## thresh beta NT_COMPETENCIA.c2   -0.33281   0.18227 -0.69005
0.02444
## thresh beta NT_COMPETENCIA.c3    1.98099   0.14694  1.69299
2.26899
## thresh beta NT_COMPETENCIA.c4    4.33582   0.15908  4.02403
4.64761
## thresh beta NT_FLUENCIA.c1      -2.19776   0.30472 -2.79500
-1.60052
## thresh beta NT_FLUENCIA.c2       0.10041   0.16942 -0.23166
0.43247
## thresh beta NT_FLUENCIA.c3       2.57085   0.14678  2.28317
2.85853
## thresh beta NT_FLUENCIA.c4       4.85844   0.16606  4.53296
5.18392
## thresh beta NT_AD_GRAMATICAL.c1 -1.51838   0.22381 -1.95705
-1.07972
## thresh beta NT_AD_GRAMATICAL.c2  1.35099   0.14679  1.06329
1.63869
## thresh beta NT_AD_GRAMATICAL.c3  3.96201   0.15408  3.66001
4.26401
## thresh beta NT_AD_GRAMATICAL.c4  7.25199   0.20586  6.84850
7.65548
## thresh beta NT_AD_LEXICAL.c1    -1.81115   0.24110 -2.28370
-1.33860
```

```
## thresh beta NT_AD_LEXICAL.c2     1.25494   0.14744  0.96597
1.54390
## thresh beta NT_AD_LEXICAL.c3     3.78803   0.15255  3.48905
4.08702
## thresh beta NT_AD_LEXICAL.c4     7.25356   0.20449  6.85277
7.65436
## thresh beta NT_PRONUNCIA.c1     -2.47080   0.29720 -3.05331
-1.88829
## thresh beta NT_PRONUNCIA.c2      0.74683   0.15561  0.44183
1.05182
## thresh beta NT_PRONUNCIA.c3      2.94735   0.14696  2.65930
3.23539
## thresh beta NT_PRONUNCIA.c4      6.75173   0.18914  6.38103
7.12243
## thresh beta NT_ENTREVISTADOR.c1 -3.15182   0.37737 -3.89144
-2.41219
## thresh beta NT_ENTREVISTADOR.c2 -0.15039   0.16720 -0.47811
0.17732
## thresh beta NT_ENTREVISTADOR.c3  2.88274   0.14338  2.60172
3.16376
## thresh beta NT_ENTREVISTADOR.c4  6.80387   0.18966  6.43215
7.17559
```

```
  ifit <- itemfit(p.res)
ifit
```

```
  ##
## Itemfit Statistics:
##                     Chisq  df p-value Outfit MSQ Infit MSQ
Outfit t
## NT_COMPREENSAO   2113.120 898   0.000      2.351     1.386
4.22
## NT_COMPETENCIA   1172.627 898   0.000      1.304     1.114
3.67
## NT_FLUENCIA       652.512 898   1.000      0.726     0.737
-4.72
## NT_AD_GRAMATICAL  672.765 898   1.000      0.748     0.744
-5.65
```

```
## NT_AD_LEXICAL     561.217 898   1.000      0.624     0.605
-9.00
## NT_PRONUNCIA     1041.879 898   0.001      1.159     1.132
3.16
## NT_ENTREVISTADOR  378.713 898   1.000      0.421     0.431
-15.85
##                   Infit t
## NT_COMPREENSAO      5.61
## NT_COMPETENCIA      2.20
## NT_FLUENCIA        -5.88
## NT_AD_GRAMATICAL   -5.96
## NT_AD_LEXICAL      -9.80
## NT_PRONUNCIA        2.71
## NT_ENTREVISTADOR  -15.74
```

```
p.IC <- eRm::IC(p.res)
```

```
p.IC
```

```
##
## Information Criteria:
##                         value npar       AIC       BIC
cAIC
## joint log-lik       -4460.245   54  9028.490  9287.759
9341.759
## marginal log-lik    -6364.368   27 12782.736 12915.246
12942.246
## conditional log-lik -3318.609   27  6691.217  6823.727
6850.727
```

```
eRm::MLoef(Celpe.pcm.fit)
```

```
##
## Martin-Loef-Test (split criterion: median)
## LR-value: 501.595
## Chi-square df: 191
## p-value: 0
```

tem info

```
  eRm::plotINFO(Celpe.pcm.fit, type="item", theta =seq(-10,10,
length.out = 100000L))

  p.fit <- personfit(p.res)
## Personfit Statistics:
##         Chisq df p-value Outfit MSQ Infit MSQ Outfit t Infit
t
## P1      3.372  6   0.761      0.482     0.854     1.03
0.02
## P2      3.372  6   0.761      0.482     0.854     1.03
0.02
## P4      2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P5      6.118  6   0.410      0.874     0.920     0.16
0.08
## P6      5.317  6   0.504      0.760     1.090     1.14
0.36
## P7      5.420  6   0.491      0.774     0.817    -0.31
-0.21
## P8      3.946  6   0.684      0.564     0.846     0.50
-0.19
## P9      5.220  6   0.516      0.746     0.915    -0.12
0.05
## P10     4.250  6   0.643      0.607     0.641    -0.72
-0.63
## P12     4.479  6   0.612      0.640     0.678    -0.54
-0.51
## P13     4.990  6   0.545      0.713     0.747    -0.43
-0.36
## P14     7.962  6   0.241      1.137     1.261     0.45
0.59
## P15     2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P16     3.372  6   0.761      0.482     0.854     1.03
0.02
## P17     3.372  6   0.761      0.482     0.854     1.03
0.02
## P18     4.198  6   0.650      0.600     0.674    -0.73
```

```
-0.54
## P19     6.136  6   0.408      0.877     0.827    -0.08
-0.18
## P21     6.345  6   0.386      0.906     0.899    -0.01
-0.02
## P22     6.118  6   0.410      0.874     0.920     0.16
0.08
## P23    21.948  6   0.001      3.135     1.664     1.37
1.34
## P24     4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P25     3.970  6   0.681      0.567     0.852     0.50
-0.18
## P26    10.667  6   0.099      1.524     1.648     1.05
1.22
## P27    25.227  6   0.000      3.604     3.312     3.28
3.02
## P28     2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P29     1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P31     1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P32     5.014  6   0.542      0.716     0.807     0.10
-0.10
## P33     2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P34     1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P36    15.537  6   0.016      2.220     2.547     1.89
2.22
## P37     6.040  6   0.419      0.863     0.905    -0.10
-0.02
## P38     3.086  6   0.798      0.441     0.446    -1.21
-1.20
## P39     2.024  6   0.918      0.289     0.271    -1.74
-1.85
```

```
## P40    18.113  6   0.006      2.588     2.831     1.54
2.16
## P41     6.312  6   0.389      0.902     0.749    -0.02
-0.35
## P43    21.870  6   0.001      3.124     3.258     2.86
2.99
## P45     2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P47     7.758  6   0.256      1.108     1.118     0.57
0.39
## P48     5.014  6   0.542      0.716     0.807     0.10
-0.10
## P49     6.828  6   0.337      0.975     0.627     0.14
-0.64
## P50    10.246  6   0.115      1.464     1.676     0.75
1.10
## P52     1.507  6   0.959      0.215     0.223    -2.04
-2.00
## P53     1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P55     6.331  6   0.387      0.904     0.974    -0.02
0.12
## P56     4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P57     5.552  6   0.475      0.793     0.854    -0.26
-0.12
## P58    14.013  6   0.029      2.002     2.126     1.69
1.84
## P59     2.962  6   0.814      0.423     0.635     0.39
-0.74
## P60     2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P62     3.372  6   0.761      0.482     0.854     1.03
0.02
## P63     6.710  6   0.348      0.959     0.893     0.09
-0.04
## P64     1.470  6   0.961      0.210     0.255    -1.05
```

```
-1.51
## P65     1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P66     3.312  6   0.769      0.473     0.536    -0.83
-0.84
## P67     1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P68    17.237  6   0.008      2.462     1.144     1.65
0.43
## P69     3.970  6   0.681      0.567     0.852     0.50
-0.18
## P70     2.962  6   0.814      0.423     0.635     0.39
-0.74
## P71     1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P72     2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P73     2.894  6   0.822      0.413     0.427    -1.33
-1.27
## P74     7.489  6   0.278      1.070     1.205     0.31
0.54
## P75     3.571  6   0.734      0.510     0.522    -0.98
-0.94
## P76     3.970  6   0.681      0.567     0.852     0.50
-0.18
## P77     6.147  6   0.407      0.878     0.949    -0.06
0.07
## P78     4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P79     3.689  6   0.719      0.527     0.710     0.12
-0.40
## P80    15.320  6   0.018      2.189     2.347     1.90
2.09
## P81     4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P82     3.312  6   0.769      0.473     0.536    -0.83
-0.84
```

```
## P83     7.793  6   0.254      1.113     1.181     0.47
0.48
## P84     3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P85     7.342  6   0.290      1.049     1.181     0.29
0.49
## P86    15.702  6   0.015      2.243     1.907     1.89
1.51
## P87     5.317  6   0.504      0.760     1.090     1.14
0.36
## P88     3.372  6   0.761      0.482     0.854     1.03
0.02
## P89     2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P90     5.894  6   0.435      0.842     1.037     0.23
0.27
## P91     2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P94    10.555  6   0.103      1.508     1.505     1.03
1.02
## P95     1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P96    13.073  6   0.042      1.868     1.938     1.51
1.60
## P97     4.195  6   0.650      0.599     0.612    -0.74
-0.71
## P98     2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P99     4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P100    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P101   10.344  6   0.111      1.478     1.530     0.96
1.05
## P102  210.255  6   0.000     30.036     1.514     3.61
1.11
## P103    4.250  6   0.643      0.607     0.641    -0.72
```

```
-0.63
## P104    5.019  6   0.541      0.717     0.751    -0.43
-0.35
## P105    9.012  6   0.173      1.287     1.517     0.61
0.91
## P106    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P107    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P108    8.352  6   0.213      1.193     1.095     0.52
0.35
## P109    2.694  6   0.846      0.385     0.418    -1.39
-1.27
## P110    8.860  6   0.182      1.266     1.364     0.65
0.81
## P111    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P112    4.134  6   0.659      0.591     0.614    -0.77
-0.70
## P113    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P114    2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P115   11.845  6   0.066      1.692     1.794     1.29
1.43
## P116    2.846  6   0.828      0.407     0.403    -1.33
-1.35
## P117   15.488  6   0.017      2.213     2.379     1.94
2.12
## P118    6.852  6   0.335      0.979     0.864     0.14
-0.08
## P119    3.372  6   0.761      0.482     0.854     1.03
0.02
## P120    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P121   12.271  6   0.056      1.753     1.416     0.91
0.85
```

```
## P122    5.317  6   0.504      0.760     1.090     1.14
0.36
## P123    3.615  6   0.729      0.516     0.560    -0.97
-0.84
## P125    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P126    5.374  6   0.497      0.768     0.747    -0.32
-0.37
## P127    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P128    4.166  6   0.654      0.595     0.710    -0.23
-0.29
## P129    6.118  6   0.410      0.874     0.920     0.16
0.08
## P130    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P131    7.846  6   0.250      1.121     1.312     0.39
0.71
## P132    4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P133   13.457  6   0.036      1.922     2.023     1.60
1.72
## P134    7.567  6   0.272      1.081     1.088     0.32
0.34
## P135    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P136    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P137    3.261  6   0.775      0.466     0.573    -0.54
-0.50
## P138    6.118  6   0.410      0.874     0.920     0.16
0.08
## P139   10.558  6   0.103      1.508     1.570     1.03
1.12
## P140    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P141    6.136  6   0.408      0.877     0.827    -0.08
```

```
-0.18
## P142    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P143   10.165  6   0.118      1.452     1.369     0.93
0.81
## P144    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P145    3.235  6   0.779      0.462     0.475    -1.11
-1.07
## P146   12.696  6   0.048      1.814     2.033     0.96
1.45
## P148    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P149    5.014  6   0.542      0.716     0.807     0.10
-0.10
## P150    4.074  6   0.667      0.582     0.872     0.51
-0.13
## P151    7.057  6   0.316      1.008     1.269     0.38
0.60
## P152    1.507  6   0.959      0.215     0.223    -2.04
-2.00
## P153    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P154   51.394  6   0.000      7.342     1.425     2.89
0.80
## P155    3.009  6   0.808      0.430     0.478    -1.07
-1.02
## P156   11.318  6   0.079      1.617     1.571     1.16
1.11
## P157    2.619  6   0.855      0.374     0.377    -1.41
-1.40
## P158    2.504  6   0.868      0.358     0.381    -1.44
-1.36
## P159    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P160   99.097  6   0.000     14.157     1.355     3.32
0.76
```

```
## P161    3.968  6   0.681      0.567     0.589    -0.84
-0.77
## P162    4.160  6   0.655      0.594     0.728    -0.39
-0.32
## P163    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P164    6.569  6   0.363      0.938     0.992     0.06
0.16
## P167    8.320  6   0.216      1.189     1.232     0.52
0.59
## P168    4.624  6   0.593      0.661     0.608    -0.58
-0.71
## P169    9.813  6   0.133      1.402     1.586     0.84
1.11
## P170    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P171    2.809  6   0.832      0.401     0.482    -0.80
-0.91
## P172   30.464  6   0.000      4.352     4.581     3.76
3.92
## P173    6.040  6   0.419      0.863     0.905    -0.10
-0.02
## P174    3.693  6   0.718      0.528     0.511    -0.96
-1.00
## P175    3.970  6   0.681      0.567     0.852     0.50
-0.18
## P176    5.093  6   0.532      0.728     0.747    -0.42
-0.37
## P177    2.694  6   0.846      0.385     0.418    -1.39
-1.27
## P178    5.552  6   0.475      0.793     0.854    -0.26
-0.12
## P179    2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P180    2.809  6   0.832      0.401     0.482    -0.80
-0.91
## P182    5.317  6   0.504      0.760     1.090     1.14
```

```
0.36
## P183    4.245  6   0.644      0.606     0.666    -0.68
-0.53
## P184    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P185   57.261  6   0.000      8.180     2.219     3.08
1.63
## P186    3.372  6   0.761      0.482     0.854     1.03
0.02
## P187    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P189    2.707  6   0.845      0.387     0.428    -1.41
-1.25
## P191    5.655  6   0.463      0.808     0.843    -0.23
-0.15
## P192    4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P193    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P194   10.197  6   0.117      1.457     1.546     0.95
1.08
## P195    5.014  6   0.542      0.716     0.807     0.10
-0.10
## P196    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P197    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P198    6.770  6   0.343      0.967     1.009     0.11
0.19
## P199   10.558  6   0.103      1.508     1.570     1.03
1.12
## P200    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P201    8.292  6   0.217      1.185     1.256     0.51
0.63
## P202   20.192  6   0.003      2.885     3.150     1.71
2.40
```

```
## P204    3.372  6   0.761      0.482     0.854     1.03
0.02
## P205    2.046  6   0.915      0.292     0.319    -1.37
-1.54
## P206    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P207   11.561  6   0.073      1.652     1.712     1.23
1.31
## P208   13.794  6   0.032      1.971     1.972     1.66
1.65
## P209    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P210    4.748  6   0.576      0.678     0.690    -0.54
-0.51
## P211    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P212    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P213    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P216    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P217    5.166  6   0.523      0.738     0.914    -0.24
0.02
## P218    6.740  6   0.346      0.963     0.980     0.10
0.13
## P219    9.678  6   0.139      1.383     1.447     0.83
0.93
## P220    3.372  6   0.761      0.482     0.854     1.03
0.02
## P221    3.177  6   0.786      0.454     0.473    -1.13
-1.09
## P222    3.767  6   0.708      0.538     0.515    -0.90
-0.96
## P223    7.962  6   0.241      1.137     1.261     0.45
0.59
## P224   10.423  6   0.108      1.489     1.557     0.94
```

```
1.08
## P225    4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P226    3.136  6   0.792      0.448     0.411    -1.18
-1.30
## P227    5.433  6   0.490      0.776     0.768    -0.31
-0.32
## P228   15.861  6   0.015      2.266     1.937     2.00
1.60
## P229    4.198  6   0.650      0.600     0.674    -0.73
-0.54
## P231    4.309  6   0.635      0.616     0.627    -0.70
-0.67
## P232    6.494  6   0.370      0.928     0.989     0.22
0.20
## P233    3.261  6   0.775      0.466     0.573    -0.54
-0.50
## P234    3.645  6   0.725      0.521     0.513    -0.93
-0.95
## P235    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P236    3.689  6   0.719      0.527     0.710     0.12
-0.40
## P237    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P238    7.846  6   0.250      1.121     1.312     0.39
0.71
## P239    6.908  6   0.329      0.987     1.025     0.15
0.22
## P240    5.317  6   0.504      0.760     1.090     1.14
0.36
## P241    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P242    5.326  6   0.503      0.761     0.753    -0.34
-0.36
## P243    3.812  6   0.702      0.545     0.681    -0.48
-0.42
```

```
## P244    3.813  6   0.702      0.545     0.485    -0.83
-1.01
## P246    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P247    5.894  6   0.435      0.842     1.037     0.23
0.27
## P248    5.517  6   0.479      0.788     1.227     0.00
0.54
## P249    4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P250    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P251    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P252    2.991  6   0.810      0.427     0.429    -1.27
-1.25
## P253   78.247  6   0.000     11.178     3.887     2.42
3.82
## P254    7.342  6   0.290      1.049     1.181     0.29
0.49
## P255   13.420  6   0.037      1.917     1.377     1.06
0.88
## P256    6.962  6   0.324      0.995     1.184     0.17
0.50
## P257   12.446  6   0.053      1.778     1.849     1.37
1.48
## P258    5.317  6   0.504      0.760     1.090     1.14
0.36
## P259    3.968  6   0.681      0.567     0.589    -0.84
-0.77
## P260    1.483  6   0.961      0.212     0.203    -2.17
-2.23
## P261    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P263    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P264    3.813  6   0.702      0.545     0.485    -0.83
```

```
-1.01
## P265   10.746  6   0.097      1.535     1.592     0.78
0.99
## P266   13.457  6   0.036      1.922     2.023     1.60
1.72
## P267    2.694  6   0.846      0.385     0.418    -1.39
-1.27
## P269    3.177  6   0.786      0.454     0.473    -1.13
-1.09
## P270    2.894  6   0.822      0.413     0.427    -1.33
-1.27
## P271    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P272    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P274    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P275    2.809  6   0.832      0.401     0.482    -0.80
-0.91
## P276   10.382  6   0.109      1.483     1.634     0.96
1.17
## P277   29.633  6   0.000      4.233     1.963     1.74
1.14
## P278    3.372  6   0.761      0.482     0.854     1.03
0.02
## P279    4.379  6   0.626      0.626     0.603    -0.79
-0.82
## P280    5.204  6   0.518      0.743     0.749    -0.13
-0.27
## P281    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P282    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P283    4.374  6   0.626      0.625     0.663    -0.65
-0.56
## P284    5.589  6   0.471      0.798     0.837    -0.25
-0.16
```

```
## P285    9.621  6   0.142      1.374     1.463     0.82
0.96
## P286    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P287   10.673  6   0.099      1.525     1.683     1.03
1.25
## P288    2.905  6   0.821      0.415     0.429    -1.31
-1.26
## P289    4.479  6   0.612      0.640     0.678    -0.54
-0.51
## P290   12.284  6   0.056      1.755     1.725     1.37
1.33
## P291    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P292    5.166  6   0.523      0.738     0.914    -0.24
0.02
## P295    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P296    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P297    6.482  6   0.371      0.926     1.093     0.03
0.35
## P298    8.712  6   0.190      1.245     0.908     0.59
0.00
## P299    4.195  6   0.650      0.599     0.612    -0.74
-0.71
## P300    8.472  6   0.206      1.210     1.536     0.55
0.93
## P301    8.885  6   0.180      1.269     1.343     0.65
0.77
## P302    6.118  6   0.410      0.874     0.920     0.16
0.08
## P303    5.907  6   0.434      0.844     0.901    -0.19
-0.05
## P304    2.619  6   0.855      0.374     0.377    -1.41
-1.40
## P305   13.563  6   0.035      1.938     2.205     1.57
```

```
1.91
## P306    6.207  6   0.400      0.887     0.948     0.09
0.11
## P307   11.817  6   0.066      1.688     1.803     0.88
1.22
## P309    3.813  6   0.702      0.545     0.485    -0.83
-1.01
## P310    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P311    7.585  6   0.270      1.084     1.163     0.33
0.47
## P312    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P313   99.097  6   0.000     14.157     1.355     3.32
0.76
## P314    3.968  6   0.681      0.567     0.589    -0.84
-0.77
## P315    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P316    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P317    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P318    4.129  6   0.659      0.590     0.571    -0.72
-0.77
## P319   12.157  6   0.059      1.737     1.717     1.32
1.30
## P320    5.198  6   0.519      0.743     0.746    -0.46
-0.43
## P321    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P322    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P323    7.384  6   0.287      1.055     1.291     0.28
0.68
## P324    4.250  6   0.643      0.607     0.641    -0.72
-0.63
```

```
## P325   49.806  6   0.000      7.115     1.517     1.97
0.86
## P326    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P327    9.401  6   0.152      1.343     1.345     0.75
0.76
## P328    3.689  6   0.719      0.527     0.710     0.12
-0.40
## P329    3.624  6   0.727      0.518     0.518    -0.98
-0.98
## P330    7.192  6   0.303      1.027     1.278     0.40
0.61
## P331    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P332   12.268  6   0.056      1.753     1.690     1.37
1.29
## P333    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P334    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P335    6.118  6   0.410      0.874     0.920     0.16
0.08
## P336    5.635  6   0.465      0.805     0.824    -0.23
-0.19
## P337   44.434  6   0.000      6.348     3.736     3.13
2.81
## P338    2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P339    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P340   12.751  6   0.047      1.822     1.988     1.14
1.51
## P341    5.220  6   0.516      0.746     0.915    -0.12
0.05
## P342    9.929  6   0.128      1.418     1.535     0.74
0.98
## P343    1.775  6   0.939      0.254     0.248    -1.98
```

```
-2.00
## P344   15.488  6   0.017      2.213     2.379     1.94
2.12
## P345    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P346    3.261  6   0.775      0.466     0.573    -0.54
-0.50
## P347    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P349    5.220  6   0.516      0.746     0.915    -0.12
0.05
## P350   14.295  6   0.027      2.042     2.127     1.72
1.81
## P351    2.726  6   0.842      0.389     0.407    -1.41
-1.34
## P352    9.144  6   0.166      1.306     1.363     0.68
0.79
## P353    2.448  6   0.874      0.350     0.407    -1.45
-1.25
## P354    6.631  6   0.356      0.947     0.978     0.07
0.13
## P355    7.758  6   0.256      1.108     1.118     0.57
0.39
## P357    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P358    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P359    5.553  6   0.475      0.793     0.848    -0.31
-0.17
## P360    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P361    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P362    4.071  6   0.667      0.582     0.592    -0.78
-0.75
## P363    5.339  6   0.501      0.763     0.770    -0.33
-0.31
```

```
## P364    5.030  6   0.540      0.719     0.682    -0.41
-0.50
## P365    5.481  6   0.484      0.783     0.779    -0.28
-0.29
## P366    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P367    3.968  6   0.681      0.567     0.589    -0.84
-0.77
## P368    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P369    3.098  6   0.796      0.443     0.468    -1.17
-1.09
## P370    7.267  6   0.297      1.038     1.135     0.30
0.42
## P372    4.134  6   0.659      0.591     0.614    -0.77
-0.70
## P374    6.118  6   0.410      0.874     0.920     0.16
0.08
## P375    6.034  6   0.419      0.862     0.920    -0.11
0.01
## P376    9.147  6   0.165      1.307     1.340     0.70
0.76
## P378    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P379   21.948  6   0.001      3.135     1.664     1.37
1.34
## P380    5.227  6   0.515      0.747     0.820    -0.36
-0.20
## P381    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P382    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P385   16.177  6   0.013      2.311     0.957     1.54
0.12
## P386    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P387    6.345  6   0.386      0.906     0.899    -0.01
```

```
-0.02
## P388    7.469  6   0.280      1.067     0.951     0.30
0.08
## P389    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P390    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P391   19.883  6   0.003      2.840     1.378     1.91
0.77
## P392    7.089  6   0.313      1.013     1.049     0.20
0.27
## P393    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P394    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P395    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P396    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P397    1.024  6   0.985      0.146     0.306    -0.46
-0.78
## P398    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P399    6.482  6   0.371      0.926     1.093     0.03
0.35
## P400    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P401    5.317  6   0.504      0.760     1.090     1.14
0.36
## P402    4.074  6   0.667      0.582     0.872     0.51
-0.13
## P403    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P404    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P405    3.970  6   0.681      0.567     0.852     0.50
-0.18
```

```
## P406    4.719  6   0.580      0.674     0.764    -0.46
-0.31
## P407   11.761  6   0.068      1.680     1.831     0.88
1.25
## P408    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P409    3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P410    7.919  6   0.244      1.131     1.110     0.42
0.38
## P411    3.968  6   0.681      0.567     0.589    -0.84
-0.77
## P412    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P413   10.959  6   0.090      1.566     1.941     0.89
1.46
## P414    5.894  6   0.435      0.842     1.037     0.23
0.27
## P416   15.980  6   0.014      2.283     2.125     1.93
1.76
## P417    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P418    4.719  6   0.580      0.674     0.764    -0.46
-0.31
## P419   27.530  6   0.000      3.933     4.085     3.46
3.57
## P420    3.261  6   0.775      0.466     0.573    -0.54
-0.50
## P421    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P422    5.552  6   0.475      0.793     0.854    -0.26
-0.12
## P423    4.198  6   0.650      0.600     0.674    -0.73
-0.54
## P424    4.502  6   0.609      0.643     0.648    -0.63
-0.61
## P425    8.684  6   0.192      1.241     1.298     0.60
```

```
0.69
## P426   50.379  6   0.000      7.197     1.186     2.85
0.49
## P427    3.372  6   0.761      0.482     0.854     1.03
0.02
## P428    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P431   17.124  6   0.009      2.446     2.656     2.13
2.33
## P432    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P433   13.141  6   0.041      1.877     1.971     1.53
1.64
## P434    4.270  6   0.640      0.610     0.735    -0.50
-0.35
## P437    4.508  6   0.608      0.644     0.655    -0.63
-0.60
## P438    4.479  6   0.612      0.640     0.678    -0.54
-0.51
## P439    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P440   12.406  6   0.053      1.772     1.886     1.00
1.32
## P441    5.998  6   0.423      0.857     0.958    -0.08
0.10
## P443    3.754  6   0.710      0.536     0.670    -0.32
-0.38
## P444    4.878  6   0.560      0.697     0.798     0.07
-0.12
## P445    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P448    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P449    7.450  6   0.281      1.064     1.082     0.29
0.33
## P450    4.499  6   0.609      0.643     0.475    -0.60
-1.09
```

```
## P451    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P452    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P453    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P454    6.454  6   0.374      0.922     0.927     0.01
0.03
## P455    7.384  6   0.287      1.055     1.291     0.28
0.68
## P456    5.204  6   0.518      0.743     0.749    -0.13
-0.27
## P457   13.047  6   0.042      1.864     1.892     1.48
1.52
## P458   13.037  6   0.042      1.862     1.918     1.50
1.56
## P459    5.317  6   0.504      0.760     1.090     1.14
0.36
## P460    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P461    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P462    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P463    5.894  6   0.435      0.842     1.037     0.23
0.27
## P464    4.198  6   0.650      0.600     0.674    -0.73
-0.54
## P465    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P466    8.711  6   0.191      1.244     1.437     0.59
0.90
## P467    7.550  6   0.273      1.079     1.221     0.39
0.54
## P468   10.989  6   0.089      1.570     1.667     1.12
1.26
## P469    5.284  6   0.508      0.755     0.766    -0.35
```

```
-0.33
## P470    2.670  6   0.849      0.381     0.371    -1.43
-1.46
## P471    6.778  6   0.342      0.968     0.942     0.35
0.12
## P473    4.508  6   0.608      0.644     0.655    -0.63
-0.60
## P474    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P475    3.372  6   0.761      0.482     0.854     1.03
0.02
## P476    6.672  6   0.352      0.953     0.979     0.08
0.13
## P477    5.949  6   0.429      0.850     1.009     0.24
0.23
## P478    5.517  6   0.479      0.788     0.794    -0.28
-0.26
## P479    2.991  6   0.810      0.427     0.429    -1.27
-1.25
## P481   62.029  6   0.000      8.861     3.062     3.22
2.32
## P483    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P484    4.270  6   0.640      0.610     0.735    -0.50
-0.35
## P485    2.991  6   0.810      0.427     0.429    -1.27
-1.25
## P486    2.991  6   0.810      0.427     0.429    -1.27
-1.25
## P487    3.394  6   0.758      0.485     0.856     1.03
0.03
## P489   20.836  6   0.002      2.977     1.427     1.33
0.96
## P490    3.593  6   0.732      0.513     0.518    -0.87
-0.93
## P491    4.938  6   0.552      0.705     0.732    -0.47
-0.40
```

```
## P492   12.274  6   0.056      1.753     1.432     0.91
0.87
## P493    3.372  6   0.761      0.482     0.854     1.03
0.02
## P494   12.369  6   0.054      1.767     1.448     0.92
0.90
## P495    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P496    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P497    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P498   17.411  6   0.008      2.487     2.455     2.25
2.21
## P499    5.655  6   0.463      0.808     0.843    -0.23
-0.15
## P502    4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P503    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P505    4.679  6   0.586      0.668     0.702    -0.54
-0.46
## P506    3.412  6   0.756      0.487     0.538    -1.01
-0.87
## P509    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P510  561.906  6   0.000     80.272     1.564     3.95
0.91
## P511    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P512    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P513    5.317  6   0.504      0.760     1.090     1.14
0.36
## P514   10.264  6   0.114      1.466     1.357     0.97
0.80
## P515    4.270  6   0.640      0.610     0.735    -0.50
```

```
-0.35
## P516    9.313  6   0.157      1.330     1.393     0.80
0.90
## P517    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P518    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P519    8.926  6   0.178      1.275     1.336     0.59
0.71
## P520    3.157  6   0.789      0.451     0.530    -0.68
-0.78
## P521   10.555  6   0.103      1.508     1.505     1.03
1.02
## P522    6.494  6   0.370      0.928     0.989     0.22
0.20
## P523    5.894  6   0.435      0.842     1.037     0.23
0.27
## P524    4.430  6   0.619      0.633     0.677    -0.65
-0.54
## P526    9.807  6   0.133      1.401     1.559     0.74
1.05
## P527    5.144  6   0.525      0.735     0.752    -0.39
-0.34
## P529    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P531    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P532    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P533    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P534    3.372  6   0.761      0.482     0.854     1.03
0.02
## P535    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P536    2.372  6   0.883      0.339     0.373    -1.51
-1.39
```

```
## P537    2.370  6   0.883      0.339     0.313    -1.54
-1.67
## P538    7.585  6   0.270      1.084     1.163     0.33
0.47
## P539   16.938  6   0.010      2.420     2.307     2.18
2.05
## P540    8.711  6   0.191      1.244     1.437     0.59
0.90
## P541    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P542    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P544    4.249  6   0.643      0.607     0.601    -0.73
-0.74
## P546    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P547    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P548    7.615  6   0.268      1.088     1.183     0.36
0.49
## P549    5.321  6   0.503      0.760     0.795    -0.34
-0.26
## P550    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P551    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P552   16.489  6   0.011      2.356     2.528     2.11
2.29
## P553    2.619  6   0.855      0.374     0.377    -1.41
-1.40
## P554    6.494  6   0.370      0.928     0.989     0.22
0.20
## P555    6.129  6   0.409      0.876     1.081     0.01
0.33
## P556    8.860  6   0.182      1.266     1.364     0.65
0.81
## P557    1.046  6   0.984      0.149     0.157    -0.87
```

```
-1.92
## P558    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P559    5.317  6   0.504      0.760     1.090     1.14
0.36
## P560    5.481  6   0.484      0.783     0.779    -0.28
-0.29
## P562    4.634  6   0.592      0.662     0.684    -0.58
-0.52
## P563    3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P564   11.116  6   0.085      1.588     1.348     1.07
0.77
## P565    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P566   49.806  6   0.000      7.115     1.517     1.97
0.86
## P567    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P568    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P569    3.945  6   0.684      0.564     0.570    -0.84
-0.83
## P570    8.712  6   0.190      1.245     0.908     0.59
0.00
## P571    2.707  6   0.845      0.387     0.428    -1.41
-1.25
## P572    5.014  6   0.542      0.716     0.807     0.10
-0.10
## P573    5.220  6   0.516      0.746     0.915    -0.12
0.05
## P574    5.517  6   0.479      0.788     0.794    -0.28
-0.26
## P575    5.317  6   0.504      0.760     1.090     1.14
0.36
## P576    2.046  6   0.915      0.292     0.319    -1.37
-1.54
```

```
## P577    6.312  6   0.389      0.902     0.749    -0.02
-0.35
## P578    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P579    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P580    8.953  6   0.176      1.279     1.297     0.67
0.69
## P581    4.160  6   0.655      0.594     0.728    -0.39
-0.32
## P582   30.216  6   0.000      4.317     4.168     2.05
3.05
## P583   25.795  6   0.000      3.685     0.800     2.11
-0.12
## P584    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P585    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P586    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P588    4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P590    9.012  6   0.173      1.287     1.517     0.61
0.91
## P591   19.191  6   0.004      2.742     2.944     2.39
2.58
## P592    3.384  6   0.759      0.483     0.483    -1.09
-1.09
## P593   10.165  6   0.118      1.452     1.369     0.93
0.81
## P594    5.166  6   0.523      0.738     0.914    -0.24
0.02
## P595    2.809  6   0.832      0.401     0.482    -0.80
-0.91
## P596    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P597    7.340  6   0.291      1.049     1.138     0.27
```

```
0.43
## P598    5.317  6   0.504      0.760     1.090     1.14
0.36
## P599    6.234  6   0.398      0.891     0.959    -0.05
0.09
## P600    3.240  6   0.778      0.463     0.427    -1.16
-1.27
## P601   16.716  6   0.010      2.388     2.387     2.08
2.08
## P602    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P603    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P604    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P606    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P607    4.878  6   0.560      0.697     0.798     0.07
-0.12
## P608    5.589  6   0.471      0.798     0.837    -0.25
-0.16
## P609    4.074  6   0.667      0.582     0.872     0.51
-0.13
## P610    3.813  6   0.702      0.545     0.485    -0.83
-1.01
## P611    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P612    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P613    7.125  6   0.309      1.018     0.675     0.22
-0.52
## P614   10.746  6   0.097      1.535     1.592     0.78
0.99
## P615    2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P616    8.320  6   0.216      1.189     1.232     0.52
0.59
```

```
## P617    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P618    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P619    8.320  6   0.216      1.189     1.232     0.52
0.59
## P620    7.075  6   0.314      1.011     0.833     0.29
-0.02
## P621    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P622   18.388  6   0.005      2.627     2.265     1.26
1.89
## P623    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P624    2.894  6   0.822      0.413     0.427    -1.33
-1.27
## P625    5.446  6   0.488      0.778     0.844    -0.30
-0.15
## P626    9.838  6   0.132      1.405     1.154     0.83
0.45
## P627    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P628    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P629    2.370  6   0.883      0.339     0.313    -1.54
-1.67
## P630    3.394  6   0.758      0.485     0.856     1.03
0.03
## P631    2.558  6   0.862      0.365     0.398    -1.40
-1.28
## P632    5.655  6   0.463      0.808     0.843    -0.23
-0.15
## P633    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P634    6.118  6   0.410      0.874     0.920     0.16
0.08
## P635    7.094  6   0.312      1.013     1.240     0.32
```

```
0.56
## P637    6.584  6   0.361      0.941     0.933     0.06
0.05
## P638    3.693  6   0.718      0.528     0.511    -0.96
-1.00
## P639   13.344  6   0.038      1.906     1.653     0.98
1.18
## P640   11.362  6   0.078      1.623     1.772     1.16
1.35
## P641    7.855  6   0.249      1.122     1.180     0.40
0.50
## P642    3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P643    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P644    7.384  6   0.287      1.055     1.291     0.28
0.68
## P645    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P646    3.946  6   0.684      0.564     0.846     0.50
-0.19
## P647    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P648    5.572  6   0.473      0.796     0.929    -0.23
0.04
## P649    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P650    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P651    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P652    4.261  6   0.641      0.609     0.633    -0.69
-0.63
## P653    4.261  6   0.641      0.609     0.633    -0.69
-0.63
## P654   13.913  6   0.031      1.988     2.010     1.66
1.70
```

```
## P655    3.177  6   0.786      0.454     0.473    -1.13
-1.09
## P656    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P657    6.744  6   0.345      0.963     0.951     0.07
0.05
## P658    2.348  6   0.885      0.335     0.369    -1.22
-1.36
## P659    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P660    4.479  6   0.612      0.640     0.678    -0.54
-0.51
## P661    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P662    7.741  6   0.258      1.106     1.223     0.37
0.58
## P664    6.494  6   0.370      0.928     0.989     0.22
0.20
## P665   15.237  6   0.018      2.177     2.249     1.91
1.98
## P666    3.782  6   0.706      0.540     0.542    -0.88
-0.87
## P667    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P668    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P669   10.989  6   0.089      1.570     1.667     1.12
1.26
## P670    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P671    6.118  6   0.410      0.874     0.920     0.16
0.08
## P673    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P675    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P676    8.085  6   0.232      1.155     1.217     0.46
```

```
0.57
## P677   13.243  6   0.039      1.892     1.918     1.52
1.55
## P678    5.317  6   0.504      0.760     1.090     1.14
0.36
## P679    4.249  6   0.643      0.607     0.601    -0.73
-0.74
## P681    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P682    5.949  6   0.429      0.850     1.009     0.24
0.23
## P683    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P684    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P685    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P687    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P688    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P689    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P690    4.166  6   0.654      0.595     0.710    -0.23
-0.29
## P692    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P694    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P695    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P696    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P697   10.989  6   0.089      1.570     1.667     1.12
1.26
## P698    4.153  6   0.656      0.593     0.598    -0.54
-0.68
```

```
## P699    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P700    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P701    2.504  6   0.868      0.358     0.381    -1.44
-1.36
## P702    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P703   14.511  6   0.024      2.073     2.207     1.77
1.93
## P704    8.535  6   0.201      1.219     1.387     0.56
0.83
## P705    5.204  6   0.518      0.743     0.749    -0.13
-0.27
## P706    1.775  6   0.939      0.254     0.248    -1.98
-2.00
## P707    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P710    1.738  6   0.942      0.248     0.237    -1.98
-2.05
## P711    4.166  6   0.654      0.595     0.710    -0.23
-0.29
## P712    5.174  6   0.522      0.739     0.747    -0.38
-0.36
## P713    3.689  6   0.719      0.527     0.710     0.12
-0.40
## P714    4.812  6   0.568      0.687     0.668    -0.57
-0.60
## P715    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P716    4.074  6   0.667      0.582     0.872     0.51
-0.13
## P718    7.423  6   0.284      1.060     1.085     0.29
0.33
## P719   14.013  6   0.029      2.002     2.209     1.67
1.91
## P720    2.714  6   0.844      0.388     0.505    -0.03
```

```
-0.92
## P721    7.057  6   0.316      1.008     1.269     0.38
0.60
## P722    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P723    3.240  6   0.778      0.463     0.427    -1.16
-1.27
## P724    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P725    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P726    6.740  6   0.346      0.963     0.980     0.10
0.13
## P727    1.470  6   0.961      0.210     0.255    -1.05
-1.51
## P729    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P730   15.174  6   0.019      2.168     0.759     1.43
-0.25
## P732    3.693  6   0.718      0.528     0.725     0.12
-0.37
## P733    5.827  6   0.443      0.832     0.831    -0.18
-0.18
## P735    9.991  6   0.125      1.427     1.506     0.89
1.01
## P736   11.609  6   0.071      1.658     1.762     1.25
1.38
## P737    3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P738    5.533  6   0.478      0.790     0.810    -0.27
-0.22
## P739    1.601  6   0.952      0.229     0.235    -2.10
-2.07
## P741    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P742    2.441  6   0.875      0.349     0.343    -1.57
-1.58
```

```
## P744    3.693  6   0.718      0.528     0.511    -0.96
-1.00
## P745    4.074  6   0.667      0.582     0.872     0.51
-0.13
## P746    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P747    6.432  6   0.377      0.919     0.963     0.02
0.11
## P748    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P749   11.828  6   0.066      1.690     1.913     1.20
1.54
## P750    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P751    3.515  6   0.742      0.502     0.505    -1.03
-1.02
## P752    9.097  6   0.168      1.300     1.285     0.69
0.67
## P753    1.314  6   0.971      0.188     0.177    -2.30
-2.36
## P754   18.433  6   0.005      2.633     2.770     2.29
2.42
## P755    6.482  6   0.371      0.926     1.093     0.03
0.35
## P756   15.174  6   0.019      2.168     0.759     1.43
-0.25
## P757    9.334  6   0.156      1.333     1.451     0.75
0.93
## P758    3.086  6   0.798      0.441     0.446    -1.21
-1.20
## P759    3.296  6   0.771      0.471     0.470    -0.99
-1.08
## P761    5.638  6   0.465      0.805     0.695    -0.21
-0.46
## P762   17.501  6   0.008      2.500     2.740     2.16
2.39
## P763    3.693  6   0.718      0.528     0.725     0.12
```

```
-0.37
## P764   49.806  6   0.000      7.115     1.517     1.97
0.86
## P765    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P766   12.199  6   0.058      1.743     1.756     1.36
1.38
## P767    3.501  6   0.744      0.500     0.432    -1.03
-1.24
## P768    4.422  6   0.620      0.632     0.716    -0.56
-0.42
## P769    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P770    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P771    8.320  6   0.216      1.189     1.232     0.52
0.59
## P772    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P773    4.270  6   0.640      0.610     0.735    -0.50
-0.35
## P774    3.501  6   0.744      0.500     0.432    -1.03
-1.24
## P775    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P776    3.312  6   0.769      0.473     0.536    -0.83
-0.84
## P777    6.912  6   0.329      0.987     1.070     0.17
0.31
## P778    6.687  6   0.351      0.955     0.926     0.10
0.04
## P779    5.317  6   0.504      0.760     1.090     1.14
0.36
## P780   21.946  6   0.001      3.135     1.764     2.10
1.26
## P781    1.709  6   0.944      0.244     0.237    -1.80
-1.98
```

```
## P783    6.118  6   0.410      0.874     0.920     0.16
0.08
## P784    5.317  6   0.504      0.760     1.090     1.14
0.36
## P785    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P786    8.723  6   0.190      1.246     1.228     0.62
0.58
## P787    3.754  6   0.710      0.536     0.670    -0.32
-0.38
## P788    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P789   18.154  6   0.006      2.593     2.719     2.29
2.45
## P790    3.312  6   0.769      0.473     0.536    -0.83
-0.84
## P791   15.488  6   0.017      2.213     2.379     1.94
2.12
## P792   12.406  6   0.053      1.772     1.886     1.00
1.32
## P793    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P794   10.531  6   0.104      1.504     1.592     1.02
1.15
## P795    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P796   10.264  6   0.114      1.466     1.357     0.97
0.80
## P797    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P798    5.533  6   0.478      0.790     0.810    -0.27
-0.22
## P799    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P800    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P801    5.323  6   0.503      0.760     0.763    -0.34
```

```
-0.33
## P802    3.322  6   0.768      0.475     0.511    -1.10
-0.99
## P803   26.502  6   0.000      3.786     3.898     3.41
3.51
## P804    5.220  6   0.516      0.746     0.915    -0.12
0.05
## P805    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P806    3.157  6   0.789      0.451     0.530    -0.68
-0.78
## P808    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P809    5.669  6   0.461      0.810     0.823    -0.21
-0.19
## P810    2.846  6   0.828      0.407     0.403    -1.33
-1.35
## P811    8.723  6   0.190      1.246     1.228     0.62
0.58
## P813    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P814    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P816    5.258  6   0.511      0.751     0.767    -0.35
-0.32
## P817    7.615  6   0.268      1.088     1.183     0.36
0.49
## P818    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P819    5.884  6   0.436      0.841     0.915    -0.16
0.01
## P820   30.216  6   0.000      4.317     4.168     2.05
3.05
## P821    4.719  6   0.580      0.674     0.764    -0.46
-0.31
## P822    3.372  6   0.761      0.482     0.854     1.03
0.02
```

```
## P823    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P824    3.394  6   0.758      0.485     0.856     1.03
0.03
## P825    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P827    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P828    9.334  6   0.156      1.333     1.451     0.75
0.93
## P829    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P830    3.394  6   0.758      0.485     0.856     1.03
0.03
## P832    6.497  6   0.370      0.928     0.872     0.04
-0.08
## P833    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P834    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P835    6.740  6   0.346      0.963     0.980     0.10
0.13
## P836    7.855  6   0.249      1.122     1.180     0.40
0.50
## P837    7.615  6   0.268      1.088     1.183     0.36
0.49
## P838    8.712  6   0.190      1.245     0.908     0.59
0.00
## P839    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P840    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P841   11.150  6   0.084      1.593     1.689     1.13
1.27
## P842    3.322  6   0.768      0.475     0.511    -1.10
-0.99
## P843    1.867  6   0.932      0.267     0.253    -1.90
```

```
-1.96
## P844    3.322  6   0.768      0.475     0.511    -1.10
-0.99
## P845    5.553  6   0.475      0.793     0.801    -0.27
-0.25
## P846    3.086  6   0.798      0.441     0.446    -1.21
-1.20
## P847    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P848   11.917  6   0.064      1.702     1.679     1.25
1.23
## P849    9.838  6   0.132      1.405     1.154     0.83
0.45
## P850    7.057  6   0.316      1.008     1.269     0.38
0.60
## P851    6.291  6   0.391      0.899     0.861    -0.02
-0.10
## P852    3.394  6   0.758      0.485     0.856     1.03
0.03
## P853    6.494  6   0.370      0.928     0.989     0.22
0.20
## P854    4.878  6   0.560      0.697     0.798     0.07
-0.12
## P856   13.028  6   0.043      1.861     1.830     1.50
1.45
## P857   15.377  6   0.018      2.197     2.338     1.93
2.08
## P858    7.130  6   0.309      1.019     1.082     0.21
0.33
## P859    8.127  6   0.229      1.161     1.480     0.51
0.86
## P860    2.707  6   0.845      0.387     0.428    -1.41
-1.25
## P862    7.267  6   0.297      1.038     1.135     0.30
0.42
## P863   26.594  6   0.000      3.799     3.764     2.31
2.63
```

```
## P864    3.632  6   0.726      0.519     0.497    -0.92
-0.98
## P865    5.552  6   0.475      0.793     0.854    -0.26
-0.12
## P866    7.788  6   0.254      1.113     1.250     0.39
0.63
## P867    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P868    3.970  6   0.681      0.567     0.852     0.50
-0.18
## P869    2.046  6   0.915      0.292     0.319    -1.37
-1.54
## P870    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P871    6.770  6   0.343      0.967     1.009     0.11
0.19
## P872    2.991  6   0.810      0.427     0.429    -1.27
-1.25
## P873    2.619  6   0.855      0.374     0.489    -0.05
-0.97
## P874   11.995  6   0.062      1.714     1.691     1.31
1.28
## P875    3.689  6   0.719      0.527     0.710     0.12
-0.40
## P876    3.001  6   0.809      0.429     0.455    -1.19
-1.11
## P877    5.155  6   0.524      0.736     0.680    -0.36
-0.49
## P879    6.589  6   0.360      0.941     1.008     0.06
0.19
## P880    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P881   11.318  6   0.079      1.617     1.571     1.16
1.11
## P883    1.961  6   0.923      0.280     0.291    -1.75
-1.71
## P884    5.552  6   0.475      0.793     0.854    -0.26
```

```
-0.12
## P885    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P886    6.942  6   0.326      0.992     1.014     0.16
0.20
## P887   15.051  6   0.020      2.150     2.424     1.82
2.12
## P888    2.441  6   0.875      0.349     0.343    -1.57
-1.58
## P889    3.501  6   0.744      0.500     0.432    -1.03
-1.24
## P890   11.537  6   0.073      1.648     1.779     1.23
1.41
## P891    6.034  6   0.419      0.862     0.920    -0.11
0.01
## P892    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P893    3.372  6   0.761      0.482     0.854     1.03
0.02
## P894    4.499  6   0.609      0.643     0.475    -0.60
-1.09
## P895    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P896   11.318  6   0.079      1.617     1.571     1.16
1.11
## P897    3.821  6   0.701      0.546     0.490    -0.89
-1.05
## P898    7.125  6   0.309      1.018     0.675     0.22
-0.52
## P899    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P900    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P901    1.024  6   0.985      0.146     0.306    -0.46
-0.78
## P902    2.846  6   0.828      0.407     0.403    -1.33
-1.35
```

```
## P903    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P904   29.847  6   0.000      4.264     1.483     1.72
0.83
## P905    4.878  6   0.560      0.697     0.798     0.07
-0.12
## P906   11.089  6   0.086      1.584     1.887     1.10
1.50
## P907    4.540  6   0.604      0.649     0.653    -0.57
-0.56
## P908   19.883  6   0.003      2.840     1.378     1.91
0.77
## P909    5.336  6   0.501      0.762     0.789    -0.31
-0.26
## P910    8.777  6   0.187      1.254     1.404     0.56
0.78
## P911    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P912    3.312  6   0.769      0.473     0.536    -0.83
-0.84
## P913    5.014  6   0.542      0.716     0.807     0.10
-0.10
## P915   10.558  6   0.103      1.508     1.570     1.03
1.12
## P916    4.492  6   0.610      0.642     0.659    -0.63
-0.59
## P917    6.118  6   0.410      0.874     0.920     0.16
0.08
## P918    5.773  6   0.449      0.825     0.845    -0.19
-0.15
## P919    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P920    9.045  6   0.171      1.292     1.567     0.61
1.02
## P921    2.041  6   0.916      0.292     0.298    -1.80
-1.77
## P922    6.118  6   0.410      0.874     0.920     0.16
```

```
0.08
## P923    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P924    2.801  6   0.833      0.400     0.405    -1.36
-1.35
## P925   10.425  6   0.108      1.489     1.581     1.00
1.13
## P926   25.795  6   0.000      3.685     0.800     2.11
-0.12
## P927    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P928   51.394  6   0.000      7.342     1.425     2.89
0.80
## P929    7.793  6   0.254      1.113     1.181     0.47
0.48
## P930    2.024  6   0.918      0.289     0.271    -1.74
-1.85
## P931    4.071  6   0.667      0.582     0.592    -0.78
-0.75
## P932    8.617  6   0.196      1.231     1.350     0.59
0.78
## P933    1.749  6   0.941      0.250     0.295    -1.23
-1.51
## P934    1.709  6   0.944      0.244     0.237    -1.80
-1.98
## P935    3.874  6   0.694      0.553     0.595    -0.86
-0.74
## P936    9.895  6   0.129      1.414     1.116     0.84
0.39
## P937    8.723  6   0.190      1.246     1.228     0.62
0.58
## P938    2.962  6   0.814      0.423     0.635     0.39
-0.74
## P939    3.412  6   0.756      0.487     0.538    -1.01
-0.87
## P941    6.629  6   0.357      0.947     0.978     0.07
0.13
```

```
## P942    4.153  6   0.656      0.593     0.598    -0.54
-0.68
## P943    4.071  6   0.667      0.582     0.592    -0.78
-0.75
## P944    9.772  6   0.135      1.396     0.693     0.77
-0.44
## P945    7.342  6   0.290      1.049     1.181     0.29
0.49
## P946    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P947    5.701  6   0.458      0.814     0.910    -0.16
0.00
## P948   31.294  6   0.000      4.471     2.169     1.62
2.04
## P949    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P950    5.769  6   0.450      0.824     0.774     0.05
-0.12
## P951    2.046  6   0.915      0.292     0.319    -1.37
-1.54
## P952    9.521  6   0.146      1.360     1.443     0.79
0.92
## P954    3.898  6   0.690      0.557     0.583    -0.86
-0.79
## P955    5.117  6   0.529      0.731     0.766    -0.25
-0.28
## P956    4.634  6   0.592      0.662     0.684    -0.58
-0.52
## P957    6.615  6   0.358      0.945     1.113     0.12
0.38
## P959    5.433  6   0.490      0.776     0.768    -0.31
-0.32
## P960   12.979  6   0.043      1.854     1.454     1.50
0.94
## P961   12.696  6   0.048      1.814     2.033     0.96
1.45
## P962   16.693  6   0.010      2.385     2.422     1.79
```

```
2.05
## P963    3.372  6   0.761      0.482     0.854     1.03
0.02
## P964    0.753  6   0.993      0.108     0.101    -2.76
-2.80
## P965   11.561  6   0.073      1.652     1.712     1.23
1.31
## P966    1.046  6   0.984      0.149     0.157    -0.87
-1.92
## P968    4.565  6   0.601      0.652     0.691    -0.49
-0.46
## P969    3.750  6   0.710      0.536     0.563    -0.93
-0.85
## P970    1.214  6   0.976      0.173     0.175    -2.40
-2.40
## P971   10.746  6   0.097      1.535     1.592     0.78
0.99
## P972    1.507  6   0.959      0.215     0.223    -2.04
-2.00
## P974    4.131  6   0.659      0.590     0.651    -0.76
-0.60
## P975    5.204  6   0.518      0.743     0.749    -0.13
-0.27
## P976    4.134  6   0.659      0.591     0.614    -0.77
-0.70
## P977    2.714  6   0.844      0.388     0.505    -0.03
-0.92
## P978    1.572  6   0.955      0.225     0.219    -2.08
-2.10
## P980    7.846  6   0.250      1.121     1.312     0.39
0.71
## P982    1.961  6   0.923      0.280     0.291    -1.75
-1.71
## P983   20.791  6   0.002      2.970     3.401     1.75
2.58
## P984    6.913  6   0.329      0.988     0.951     0.36
0.14
```

```
## P985    7.929  6   0.243      1.133     1.205     0.42
0.55
## P986    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P988    2.484  6   0.870      0.355     0.335    -1.65
-1.69
## P990    8.845  6   0.182      1.264     1.322     0.64
0.74
## P991   16.138  6   0.013      2.305     2.517     2.00
2.24
## P992    6.584  6   0.361      0.941     0.933     0.06
0.05
## P993    7.457  6   0.281      1.065     1.297     0.43
0.64
## P994    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P995    1.867  6   0.932      0.267     0.253    -1.90
-1.96
## P996    2.809  6   0.832      0.401     0.482    -0.80
-0.91
## P997    4.049  6   0.670      0.578     0.866     0.50
-0.15
## P998    1.314  6   0.971      0.188     0.177    -2.30
-2.36
## P999    2.697  6   0.846      0.385     0.438    -0.60
-0.92
## P1000   1.046  6   0.984      0.149     0.157    -0.87   -
```

```
  devtools::session_info()
```

```
## - Session info
-------------------------------------------------------------
##  setting  value
##  version  R version 3.5.0 Patched (2018-04-23 r74633)
##  os       Windows 10 x64
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
```

```
##  collate  Portuguese_Brazil.1252
##  tz       America/Sao_Paulo
##  date     2018-05-22
##
## - Packages
-------------------------------------------------------------
##  package     * version     date       source

##  assertthat    0.2.0       2017-04-11 CRAN (R 3.5.0)

##  backports     1.1.2       2017-12-13 CRAN (R 3.5.0)

##  bindr         0.1.1.9000  2018-05-12 Github
(krlmlr/bindr@b6e6fd6)
##  bindrcpp    * 0.2.2.9000  2018-05-12 Github
(krlmlr/bindrcpp@bd5ae73)
##  broom         0.4.4       2018-05-12 Github
(tidyverse/broom@570b25a)
##  callr         2.0.3       2018-04-11 CRAN (R 3.5.0)

##  cellranger    1.1.0       2016-07-27 CRAN (R 3.5.0)

##  cli           1.0.0       2017-11-05 CRAN (R 3.5.0)

##  clisymbols    1.2.0       2017-05-21 CRAN (R 3.5.0)

##  colorspace    1.3-2       2016-12-14 CRAN (R 3.5.0)

##  crayon        1.3.4       2017-09-16 CRAN (R 3.5.0)

##  debugme       1.1.0       2017-10-22 CRAN (R 3.5.0)

##  desc          1.2.0       2018-05-01 CRAN (R 3.5.0)
```

```
##  devtools      1.13.5.9000 2018-05-12 Github
(hadley/devtools@13ee56b)
##  digest        0.6.15      2018-01-28 CRAN (R 3.5.0)

##  dplyr       * 0.7.5.9000  2018-05-12 Github
(tidyverse/dplyr@09209ae)
##  eRm         * 0.16-0      2018-03-11 CRAN (R 3.5.0)

##  evaluate      0.10.3      2018-05-12 Github
(hadley/evaluate@06f8e24)
##  forcats     * 0.3.0.9000  2018-05-12 Github
(tidyverse/forcats@f4a7fd1)
##  foreign       0.8-70      2017-11-28 CRAN (R 3.5.0)

##  ggplot2     * 2.2.1.9000  2018-05-12 Github
(tidyverse/ggplot2@4463da6)
##  glue          1.2.0       2017-10-29 CRAN (R 3.5.0)

##  gtable        0.2.0.9000  2018-05-12 Github
(hadley/gtable@0ed36a4)
##  haven         1.1.1.9000  2018-05-12 Github
(tidyverse/haven@746eb3e)
##  hms           0.4.2       2018-05-12 Github
(tidyverse/hms@c0cfc01)
##  htmltools     0.3.6       2017-04-28 CRAN (R 3.5.0)

##  httr          1.3.1       2018-05-12 Github (r-
lib/httr@6b2dadc)
##  jsonlite      1.5         2017-06-01 CRAN (R 3.5.0)

##  knitr         1.20.3      2018-05-12 Github
(yihui/knitr@dc028f4)
##  lattice       0.20-35     2017-03-25 CRAN (R 3.5.0)

##  lazyeval      0.2.1.9000  2018-05-12 Github
```

```
(hadley/lazyeval@93c455c)
##  lubridate     1.7.4       2018-05-12 Github
(tidyverse/lubridate@45395b4)
##  magrittr      1.5.0       2018-05-12 Github
(tidyverse/magrittr@0a76de2)
##  MASS          7.3-50      2018-04-30 CRAN (R 3.5.0)

##  Matrix        1.2-14      2018-04-13 CRAN (R 3.5.0)

##  memoise       1.1.0       2017-04-21 CRAN (R 3.5.0)

##  mnormt        1.5-5       2016-10-15 CRAN (R 3.5.0)

##  modelr        0.1.2       2018-05-11 CRAN (R 3.5.0)

##  munsell       0.4.3       2016-02-13 CRAN (R 3.5.0)

##  nlme          3.1-137     2018-04-07 CRAN (R 3.5.0)

##  pillar        1.2.2       2018-04-26 CRAN (R 3.5.0)

##  pkgbuild      1.0.0       2018-05-12 Github (r-
lib/pkgbuild@0457039)
##  pkgconfig     2.0.1       2017-03-21 CRAN (R 3.5.0)

##  pkgload       1.0.0       2018-05-12 Github (r-
lib/pkgload@35efedd)
##  plyr          1.8.4       2016-06-08 CRAN (R 3.5.0)

##  psych         1.8.4       2018-05-06 CRAN (R 3.5.0)

##  purrr       * 0.2.4.9000  2018-05-12 Github
(tidyverse/purrr@fda4bbe)
##  R6            2.2.2.9000  2018-05-12 Github (r-
```

```
lib/R6@a9eb0f1)
##  Rcpp          0.12.17     2018-05-12 Github
(RcppCore/Rcpp@001db74)
##  readr       * 1.2.0       2018-05-12 Github
(tidyverse/readr@d6d622b)
##  readxl        1.1.0.9000  2018-05-12 Github
(tidyverse/readxl@b10a1a8)
##  reshape2      1.4.3       2018-05-12 Github
(hadley/reshape@777638a)
##  rlang         0.2.0.9001  2018-05-12 Github
(tidyverse/rlang@ccdbd8b)
##  rmarkdown     1.9.11      2018-05-12 Github
(rstudio/rmarkdown@41b2bab)
##  rprojroot     1.3-2       2018-01-03 CRAN (R 3.5.0)

##  rstudioapi    0.7         2017-09-07 CRAN (R 3.5.0)

##  rvest         0.3.2.9000  2018-05-12 Github
(hadley/rvest@9a51a5d)
##  scales        0.5.0.9000  2018-05-12 Github
(hadley/scales@d767915)
##  sessioninfo   1.0.0       2017-06-21 CRAN (R 3.5.0)

##  stringi       1.2.2       2018-05-02 CRAN (R 3.5.0)

##  stringr     * 1.3.1       2018-05-12 Github
(tidyverse/stringr@eff4e4d)
##  testthat      2.0.0       2017-12-13 CRAN (R 3.5.0)

##  tibble      * 1.4.2       2018-01-22 CRAN (R 3.5.0)

##  tidyr       * 0.8.0       2018-01-29 CRAN (R 3.5.0)

##  tidyselect    0.2.4       2018-02-26 CRAN (R 3.5.0)
```

```
##  tidyverse   * 1.2.1.9000  2018-05-12 Github
(tidyverse/tidyverse@83f6ec3)
##  usethis       1.3.0       2018-02-24 CRAN (R 3.5.0)

##  withr         2.1.2       2018-05-12 Github
(jimhester/withr@79d7b0d)
##  xml2          1.2.0.9000  2018-05-12 Github (r-
lib/xml2@ba3511f)
##  yaml          2.1.19      2018-05-01 CRAN (R 3.5.0)
```