REVISTA DA **ABRALIN**

RELATÓRIO DE PESQUISA

Componentes da habilidade oral: uma análise das propriedades dos itens analíticos do exame Celpe-Bras

Laura Marcia Luiza FERREIRA
Universidade Federal da Integração Latino-Americana (UNILA)

RESUMO

As avaliações externas são constantemente desafiadas a informar exatamente o que o teste avalia e a pertinência do uso de seus resultados. Neste trabalho, discuto como os subcomponentes da habilidade oral do Celpe-bras interagem para construção da nota do avaliador-observador. A escala é compostas pelos seguintes itens: *compreensão*, *competência interacional*, *fluência*, *adequação lexical*, *adequação gramatical* e *pronúncia.* Com o objetivo de estudar as propriedades desses itens e relacioná-las aos descritores da grade e à tarefa de avaliação oral, apresentei uma análise das medidas de dificuldade e discriminação de cada um dos itens, apoiando-me no modelo da habilidade comunicativa da linguagem de Bachman (1990). O modelo Rasch básico na extensão *Partial Credit Model* foi utilizado para a análise dos itens. O *corpus* é composto por notas de 1.000 participantes da primeira edição de 2016. Após a análise, argumento que a tarefa interfere na forma como os subcomponentes da habilidade linguística interagem. Além disso, os dados sugerem que a competência linguística seja mais determinante para classificar os examinandos nas faixas mais elevadas do exame do que os subcomponentes da competência estratégica, e ainda, que a compreensão seja um requisito para o examinando demonstrar um desempenho adequado nas faixas de classificação do exame.

ABSTRACT

Large scale tests are frequently under a challenge as they have to inform

OPEN ACCESS

EDITADO POR
- Luiz Amaral (UMASS)
- Ricardo de Souza (UFMG)
- Thaís Maíra de Sá (CEFET-MG)

AVALIADO POR
- Lilian Hübner (PUC-RS)
- Lêda Tomitch (UFSC)

DATAS
- Recebido: 19/08/2020
- Aceito: 03/11/2020
- Publicado: 23/12/2020

COMO CITAR
Ferreira, L. M. L (2020) Componentes da habilidade oral: uma análise das propriedades dos itens analíticos do exame Celpe-Bras. *Revista da Abralin*, v. 19, n. 3, p. 799-824, 2020.

what is being tested and also argue about their social accountability. In this paper, I discuss how the oral ability components are being evaluated by the rater *avaliador-observador* at the Celpe-bras exam. The analytic rubric was combined by the following features or items: *comprehension*, *interactional competence*, *fluency*, *lexical adequacy*, *grammatical adequacy* and *pronunciation*. The aim of this paper is to study the items properties in line with the rubrics descriptions and tasks. The analyses of the item difficulty index and the item discrimination index were made in the light of Bachman's (1990) model of communicative language ability and was calculated using Rasch's Partial Credit Model. One-thousand examinees' scores were analyzed from the first oral exam edition of 2016. Item analysis endorses the idea that tasks influence the way components interact. By analyzing the items indexes, I argue that linguistic competence is more important to place examinees at the higher proficiency bands compared to the strategic competence related items. I also discuss that *comprehension* may be a requirement for the examinee to develop oral performance at the Celpe-Bras oral exam.

PALAVRAS-CHAVE
Proficiência oral. Celpe-Bras. Teoria de Resposta ao Item.

KEYWORDS
Oral proficiency. Celpe-Bras. Item response theory.

Introdução

A virada crítica no campo dos estudos da linguística aplicada tem posicionado o debate sobre usos da linguagem em linha com os direitos humanos fundamentais tais como direito à educação e direitos linguísticos, ao mesmo tempo em que tem ampliado as discussões sobre como a língua pode ser usada como instrumento de políticas linguísticas. Uma alternativa que promova a diversidade e a justiça social por meio do ensino de línguas é a aposta, por exemplo, da pedagogia de repertórios, em detrimento das práticas de ensino e aprendizagem cujo foco é a aquisição do sistema linguístico de uma determinada língua. Exemplifica tal diretriz, a noção de translinguagem que concebe todo repertório linguístico do indivíduo como um sistema linguístico independente da quantidade de línguas e seus níveis de desempenho (CANAGARAJAH, 2018). Jaspers (2017) explica que o termo *translanguaging* tem a origem no termo *languaging* cunhado pelos teóricos da abordagem comunicativa para o ensino de línguas. *Languaging* se refere ao processo dinâmico de uso da linguagem para

produção de sentidos, explica o autor. *Languaging*, no entanto, tem como pressuposto que a produção de sentidos se dá com a finalidade de aquisição, ensino ou avaliação de um língua específica, sendo estrangeira, adicional ou de herança a depender do contexto.

Interessa neste trabalho, investigar como a língua portuguesa que está sendo avaliada no contexto de um teste de larga escala, o exame Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). Ao mesmo tempo em que há um crescente debate sobre a valorização das subjetividades no processo de ensino de línguas estrangeiras que celebra o hibridismo linguístico, cada vez mais se discute o complexo papel das avaliações de proficiência para o ensino e para a formação de professores, situando-as em contextos mais amplos de políticas educacionais e internacionais. Além dos debates acerca das pedagogias de repertório, que desconstroem pressupostos de língua em que são normalmente baseados os instrumentos de avaliação, McNamara (2004) afirma que os testes de larga escala são também responsabilizados pelos usos sociais que são feitos do seu resultado e ainda sobre o impacto na construção de políticas linguísticas locais, regionais e internacionais. Nesta perspectiva, os instrumentos devem ser constantemente avaliados e revisados quanto aos modelos teóricos que mobilizam e ao significado dos resultados gerados. Avaliar a avaliação, ou seja, fazer a validação de testes é um processo em construção constante que carece de evidências teóricas e empíricas (MESSICK, 1987; FULCHER, 2003, BACHMAN, 1990; MCNAMARA, 2004). O processo em construção constante da validação não tem prazo para terminar, pois os testes podem ser a qualquer momento colocados em prova sob a luz de novos dados ou debates teóricos, mesmo depois de terem sido validados antes de sua implementação. McNamara (2004) explica que o processo de análise dos dados gerados nos exames de proficiência linguística podem gerar informações sobre o quanto o uso dos modelos teóricos está adequado para subsidiar a construção de instrumentos, retroalimentando assim discussões mais amplas sobre ensino e aprendizagem de línguas. Por meio da análise das propriedades dos itens avaliados na prova oral, discuto o significado da nota analítica, atribuída pelo avaliador-observador do Celpe-Bras, de forma a relacionar com os conceitos teóricos operacionalizados na grade com o modelo da habilidade comunicativa da linguagem de Bachman (1990).

1. O Celpe-Bras: construtos e critério

O Celpe-Bras é o exame oficial do Governo do Brasil para a comprovação da proficiência em Língua Portuguesa de cidadãos de outros países que não têm o português como língua oficial. Juntamente com o exame argentino Certificado de *Español Lengua y Uso* (CELU), o Celpe-Bras foi criado no contexto do Mercosul com o objetivo de fomentar e promover o uso do português e do espanhol em contextos acadêmicos e econômicos. Tais exames formam parte de relações políticas mais amplas cujo objetivo é integrar os países que formam o bloco. O objeto de avaliação dos exames é a língua portuguesa ou espanhola faladas na América Latina em contextos profissionais e acadêmicos, portanto, pode ser que a noção de língua híbrida, como o portunhol, não seja bem vinda neste contexto. É preciso investigar, no entanto, quais e como os aspectos dessas línguas estão sendo avaliados.

McNamara (2004) define os testes de língua como um processo de coleta de informações sobre o que os examinandos sabem fazer na língua, ou seja, de sua performance, a partir de condições estabelecidas no instrumento. Dois conceitos fundamentais para entender o significado dessas informações coletadas nos testes de língua são o de critério e o de construto. O critério é um termo da área de avaliação que diz respeito ao conjunto de usos reais da língua de interesse do teste. No caso do Celpe-Bras e do CELU, as tarefas comunicativas operacionalizam o critério ao delimitarem o uso da linguagem que se espera dos examinandos. Nos exames em que se usa os critérios como referência para dar notas (*criterion-referenced*), a classificação da proficiência ou mensuração dos escores são feitos a partir da descrição verbal dos desempenhos esperados que formam os descritores das grades e escalas de itens. A produção verbal do examinando é avaliada a partir de um número possível de desempenhos previstos. Schlatter *et al.* (2008) explicam que as tarefas comunicativas avaliam de forma integrada habilidades e competências envolvidas no uso da linguagem. As tarefas, bem como os itens por meio delas avaliados, são elaboradas a partir dos construtos teóricos. Os construtos teóricos se referem aos modelos teóricos sobre os desempenhos que foram operacionalizados no exame. No caso de exames de proficiência linguística, os testes operacionalizam visões de língua nas tarefas e nas grades de correção, ao especificar como a linguagem deve ser usada para atingir as faixas de certificação previstas. Sobre o construto da linguagem operacionalizado no Celpe-Bras, as elaboradoras afirmam que

> não se busca aferir conhecimentos sobre a língua, como é o caso de exames tradicionais que formulam questões sobre morfologia e sintaxe, porém, sim, a capacidade de uso dessa língua, já que a competência lingüística é um dos componentes da comunicativa. Assim, o exame está centrado no desenvolvimento de uma competência de uso que requer muito mais do que a manipulação de formas e regras lingüísticas, exigindo também o conhecimento de regras de comunicação e de formas que sejam não apenas gramaticalmente corretas, mas socialmente adequadas. (DELL'ISOLA *et al.* 2003, p. 155)

A partir da afirmação acima, há uma clara intenção de assumir uma perspectiva teórica descolada da estruturalista, uma vez que não apenas as "formas e regras linguísticas", mas também as "regras de comunicação" são levadas em conta quando se avalia o uso da língua. Além disso, destaco a diretriz de interpretação do desempenho pautado no que é 'socialmente adequado' e não apenas na adequação à norma gramatical. Sobre a língua portuguesa que se pretende certificar por meio do Celpe-Bras, o exame se alinha à perspectiva comunicativa, uma vez que a competência de uso é o elemento principal para elaboração do instrumento. No caso o *languaging*, ou competência de uso da língua portuguesa, parece ser o principal construto operacionalizado no exame. Em linhas gerais, o foco do ensino comunicativo é o fomento da produção de significados em língua estrangeira e, para fazê-lo, é preciso negociar sentidos durante o uso situado da linguagem por meio da interação com os outros, seja lendo e se posicionando sobre o que leu ou conversando, por exemplo. Uma implicação para elaboração de testes é considerar no desenho dos instrumentos a simulação de usos da linguagem em que a interação é o elemento principal por meio do qual os sentidos são construídos. McNamara (2004) afirma que a perspectiva interacionista da linguagem desafia as noções individualistas de desempenho avaliadas normalmente em testes de larga escala, uma vez que se

considera um amplo leque de possibilidades de produção e negociação de sentidos por meio do uso da língua. Ao operacionalizar um construto alinhado à perspectiva comunicativa e interacionista da linguagem, o desenho das tarefas comunicativas do Celpe-Bras delimitam um objetivo de interação e papéis para os interagentes, que utilizarão a língua para simular situações de uso da linguagem. A partir do desempenho nas situações simuladas, a competência de uso da língua portuguesa que o examinando faz é avaliada. No desenho da prova oral do Celpe-Bras, há uma coerência com a perspectiva teórica adotada uma vez que a produção e compreensão oral são avaliadas a partir de uma interação face a face, em que avaliador e examinando interagem em uma situação definida.

O maior desafio de avaliar a proficiência oral, tendo em conta os níveis de adequação da produção e negociação de sentidos pelos examinandos em situação de prova, é garantir que não só a tarefa, mas também o significado da nota esteja coerente com as escolhas teóricas que subsidiaram a elaboração do exame. No caso do Celpe-Bras, espera-se que no processo de composição da nota oral a capacidade de uso da língua, conforme explicam as elaboradoras, seja a diretriz principal para formação do escore. Diversos teóricos da abordagem comunicativa propuseram subdivisões distintas das competências que fazem parte da capacidade de uso da língua. Canale (1983) divide a competência comunicativa em competência gramatical, competência sociolinguística, competência discursiva e competência estratégica. O autor explica que sua proposta não se trata de um modelo, justamente porque lhe falta especificar como os subcomponentes interagem entre si. Fulcher e Davidson (2007) resenham vários modelos que se desenvolveram a partir da proposta de Canale (1983), dentre os quais cito aqui o de Bachman (1990). Bachman (1990) estava interessado em saber como construir testes de línguas tendo como construto os quadros teóricos da competência comunicativa. O autor explorou a interação entre as subcompetências e definiu que havia a competência linguística, relacionada ao uso dos componentes linguísticos na interação; a competência estratégica, que diz respeito à capacidade mental para escolha dos componentes ao interagir e a competência psicofisiológica, que seria a performance em si ou a execução em si da língua. Ao propor uma maneira de interação entre os componentes, o autor explica que a tarefa dos testes, ao definirem um contexto de uso da língua, interage com todos os três componentes, podendo alterar a forma como estão relacionados. Além disso, embora pouco se defina sobre como os componentes interagem, o modelo de Bachman (1990) sugere que a tarefa do teste direciona a competência estratégica para escolhas adequadas de recursos linguísticos e outros conhecimentos de mundo ao executar a língua.

A reflexão e análise empírica sobre a forma como tais componentes da habilidade interagem estão diretamente relacionadas com a forma como os escores serão compostos. Sobre a mensuração da nota oral analítica do Celpe-Bras, optou-se pela elaboração de grades e escalas baseadas no critério da proficiência oral demonstrada durante uma interação face a face. A nota é composta a partir de subcomponentes da habilidade oral, a saber: *compreensão*, *competência interacional*, *fluência*, *adequação gramatical*, *adequação lexical* e *pronúncia*. Entendo aqui que os três primeiros componentes estariam mais próximos da competência estratégica e os três últimos, da competência linguística, definida no modelo de Bachman (1990). Tais subcomponentes da proficiência oral são descritos em escalas de cinco pontos. Sobre a importância das grades e escalas, Fulcher e Davidson (2007) afirmam que a grade ajuda a

determinar o significado da nota. Neste sentido, tirar uma nota máxima na prova oral significaria o quê, exatamente? Que o examinando compreendeu bem o avaliador? Que usou palavras específicas sem interferência da língua materna sobre um assunto específico? Que fala por muito tempo sobre o assunto abordado? Ou seja, por meio de quais evidências do desempenho oral descritas na grade podemos inferir que o examinando fala muito bem ou quase nada? Se a proficiência oral é decomposta em várias subcompetências, como é a interação entre elas?

Por meio do estudo da composição das notas é possível avaliar a interação entre os diferentes construtos teóricos operacionalizados na grade que compõem a nota. Para avaliar o quanto os resultados refletem as diretrizes do exame, uma das maneiras é investigar os itens avaliados. Descrevo abaixo como funciona a composição das notas.

1.1 O Celpe-Bras: sistema de certificação

O Celpe-Bras certifica as seguintes faixas de proficiência *intermediário*, *intermediário superior*, *avançado* e *avançado superior* e cada faixa corresponde a um intervalo de nota de 1,99 a 5 pontos. Os examinandos que obtêm uma nota igual ou inferior a 1.99 pontos não são certificados (BRASIL, 2016). Os examinandos são avaliados quanto à proficiência oral e escrita. A nota final para fins de certificação é a menor nota atribuída ao desempenho oral ou escrito do examinando. A prova escrita é composta por quatro tarefas comunicativas no formato de questões dissertativas[1], uma nota de 0 a 5 é atribuída a cada uma das quatro tarefas. Para o cálculo da nota final da prova escrita, somam-se as notas e divide-se por quatro.

A prova oral é uma interação face a face 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). A situação de avaliação é uma entrevista de proficiência oral com duas tarefas, espera-se que o examinando se apresente durante a fase do quebra-gelo e que, em seguida, fale sobre três temas diferentes veiculados em um material impresso chamado Elemento Provocador[2]. No Manual do examinando (2015) afirma-se que são avaliadas a capacidade de produção e de 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 oral são feitas de forma independente. A nota final da prova oral é calculada a partir da média entre as notas dos dois avaliadores. Se as notas forem divergentes por mais de um ponto e meio (1.5), a interação é reavaliada (BRASIL, 2016).

---

1 Para mais informações sobre as tarefas comunicativas da prova escrita, consultar Scaramucci (2001) e Schoffen (2009).

2 O Elemento Provocador é formado, na maioria das vezes, por recortes de reportagens veiculadas na mídia brasileira (FERREIRA, 2012)

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, domínio de estruturas linguísticas ao falar sobre diferentes temas e adequada pronúncia. Tais evidências de desempenho são avaliadas por meio dos itens que compõem as duas grades de avaliação. Na escala de avaliação do avaliador-interlocutor, há uma descrição holística do desempenho oral com o objetivo de avaliá-lo globalmente. A grade holística é graduada em uma escala de 0 a 5 pontos.

Fulcher (2003) explica que itens holísticos, que avaliam de forma global a qualidade do desempenho, são interessantes em contextos de avaliação de larga escala em que o avaliador tem pouco tempo para avaliar muitos examinandos. Para fins de avaliação de aprendizagem, uma grade com itens analíticos, como a grade proposta, tende a ser mais adequada porque oferece mais informações sobre as evidências que o estudante deve demonstrar em suas interações, sejam orais ou escritas. Embora o contexto seja de avaliação educacional, optou-se pela grade analítica. Dessa forma, a partir da análise dos itens analíticos da prova do Celpe-Bras é possível fazer uma discussão mais ampla sobre a interação entre os subcomponentes da habilidade oral avaliada neste contexto.

A grade de avaliação do avaliador-observador é analítica, ou seja, uma nota é atribuída para cada um dos subcomponentes avaliados, doravante, chamarei de item os subcomponentes da habilidade oral avaliada no exame Celpe-Bras. Fazem parte da grade analítica os seguintes itens: *compreensão*, *competência interacional*, *fluência*, *adequação gramatical*, *adequação lexical* e *pronúncia*. Por exemplo, uma nota 0 em *adequação lexical* é caracterizada a partir de 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.

Além dos seis itens analíticos, faz parte da nota oral final também a nota do observador-interlocutor. Neste trabalho, a análise se limita à grade analítica, com vistas a responder as seguintes perguntas:

- Qual é o nível de dificuldade de cada um dos itens que compõem a grade do avaliador-observador?

- Como cada um dos itens analíticos contribuem para a discriminação de cada uma das faixas de certificação do Celpe-Bras?

- Qual a relação entre as propriedades dos itens com a interação entre os subcomponentes da habilidade oral avaliados no exame Celpe-Bras?

McNamara (2000) advoga a favor do uso de técnicas estatísticas utilizadas pelas teorias da testagem para fins de revisão e análise dos testes de língua. O autor explica que a Psicometria analisa a qualidade do processo de avaliação ao investigar as notas. As teorias de testagem podem revelar propriedades dos itens, tais como a sua dificuldade ou o quanto cada item gera de informação relevante sobre o desempenho que se pretende medir. Apresento uma análise de discriminação de itens

calculada por meio de um dos modelos da Teoria de Resposta ao Item, o Rasch, para discussão do significado nas notas graduadas na escala analítica do Celpe-Bras e sua relação com os modelos e teorias de ensino e aquisição de línguas.

2. Avaliações de língua e a teoria de resposta ao item ou modelo Rasch

Szabó (2007) afirma que, embora os estudos sobre o significado da nota em testes educacionais no final dos anos sessenta já utilizassem a Teoria de Resposta ao Item (TRI) nas análises, os estudos sobre exames de línguas começaram a incorporar essa metodologia nos anos oitenta. Segundo os estudos resenhados por Szabó (2007), os objetivos 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.

Os modelos Rasch pertencem a uma família de modelos estatísticos baseados na Teoria de Resposta ao Item (TRI). A TRI tem o objetivo de analisar as qualidades psicométricas dos testes, ou seja, o quanto o sentido das notas geradas por um sistema são válidas para um determinado propósito de avaliação. A TRI viabiliza ferramentas estatísticas para calcular a dificuldade e a discriminação dos itens de um teste.

Os itens podem ser dicotômicos ou politômicos. 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 em mais de uma possibilidade. 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. Segundo Szabó (2007), como só as respostas dos examinandos constituem a informação disponível para o cálculo, a partir das notas cruas é estimada a dificuldade do item em função da habilidade do examinando. A habilidade do examinando no contexto da análise da TRI não é produto de 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. 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. A análise da habilidade e da dificuldade de item gera a função da informação que está relacionada à construção de intervalos confiáveis no cálculo das faixas da proficiência.

Segundo DeMars (2010), é desejável que os itens diferenciem variados níveis de proficiência. 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 deve estar relacionadas com o propósito do teste. No caso do Celpe-Bras, que se propõe a diferenciar e certificar várias faixas de proficiência, espera-se que estejam contemplados nos itens a diferenciação tanto das faixas *avançado* da *intermediário superior* quanto da *sem certificação* da *intermediário*, por exemplo. Se o teste falha na diferenciação das faixas, pode ser que muitos examinandos com habilidades distintas possam estar sendo classificados em uma mesma faixa. Ou seja, se interessa diferenciar quatro faixas de proficiência, é preciso ter itens que sejam eficientes para classificar as habilidades dos examinandos corretamente nestas faixas.

Além dos modelos da TRI permitirem estudar não só a qualidade psicométrica do teste como um todo, é possível também analisar 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 da TRI 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*) de cada item avaliado pelo observador (*compreensão*, *competência interacional*, *fluência*, *adequação gramatical*, *adequação lexical* e *pronúncia*).

Na perspectiva das teorias psicométricas, 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 para cada item analítico 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? Para uma pessoa que tirou nota máxima, a nota máxima em *adequação lexical* foi fácil ou mais difícil de alcançar quando comparamos com o item *compreensão*?

Uma medida que ajuda a responder estas perguntas ao inferir a qualidade da discriminação dos itens em uma escala são os valores de *threshold*. A medida diz respeito ao intervalo entre as faixas de classificação dispostas no gráfico da curva de informação do item. Eckes (2015) defende que o intervalo entre os valores de *threshold* deveria ser homogêneo. Verificar a homogeneidade da escala é uma tarefa empírica possível a partir da análise da relação da medida *dificuldade do item* em função dos perfis de habilidade. Ao considerar as potencialidades de análise do conjunto de respostas da prova oral do Celpe-Bras, justifica-se a pertinência do uso da TRI para o debate sobre a relação entre os sucomponentes da habilidade oral e sua interação.

A seguir apresento a análise e discussão dos resultados.

3. Análise

Para analisar o quão discriminantes são os itens de avaliação oral, utilizo, na análise, a extensão *Partial Credit Model* do Rasch (MAIR *et al.*, 2018) versão 0.16-0. Os itens se referem a cada um dos itens analíticos. Os dados utilizados foram as notas de 1.000 examinandos que dizem respeito à prova oral, das quais faziam parte as seis notas atribuídas pelo avaliador-observador relativas à primeira edição de 2016. As notas foram selecionadas pelos servidores do Inep e disponibilizadas para esta pesquisa, após o pedido de acesso protocolado junto ao Serviço de Atendimento ao Pesquisador do INEP.

Na Tabela 1, a seguir, ao avaliar a quantidade de notas do *corpus* coletado por cada um dos itens da grade analítica, percebe-se uma carência de dados relacionados com as notas zero, um e dois. Ou seja, houve poucos examinandos com notas baixas no corpus coletado. Cabe ressaltar que o conjunto de dados corresponde à 16,07% do total de examinandos inscritos no processo de certificação do Celpe-Bras no primeiro semestre de 2016.

Item analítico	Nota 0	Nota 1	Nota 2	Nota 3	Nota 4	Nota 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

TABELA 1 - Proporção de notas atribuídas por cada item analítico
Fonte: Elaboração própria.

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 chamarei de *categoria* 0. Reagrupo as notas nas categorias que apresento abaixo (TABELA 2). Trata-se da convenção que seguirei na exposição das análises.

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

TABELA 2 - Convenção
Fonte: Elaboração própria.

3.1. Ajuste de modelo e *item fit* e *outfit statistics*

Para investigar o ajuste global, apresento o resultado do *Martin Lof Test*. Verguts e Boeck (2000) sugerem o teste *Martin Lof Test* para avaliar a unidimensionalidade de uma escala. O teste *Martin Lof* consiste em dividir o corpus ao meio e testar se há diferença entre os dois corpus. O critério escolhido para fazer o presente teste foi a mediana, por ser o critério padrão. O teste gera variados valores, dentre eles o *p-value*. O teste gerou os seguintes resultados: *LR-value*, 501.595; *Chi-square df*, 191; e *p-value*, 0.

O valor de *p-value* (valor-p) é uma medida de ajuste global ao modelo, espera-se que o 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 dos valores, conclui-se que há problemas no ajuste da escala, ou seja, as escalas da prova oral do exame Celpe-Bras não se ajustam ao modelo Rasch, porque os escores são ou muito previsíveis ou imprevisíveis, segundo os parâmetros do modelo Rasch. Cabe ressaltar aqui que esta avaliação foi feita das duas escalas, incluindo a nota holística do avaliador-interlocutor.

Ao avaliar os valores de ajuste de INFIT MSQ (*Infit mean-square*) e OUTFIT MSQ (*Outfit mean-square*) de cada um dos itens é possível investigar como cada um deles se ajustou 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 abaixo (TABELA 3), os valores de *p-value* de cada um dos itens sugerem que os itens *fluência*, *adequação gramatical* e *adequação lexical* se ajustam ao modelo, os demais itens apresentam valores insatisfatórios, de zero a próximos de zero, ou seja, não correspondem às expectativas especificadas pela extensão *Partial Credit Model* do Rasch. 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 deveriam exceder 2.0. Na tabela, apenas o item *compreensão* tem valor maior que 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 encontra-se *compreensão*.

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

Legenda:
chisq: *chi-square* ou qui-quadrado
df: *degree of freedom* ou grau de liberdade
p-value: valor-p
outfit msq: *unweighted mean-square* ou quadrado médio não ponderado
infit msq: *weighted mean-square* ou quadrado médio ponderado
outfit T: *unweighted t statistics* ou estatística de ajuste t não ponderada
infit T: *weighted t statistics* ou estatística de ajuste t ponderada

TABELA 3 – Parâmetros de ajuste de item
Fonte: Elaboração própria.

Com o objetivo de oferecer informações que complementam a análise de ajuste por item e levantar hipóteses sobre os resultados acima descritos, apresento abaixo as curvas de características dos itens da escala da prova oral do exame Celpe-Bras, buscando relacioná-las com a forma como os itens são descritos na grade analítica.

3.2 Breve explicação sobre as medidas analisadas nas curvas de características dos itens

Retomando as potencialidades de análise dos itens politômicos pelo Rasch, Eckes (2015) afirma que a qualidade de uma escala pode ser avaliada a partir dos valores de *thresholds*. Os valores de *threshold* dizem respeito ao intervalo entre as faixas de classificação dispostas no gráfico da curva de informação do item que geram a altura das linhas e a distância entre elas. Eckes (2015) defende que o intervalo entre os valores de *threshold* deveria ser homogêneo. Na perspectiva de Eckes (2015), o aspecto visual da faixa avançado, por exemplo, para todos os itens de uma mesma escala deveriam ser similares ou correspondentes. A eficiência com a qual a escala de avaliação diferencia os diversos perfis de desempenho esperados pode ser investigada também a partir da análise da relação da medida *dificuldade do item*. Eckes (2015) explica que uma boa escala distingue ou discrimina com eficiência todos os níveis de proficiência. 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 em *compreensão* e nos demais itens.

No entanto, pensando na natureza da habilidade oral e sua relação com a situação de avaliação oral do Celpe-Bras, pode ser que um item seja útil para discriminar faixas mais altas e outro item, as faixas mais baixas de proficiência. A expectativa deve ser de que a escala como um todo consiga

discriminar tanto examinandos de baixa quanto de alta proficiência. Mais importante do que a homogeneidade entre os valores de *thresholds* e de *dificuldade do item* de cada item específico é a homogeneidade da escala como um todo. Ao compor uma escala única, a partir de todas as informações que cada item gera individualmente, a curva de informação da prova oral deveria nos informar que como um todo o conjunto de itens são eficientes ao discriminar diferentes perfis de habilidade que se pretende medir por meio do exame.

A diferença entre as notas e sua implicação na eficiência de discriminação dos examinandos ao longo das faixas das escalas 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. Eckes (2015) sugere 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 escrito nos gráficos apresentados a seguir. O autor explica também que quando os *thresholds* estão muito próximos na escala da proficiência, as faixas são menos discriminantes do que deveriam ser. O intervalo da categoria de resposta informa o perfil de examinandos que a categoria abarca.

3.3 Curva de característica do item *compreensão*

Na Figura 1, apresento a curva de característica do item *compreensão*. Este subcomponente é descrito na escala quanto à compreensão ou não da fala do avaliador-interlocutor e quanto à necessidade ou não de repetição de alguma ideia[3]. O eixo horizontal refere-se à escala de proficiência do examinando. O eixo vertical se refere aos valores de probabilidade dos examinandos de serem classificados nas categorias analisadas, sendo a *categoria 0* correspondendo ao nível *sem certificação* ou *básico*, a *categoria 1*, *intermediário* e, assim por diante, conforme a tabela 2. As curvas no gráfico representam as categorias ou notas atribuídas que correspondem às faixas de certificação do exame. As curvas estão organizadas de maneira a informar a probabilidade para diferentes perfis de habilidade dos examinandos, representados no eixo horizontal, de estar em uma ou outra categoria, ou seja, de ter tirado uma determinada nota no item avaliado, no caso a *compreensão*.

3 Para acessar as grades de avaliação, acesse o Documento-base do exame Celpe-Bras em : http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/6939071

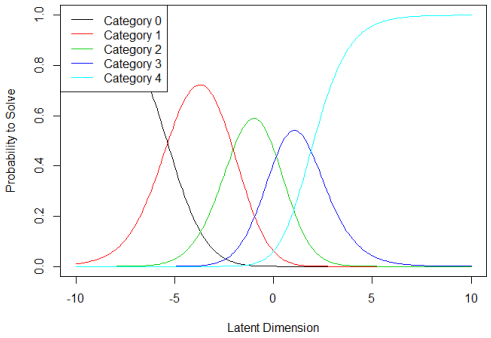

FIGURA 1 - Curva de característica do item *compreensão*
Fonte: Elaboração própria.

Quando analisamos a Figura 1, 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 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*, por exemplo. A distância da interseção entre as categorias 0 e 1 é de 3 pontos na escala de proficiência. A distância da interseção entre as categorias 1 e 2 e o encontro das categorias 2 e 4 é de 2 pontos, sugerindo que o item discrimina melhor os examinandos com proficiência baixa, que corresponde aos valores entre -6 e valores próximos de 1 do eixo horizontal da escala da habilidade. A partir do valor 2, na escala dos valores de proficiência, os examinandos tendem a tirar nota máxima em *compreensão*.

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, *sem certificação* ou *básico* e *intermediário*. Ao buscar uma explicação na forma como a habilidade de compreensão está descrita na escala, percebe-se que as notas 0 e 1 se referem a examinandos que têm muitos ou sérios problemas de compreensão da fala do avaliador-interlocutor. As faixas superiores se referem aos que compreendem o interlocutor. Na grade, diferenciam-se os níveis mais altos pela frequência das necessidades de repetição da fala do interlocutor, sendo que na faixa intermediário está previsto que o examinando possa eventualmente pedir para que o avaliador repita algo e, nas faixas avançadas, que ele peça raramente ou só algumas vezes. A partir da análise, concluo que avaliar a compreensão descrevendo-a quanto à frequência de necessidade de repetição não é eficiente para diferenciar a compreensão de examinandos que apresentam altos níveis de proficiência no contexto da situação de avaliação oral do Celpe-Bras. Essa falta de homogeneidade ao diferenciar diferentes perfis de habilidade pode explicar o baixo desempenho deste item nos testes de ajuste apresentados no começo da análise. A informações da curva reforçam a importância da revisão deste item, especialmente no que diz respeito à

reformulação da tarefa, ou ainda à retirada do item *compreensão* da grade analítica, da forma como está. Discuto mais sobre isso adiante.

3.4 Curva de característica do item *competência interacional*

Na Figura 2, apresento a curva de característica do item *competência interacional* que de forma geral é descrita na grade quanto à autonomia para desenvolver a fala ou uso de respostas breves e a frequência de uso de estratégias para compensar a falta de algum recurso linguístico. Assim como na Figura 1, o eixo horizontal refere-se à escala de habilidade do examinando, estimada empiricamente a partir do conjunto de notas e o 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*. A faixa que representa *avançado superior* do item *competência interacional* alcança valores maiores no eixo horizontal, quando comparado à curva de característica do item *compreensão*. Posso inferir que a nota 5, *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 das de 2 e 3 é de aproximadamente 2,5 pontos, ou seja, a distância entre os pontos de interseção, *thresholds*, das categorias medianas (1, 2 e 3) são mais homogêneas, quando comparadas ao padrão de 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 i*ntermediário*, *intermediário superior* e *avançado* do que o item *compreensão*. A partir do valor 7, na escala dos valores de proficiência, os examinandos têm 100% de chance de tirar nota máxima em *competência interacional*, ou seja, para examinandos que apresentam uma proficiência acima de 7, na escala de proficiência representada no eixo horizontal, o item *competência interacional* não gera informações novas sobre a proficiência do examinando. Uma hipótese para o bom desempenho do item para discriminar diversos perfis é a forma como a competência interacional está descrita: desenvoltura e autonomia são descrições reservadas às faixas avançadas; e respostas breves, às faixas intermediárias. Pressupõe-se pela descrição da grade, que o uso de estratégias, quando necessário, está relacionado às faixas superiores. Seria interessante que outros estudos pudessem explicar qual dessas informações sobre a competência interacional mais influencia o comportamento do avaliador na hora de atribuir

a nota. Os dados sugerem que os turnos de fala descritos como resposta breve em oposição à autonomia e desenvoltura possam ter influenciado mais nesta atribuição do que frequência de uso de estratégias para resolver problemas linguísticos, que é um outro aspecto avaliado também.

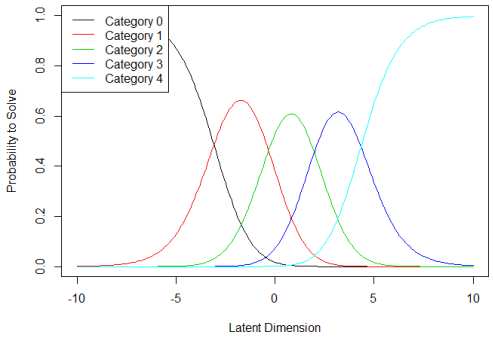

FIGURA 2- Curva de característica do item *competência interacional*
Fonte: Elaboração própria.

3.5 Curva de característica do item *fluência*

Na Figura 3, disponho a curva de característica do item *fluência.* Na grade, a *fluência* está relacionada à frequência com que pausas ou hesitações estão associadas aos problemas de construção linguística. A interrupção eventual está prevista para os níveis avançados. Para os níveis intermediários, os examinandos podem interromper sua própria fala algumas vezes ou frequentemente. Nos níveis não certificados são previstos fluxos de fala em outras línguas e pausas que exigem muito esforço do avaliador para que eles possam compreender a conversa. Pelo gráfico, posso dizer que o item *fluência* assemelha-se às características do item *competência interacional*. Assim como *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.* Ao analisarmos a forma como o item é descrito na grade, a análise sugere que descrever a fluência como sem interrupção ou poucas interrupções pode não ser suficiente para separar o avançado do avançado superior. Porém, um aspecto muito positivo deste item é o quão bem ele separa as faixas intermediárias das avançadas, pode ser que descrever a frequência de pausa e de interrupções no fluxo da conversa com “algumas pausas” “poucas

interrupções" para o *avançado* e "frequentes" pausas e "algumas interrupções" em *intermediário superior* explique a eficiência para separar as faixas. Pode ser que a frequência de pausas seja uma informação saliente na hora de atribuir a nota. A frequência de pausas é uma questão que mereceria ser investigada em futuros trabalhos.

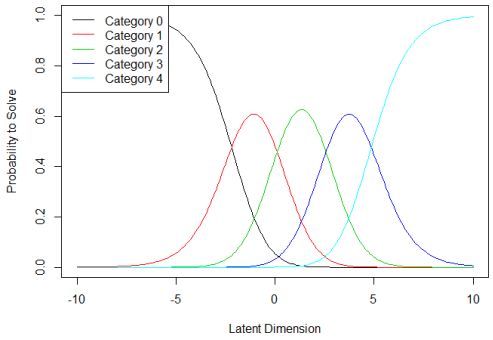

FIGURA 3 - Curva de característica do item *fluência*
Fonte: Elaboração própria.

3.6 Curva de característica do item *adequação lexical* e *adequação gramatical*

A partir daqui analiso itens que estão mais relacionados à competência linguística (BACHMAN, 1990). Como na Figura 4 e 5 da curva de característica dos itens *adequação lexical* e *adequação gramatical* são semelhantes, tratarei deles juntos, buscando explorar como eles se relacionam. Ao analisarmos a distância entre os pontos de interseção, verifica-se que ambos os itens discriminam examinandos de perfis de proficiência elevada, trata-se de uma grande novidade em termos de propriedade destes itens.

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, ou seja, *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*, a medida se refere ao intervalo estabelecido entre dois valores na métrica da habilidade estimada empiricamente pelo modelo, dispostas no eixo horizontal, e representam o encontro entre duas categorias ou faixas ou notas. Tais pontos, que determinam a área do gráfico, referem-se 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. O valor para *Category threshold values* das faixas *avançado* e *avançado superior* é de 3.3 pontos na escala. O mesmo parâmetro para o item *adequação gramatical* é de 3.5 pontos na escala para as faixas *avançado* e *avançado superior*. Tomando como referência que o intervalo não deve ser menor que 1.4 nem maior que 5, sugerido por Eckes (2015), o intervalo dos itens é adequado e, mais que isso, contribui

para diferenciar examinandos com perfil de habilidade alta, que até então os outros itens analisados, *compreensão*, *competência interacional e fluência*, se mostraram ineficientes ao fazê-lo. A *categoria* 3, que representa a nota 4, associada à faixa *avançado* da escala do exame, em ambos os itens é mais larga quando comparada aos outros itens do exame, por isso, 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*, de *compreensão* é 1.89594, *competência interacional*, 4.33582, *fluência*, 4.85844, *adequação lexical*, 7.25356, *adequação gramatical*, 7.25199. Quanto menor esse valor, maior probabilidade do item classificar examinandos de baixa habilidade na faixa *avançado*. Examinandos classificados no perfil próximos ao valor 7.25, provavelmente com proficiência mais avançada, apresentam alta probabilidade de serem classificados com nota 5 apenas nos itens *adequação lexical* e *adequação gramatical*, ao passo que examinandos localizados próximos ao valor 1.89594, pouca proficiência linguística, já tirariam nota 5 no item *compreensão*. Essa é uma informação importante revelada na análise, uma vez que apenas esses dois itens, justamente os que se referem à competência linguística, são os responsáveis por garantir que a escala da nota analítica seja homogênea para faixas mais elevadas. Discuto melhor este dado ao final da análise.

A diferença entre *adequação lexical* e *adequação gramatical* está nos valores das categorias 1 e 2, que estão relacionados 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 da categoria de respostas das categorias 1 e 3 é maior que a categoria 2, ou seja, 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 que *adequação gramatical* os examinandos cuja proficiência é alta, uma vez que as categorias que representam essas faixas são menores no gráfico 04 em relação ao 05, pois cabem um espectro menor de perfis de examinandos. O fato de o intervalo entre a categoria de respostas das categorias 1, *intermediário*, e 3, *avançado*, ser maior 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, é pouco 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." (BRASIL, 2019). 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 na grade. No entanto, o que diferencia o *avançado superior* do *avançado* é "raras interferências em outras línguas" de "poucas interferências em outras línguas", respectivamente. Parece que a avaliação de interferência de outras línguas, como o

portunhol, no contexto de uso do léxico seja algo saliente para os avaliadores-observadores ao escolher entre as faixas avançadas. Retomando a explicação das elaboradoras, temos que "o conhecimento de regras de comunicação e de formas que sejam não apenas gramaticalmente corretas, mas socialmente adequadas" (DELL'ISOLA *et all.* 2003, p.155). Até que ponto seriam as interferências lexicais, como o portunhol, formas linguísticas socialmente inadequadas para o contexto do exame? Outras pesquisas se fazem necessárias para discutir a questão do hibridismo linguístico no contexto de avaliações de larga escala.

No que diz respeito à forma como a adequação gramatical é descrita, avalia-se o uso variado de estruturas e também as suas inadequações. Pela forma como o item é descrito na grade, nas faixas avançadas e na intermediário-superior o uso variado de estruturas parece ser uma evidência comum para examinandos de alta proficiência, porém o que os diferencia é a frequência de inadequações. Como o item não é muito eficiente para discriminar as faixas avançado do avançado superior, pode ser que ao interpretar o desempenho oral, os avaliadores não diferenciem um desempenho com "poucas inadequações em estruturas complexas", descritor do avançado; do que apresentou "raras inadequações", descritor do avançado-superior. Se atentarmos sobre os descritores dos níveis avançados, percebe-se que quando se trata de avaliar a gramática utiliza-se "inadequações" e quando se trata do vocabulário utiliza-se "interferência", esta diferença na forma como os itens são descritos pode explicar porque o item *adequação lexical* tem uma performance um pouco melhor para diferenciar os examinandos avançados, seria a interferência algo mais saliente, mais perceptível para quem está avaliando a interação oral de examinandos mais proficientes? Por que a opção de descrever os itens lexicais da língua quanto a sua "interferência" e não "inadequação"? Se na grade sobre o item adequação lexical, no lugar de "interferências" estivesse a palavra "inadequações", o item teria demonstrado a mesma eficiência para separar o examinandos potencialmente entre os níveis *avançado* e *avançado-superior*? Outra hipótese diz respeito à tarefa em si, como os examinandos são provocados a falar sobre temas diversos, pode ser que o avaliador disponha de mais evidências para avaliar o uso diversificado de vocabulários do que o uso variado de estruturas gramaticais.

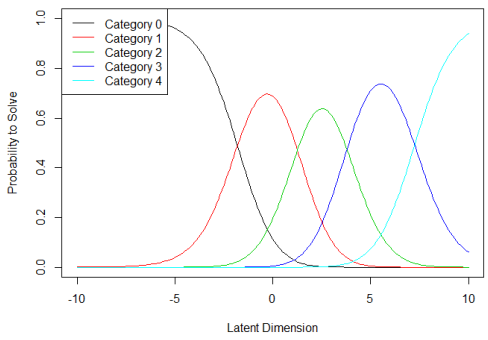

FIGURA 4 - Curva de característica do item *adequação lexical*
Fonte: Elaboração própria.

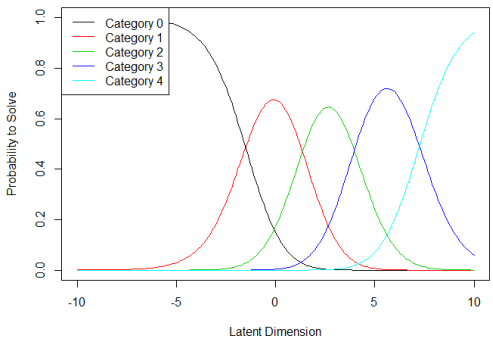

FIGURA 5 - Curva de característica do item *adequação gramatical*
Fonte: Elaboração própria.

3.7 Curva de característica do item *pronúncia*

O item *pronúncia* apresenta uma curva de característica mais próxima daquelas dos itens *adequação lexical* e *adequação gramatical*, do que dos três itens inicialmente analisados. Em pronúncia, avalia-se o quão adequado é o som, a entonação e o ritmo da fala do examinando. 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 da escala, pois a categoria está também 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, ainda que esteja próximo à sugestão de Eckes (2015) para quem o intervalo tem 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* (curva verde), 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, seja a de os descritores se referirem a algo que não é muito frequente na prática. No caso da *pronúncia*, parece provável que os avaliadores entre as notas 2, 3 e 4 optem pelas notas 2 ou 4, poucos atribuem nota 3 para pronúncia. Entre uma pronúncia "com algumas inadequações e/ou interferências de outras línguas.", "com inadequações e/ou interferências de outras línguas." e "com inadequações e/ou interferências frequentes de outras línguas.", que correspondem respectivamente aos descritores das faixas de avançado (nota 4), intermediário-superior (nota 3) e intermediário (nota 2) parece ser mais frequente classificar a pronúncia quanto às interferências frequentes ou ter interferências com algumas inadequações, alocando-os nas faixas avançado ou intermediário, ignorando a faixa intermediário-superior (nota 3). Sobre a avaliação da pronúncia, novamente a interferência de outras línguas, como o portunhol, parece estar pesando na hora de avaliar o desempenho do examinando.

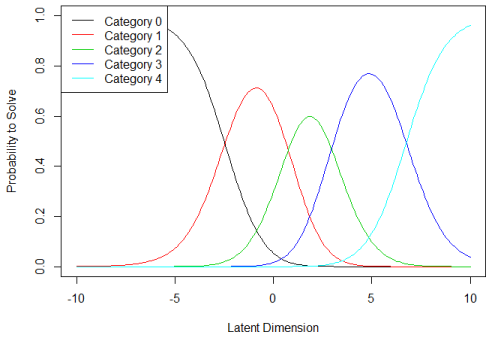

FIGURA 6 - Curva de característica do item *pronúncia*
Fonte: Elaboração própria.

3.8. Mapa Rasch e a discussão integrada dos itens da escala analítica

Uma maneira de analisar como os itens contribuem com a escala analítica como um todo é analisando o Mapa Rasch que sumariza todas as informações das curvas de característica dos itens. O mapa organiza os perfis de habilidades dos examinandos 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 são posicionadas a partir do valor de *threshold*. Pelas distâncias entre estes valores se pode avaliar a capacidade de discriminação das categorias em função dos valores de proficiência.

A figura do mapa 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 seis itens 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 escalas da prova oral do exame e a distância entre elas representa o intervalo entre os pontos de *threshold* que nos gráficos das curva de característica do item se referem à interseção entre as curvas.

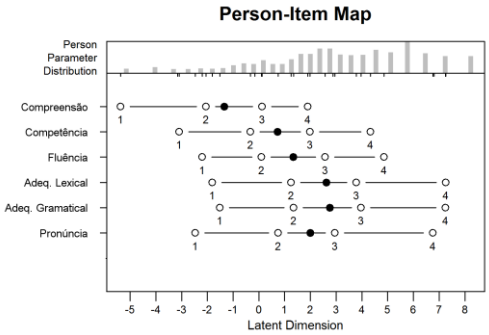

FIGURA 7 - Mapa do Rasch
Fonte: Elaboração própria.

Retomando o modelo da habilidade comunicativa proposto por Bachman (1990), teríamos os três primeiros itens ou subcomponentes da habilidade oral como parte da competência estratégica e os três últimos, da competência linguística. Os subcomponentes da competência linguística, nessa escala e nesse exame, apresentam medidas mais eficientes para diferenciar perfis mais altos da habilidade oral. 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. Salta aos olhos o deslocamento do item *compreensão*.

No contexto do exame, as habilidades são avaliadas de forma integrada, por isso a compreensão, assim como os demais aspectos da produção oral, são avaliados juntos na mesma situação e na mesma escala. No entanto, ao analisar as propriedades do item *compreensão* levanto diversos argumentos empíricos que sugerem que o item ou deva ser suprimido da escala ou deva ser feita uma grande revisão nos descritores e na tarefa ou situação de avaliação oral. Bachman (1990) afirma que o contexto de produção da linguagem interfere na interação entre os componentes da habilidade, por isso, acredito que a situação da prova oral seja insuficiente para avaliar o subcomponente *compreensão*, da mesma forma em que são avaliados os outros subcomponentes da habilidade oral, no que diz respeito à produção. Na situação da prova, está previsto que o examinando fale mais que o interlocutor, por isso, há mais evidências para avaliar a produção do que a compreensão oral. O examinando precisa apenas compreender algumas poucas perguntas colocadas pelo avaliador-interlocutor durante a interação. Se houvesse situações em que o examinando fosse desafiado a escutar falas com diversas complexidades de compreensão, talvez o item apresentaria outras propriedades. Uma hipótese possível para as propriedades desse item, apresentadas na análise, seria a de que a compreensão é um pressuposto para a interação oral no contexto da situação da prova oral, porque apenas examinandos com proficiência muito baixa tiram uma nota baixa nesse item. Ou seja, para demonstrar um desempenho adequado nas faixas mais altas de classificação poderíamos afirmar que a compreensão é um requisito.

Outros dois itens relacionados à competência estratégica (BACHMAN, 1990) são *competência interacional* e *fluência*, que apresentam um comportamento semelhante, mas quando comparados à *compreensão*, são mais difíceis. Ao analisar o mapa, posso inferir que os itens relacionados à

competência linguística tais como *adequação lexical*, *adequação gramatical* e *pronúncia* conseguem distinguir melhor as faixas *avançado* e *avançado superior*. 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*. As avaliações das não interferências de outras línguas e as adequações gramaticais e lexicais da fala do examinando no contexto da prova oral do Celpe-Bras são as que mais contribuem para diferenciar os níveis avançados. Embora Dell'Isola *et al.* (2003) tenham explicitamente afirmado que os conhecimentos da língua não seriam avaliados em questões isoladas para esse fim, após a análise das propriedades dos itens, a competência linguística é mais determinante para os examinandos que querem ser classificados em faixas mais elevadas do exame do que os subcomponentes da competência estratégica. Lembrando que os itens são avaliados dentro de um contexto de uso da língua, ou seja, a partir da interação oral entre examinando e avaliador-interlocutor avaliam-se as adequações linguísticas no que diz respeito à gramática, ao léxico e à pronúncia. Neste trabalho, discuti e levantei hipóteses para tentar explicar o funcionamento desses itens, levando em conta como o uso da gramática, léxico e pronúncia estão descritos na grade analítica. Outros estudos que explorem o comportamento do avaliador-observador ao atribuir notas podem contribuir para refutar ou explicar os dados e argumentos que sugerem que a competência linguística esteja mais relacionada às proficiências mais altas do que a competência estratégica, no contexto do Celpe-Bras.

4. Considerações finais

Neste trabalho busquei relacionar a proposta de avaliação da proficiência oral com as discussões teóricas sobre ensino e aprendizagem de línguas. O exame Celpe-Bras foi elaborado a partir das diretrizes teóricas da abordagem comunicativa cujo interesse é o uso da língua para produção e negociação de sentidos. Embora haja algumas propostas sobre a composição da competência comunicativa, pouco se discute sobre a integração entre seus subcomponentes. No contexto das avaliações de línguas, Bachman (1990) sugere que a interação entre os componentes da competência comunicativa é sensível às tarefas dos testes. Apoiando-se na proposta do autor, aproximo os itens *compreensão*, *competência interacional* e *fluência* à competência estratégica; e os demais itens, *adequação gramatical*, *adequação lexical* e *pronúncia*, à competência linguística. Tais itens, ou subcomponentes, compõem a grade analítica de avaliação da prova oral utilizada pelo avaliador-observador. Com o objetivo de estudar as propriedades desses itens e relacioná-las aos descritores da grade e à tarefa de avaliação oral, apresentei uma análise das medidas de dificuldade e discriminação de cada um dos itens.

Os três primeiros itens da grade analítica, ou seja, subcomponentes da competência estratégica, se mostraram eficientes para discriminar faixas intermediárias. O item *compreensão* foi o menos eficiente para discriminar examinandos após a primeira faixa certificada pelo exame, a *intermediário*.

Ou seja, é o item mais fácil da escala. *Competência interacional* se mostrou um pouco mais eficiente para discriminar as faixas *básico*, *intermediário* e *intermediário superior*. A nota de *fluência* diferencia melhor examinandos entre as faixas *intermediário superior* e *avançado* do que a nota de *competência interacional*. Os subcomponentes da competência linguística, os itens *adequação lexical*, *adequação gramatical* e *pronúncia* discriminam faixas de proficiências altas previstas pelo exame, embora cada item o faça de forma distinta. Tais itens são os mais difíceis da grade analítica. A análise apontou também a necessidade de revisão da faixa *intermediário* para os itens *adequação lexical* e *pronúncia*, sugerindo que seus descritores possam estar se referindo a algo que não é muito frequente na prática do contexto da avaliação oral do Celpe-Bras.

De maneira geral, a análise empírica das escalas de proficiência utilizadas para avaliação oral teve como pergunta de fundo a validade e a confiabilidade da escala analítica do exame Celpe-Bras que, no contexto deste trabalho, significa a capacidade de discriminação das faixas de proficiência. Ao avaliar separadamente cada item, a análise aponta que a nota de *compreensão* só é confiável ou eficiente quando se trata de examinandos com perfis baixos de certificação entre as faixas *sem certificação*, *básico* e *intermediário*. As notas medianas entre as faixas *intermediário e intermediário superior* são mais confiáveis quando atribuídas aos itens *competência interacional* e *fluência,* em comparação com o item *compreensão*. Notas altas ou que se referem às faixas *avançado* e *avançado superior* são confiáveis quando se trata da avaliação dos aspectos da *adequação lexical*, *adequação gramatical* e *pronúncia*.

A análise da curva de característica do item *compreensão* trouxe evidências de que a maneira como a compreensão oral está sendo avaliada precisa ser revista ou de que o item precisa ser suprimido da grade analítica. Há fortes indícios de que a compreensão, no contexto da situação da prova oral do Celpe-Bras, é um requisito para avaliação da produção oral dos examinandos.

Quanto à relação entre os resultados da análise e a forma como os itens estão descritos na grade, sugere-se que o tamanho do turno de fala seja o principal aspecto avaliado em *competência interacional* e quanta à *fluência*, as frequências de pausas. Os itens que se relacionam à competência linguística (*adequação lexical*, *adequação gramatical* e *pronúncia*) discriminam faixas de proficiências altas previstas pelo exame. Ao relacionar estes dados com a descrição da grade, pode ser que a não interferência de outras línguas, ou seja, o uso de um português menos híbrido ou o não portunhol, por exemplo, seja o aspecto mais saliente que possa estar influenciando na capacidade do item de diferenciar os examinandos avançados dos avançados superior. Outras pesquisas que investiguem o comportamento de avaliação poderiam trazer mais dados para a discussão das potencialidades e limites do portunhol e outros hibridismos linguísticos, no contexto desta prova. Verifica-se ainda a necessidade de revisão da faixa *intermediário* para os itens *adequação lexical* e *pronúncia*, sugerindo que seus descritores possam estar se referindo a algo que não é muito frequente na prática do contexto da avaliação oral do Celpe-Bras.

Concluo, a partir da análise, que a interação entre os subcomponentes da grade analítica da prova oral está diretamente relacionados com a situação de avaliação e com a forma como os subcomponentes estão descritos na escala, corroborando ainda mais com a hipótese de que a tarefa

interfere na forma como os subcomponentes da habilidade linguística interagem (BACHMAN, 1990). Neste trabalho, por meio de dados empíricos, tento costurar as relações entre os subcomponentes da habilidade oral. Uma das limitações da análise foi trabalhar com uma amostra aleatória, seria interessante que outras pesquisas replicassem a análise tendo como base um *corpus* formado por falantes de espanhol como primeira língua, por exemplo. O processo de atribuição de notas, especialmente no contexto da avaliação oral, é extremamente complexo. Ao avaliar a proficiência oral, variados aspectos estão em interação, impactando na maneira como os construtos são operacionalizados. Por isso, relacionar os construtos relacionados à proficiência oral com diversos perfis de avaliadores ou examinandos é uma das questões a serem exploradas em trabalhos futuros.

Espero ter contribuído para o debate sobre os subcomponentes da proficiência oral, em especial, no contexto do Celpe-Bras.

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.

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

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. *Certificado de Proficiência em Língua Portuguesa para Estrangeiros*: grades de avaliação da prova oral. Brasília, 2019.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Edital n. 1, de 28 de janeiro de 2016 - de abertura de inscrições do exame Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras/2016.1)*, 2016 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.

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

CANAGARAJAH, S. Translingual practice as spatial repertoires: Expanding the paradigm beyond structuralist orientations. *Applied Linguistics*, Oxford, v. 39, n. 1, p. 31-54, 2018. Doi: https://doi.org/10.1093/applin/amx041

CANALE, M. From Communicative competence to communicative language pedagogy. In: RICHARDS, J.; SCHMIDT, R. (Org.) *Language and Communication*. Londres: Longman, 1983, p. 2-27.

DELL'ISOLA. R.; SCARAMUCCI, M. R. V.; SCHLATTER, M.; JUDICE, N. A avaliação de proficiência em português língua estrangeira: o exame Celpe-Bras. *Revista Brasileira de Linguística Aplicada*, v.3, n. l, p. 153-184, 2003.

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

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

FERREIRA, L. M. L. *Habilidades de Leitura na Proposta de Interação Face a Face do Exame* CELPE-BRAS. Dissertação Mestrado. Belo Horizonte: UFMG, 2012.

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

FULCHER, G.; DAVIDSON, F. *Language testing and assessment*: an advanced resource book. Routledge: New York, 2007. p. 91-114.

JASPER. J. The transformative limits of translanguaging. *Language & Communication*. Elsevier, v. 58, p. 1-10, 2018. DOI: 10.1016/j.langcom.2017.12.001

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 mai. 2018.

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

McNAMARA. Language Testing. In: DAVIES, A.; ELDER, C. *The handbook of applied linguistics.* Oxford: Blackwell, 2004, p. 763-778.

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

SCARAMUCCI, M. V. R. 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, p.77-90.

SZABÓ, G. *Applying Item Response Theory in Language Test item bank building*. Frankfurt: PeterLang, 2007.

SCHOFFEN, J. R. *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.

SCHLATTER, M.; SCARAMUCCI, M. V. R.; PRATI, S. Celpe-Bras and CELU proficiency exams as political acts in Brazil and Argentina. In: ALTE 3rd - International Conference Cambridge 2008, 2008, Cambridge. *Programme of the ALTE 3rd - International Conference Cambridge 2008*. v. 1. p. 64-64, 2008.

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: 20 mai. 2018.

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: 20 mai. 2018.