revista.pucsp.br/esp | ISSN: 2318-7115

The Especialist

Volume 41 | Número 5 | Ano 2020

A validade segundo psicometristas e linguistas aplicados e as entrevistas de proficiência oral

Validity according to psychometric and to applied linguistics and the oral proficiency interview

Laura Márcia Luiza FERREIRA (UNILA)[1]

RESUMO

Psicometristas e linguistas aplicados colaboram para o debate sobre as avaliações de línguas; os primeiros são especialistas em como medir; e os segundos, em o quê medir. Neste artigo, argumento que há uma tensão entre os dois grupos em torno da definição do conceito de validade. Apresento uma breve revisão do conceito de validade a partir da perspectiva da psicometria de forma a relacioná-lo com as definições de validade apresentadas por linguistas aplicados, especialistas em avaliação de línguas. Em seguida, exemplifico o debate sobre a validade no contexto das avaliações de proficiência oral a partir dos conceitos de autenticidade, teste direto e indireto nos dois maiores exames de larga escala da América do Sul: o Celpe-Bras e o Celu. Ao final, argumento que os linguistas aplicados tendem a debater a validade a partir do desenho dos instrumentos, eclipsando a necessidade de levantar evidências empíricas sobre como o desenho da prova interage com outros aspectos do teste.

Palavras-Chave: Validade, Avaliação em línguas, Entrevistas de proficiência oral

ABSTRACT

While building a language test, psychometrists and applied linguists negotiate how to measure and what is to be measured, concerned with language. In this paper, I review how language assessment specialists and psychometric specialists define validity in order to investigate the face validity as a source of evidence for a language test validity argument. Then I focus on how concepts of authenticity and directness of the task are related to validity arguments of oral proficiency interviews at Celpe-Bras and Celu exams. After revising the tensions related to validity of oral tests, I argue that applied linguists tend to overestimate the role of design on the validity argument so that may overlap discussions about how this design interacts with other aspects of the test.

Keywords: *Validity, Language assessment, Oral proficiency interviews.*

[1] Universidade Federal da Integração Latino-Americana, Foz do Iguaçu, Paraná, Brasil. Centro Interdisciplinar Letras e Artes; ORCID: http://orcid.org/0000-0001-7632-0834 e-mail: laura.ferreira@unila.edu.br

1. Introdução

As avaliações de larga escala de proficiência linguística estão cada vez mais presentes no ofício dos professores de línguas estrangeiras/adicionais, devido ao efeito retroativo dos exames[2]. Os testes podem influenciar na elaboração de cursos e materiais e, consequentemente, na maneira de ensinar a língua. Nesse sentido, uma compreensão sobre os exames de proficiência que extrapole os objetivos de formar professores para a aplicação de testes e a preparação dos estudantes, se faz necessário. Se o ofício dos professores de línguas é fortemente influenciado pelos exames de larga escala, é preciso que os professores identifiquem estas influências e desenvolvam meios para ensinar línguas se posicionando conscientemente sobre as potencialidades e limitações que os testes exercem no seu trabalho.

O que tenho visto, ao longo da minha experiência como professora de línguas, é que tendemos a copiar e repetir padrões das avaliações de larga escala, sem, contudo, fazer antes uma reflexão sobre o quão adequado é, para os objetivos de ensino de um determinado curso e em um contexto institucional, elaborar um instrumento de avaliação baseado em exames de larga escala. No contexto de formação de professores de línguas adicionais, há algumas discussões sobre a aprendizagem da avaliação, ou seja, como promover o desenvolvimento por parte dos professores pré-serviço, especialmente, de habilidades e conhecimentos sobre elaboração, análise, aplicação e gestão de instrumentos de avaliação de larga escala ou avaliações de aprendizagem[3].

Luckesi (2011) adverte que ainda não sabemos avaliar e que é preciso que se invista, no âmbito da formação dos professores, na aprendizagem da avaliação. Para o autor, aprender a avaliar é, não só aprender conceitos teóricos, mas, sobretudo, desenvolver práticas de avaliação. Sobre a dicotomia teoria e prática no contexto da aprendizagem da avaliação, Luckesi (2011) afirma que "aprender conceitos é fácil, o difícil mesmo é passar da compreensão para a prática" (p.30). Entendo que tanto a teoria quanto a prática estão presente nos processos de elaborar, desenvolver, gerir ou avaliar em contextos de testes padronizados, bem como as avaliações de sala de aula; e que é um dos desafios do professor de línguas compreender estes processos. No contexto da avaliação, assim como no do ensino, a teoria e a prática são indissociáveis. A complexa tarefa de elaborar instrumentos de avaliação com variadas funções envolve a operacionalização de conceitos específicos deste campo de estudo, tais como: critério, construto, validade

[2] Efeito retroativo ou washback effect se refere ao impacto do teste no ensino de línguas ou como um teste de línguas influencia questões de ordem sociais.

[3] Segundo Luckesi (2011), avaliação de aprendizagem está subordinada ao planejamento do curso e aos contextos de ensino-aprendizagem. A partir dos resultados da avaliação da aprendizagem o professor avalia o andamento do curso e a pertinência do seu planejamento pedagógico.

e confiabilidade. Estudar tais conceitos, envolve analisar as práticas de avaliação.

Apresento um estudo de revisão bibliográfica sobre o conceito de validade por ser fundamental para a compreensão da teoria e prática das avaliações de línguas. Independente do contexto das avaliações, compreender aspectos do conceito de validade se faz necessário. O conceito de validade foi desenvolvido no campo de estudos chamado Teoria das Medidas. A partir do desenvolvimento dos exames de proficiência de larga escala com impactos sociais relevantes, os termos da área da Psicometria foram sendo apropriado pelos Linguistas Aplicados durante os processos de elaboração, análise e reflexão sobre os instrumentos de avaliação linguística externas.

Neste texto, será discutido o conceito de validade definidos por psicometristas, bem como a maneira como os linguistas aplicados se apropriaram do termo. Em seguida, um aspecto controverso da validade, que é a validade de conteúdo ou de face[4], e sua relação com a autenticidade e o formato direto das tarefas de avaliação, será debatido a partir de uma revisão bibliográfica da área e de exemplos dos desenhos de avaliação oral dos exames: Certificado de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras e o Certificado de Español: Lengua y Uso - CELU.

Abaixo defino brevemente o que é avaliação, critério e construto.

2. Avaliação, critério e construto

A avaliação é um processo de coleta sistemática de informações que é semelhante em vários contextos. Taras (2010) a define como uma "combinação de dados de desempenho com um conjunto ponderado de escalas de objetivos para gerar listas comparativas ou numéricas, com base (a) nos instrumentos de coleta, (b) nas ponderações, (c) na seleção dos objetivos." (Scriven *apud* Taras, 2010, p.125). A avaliação se assemelha a uma investigação que levanta inferências sobre alguma coisa.

No contexto de avaliação de línguas, McNamara (2008) explica que a primeira etapa seria a coleta de dados; a segunda, a de inferência sobre o que o examinando é capaz de fazer na língua; na terceira, usa-se as informações sobre o que o examinando é capaz de fazer para tomar uma decisão que pode ser aprová-lo ou não em um curso de línguas, conceder ou não uma bolsa de estudos, cumprir ou não um dos requisitos para o processo de validação de diplomas estrangeiros, nivelá-lo em alguma faixa de proficiência previamente estabelecida; ou seguir ou não com o conteúdo em um contexto de um curso.

O processo de avaliar compreende, então, a fase da coleta, que seria a elaboração e aplicação do instrumento de avaliação; a fase da inferência, que diz respeito à correção da prova e, ao final, o uso dos resultados da avaliação para nortear alguma decisão. Um instrumento deve apresentar coerência entre

[4] Assim como em Messick (1987), validade de conteúdo e de face são entendidas aqui como sinônimo.

essas três fases do processo de avaliação. A coerência entre as fases da avaliação e entre os diversos aspectos de cada fase está relacionada com a validade. A validade diz respeito a todas essas etapas de avaliação, porque tanto o instrumento, a forma de avaliar e o uso dos resultados devem ser válidos (MESSICK, 1987), ou seja, as notas geradas pelo instrumento devem ser coerentes com o quê, o como e o porquê que se avalia. A depender da fase do processo de avaliação que se discute, escolhe-se um aspecto da validade para aprofundar a análise teórica ou empírica. A validade de conteúdo ou de face, por exemplo, se refere mais ao instrumento de coleta de dados, ou seja, ao desenho da avaliação do que com as escalas e ponderações da medida ou com as consequências dos resultados do teste, embora o desenho do teste esteja conectado com todas as fases do processo de avaliação. Fulcher (2003) explica que o desenho da avaliação trata não só dos comandos das tarefas, atividades ou questões, mas também da forma como os itens são avaliados. Aqui, interessa discutir os conceitos de validade de face, autenticidade e avaliações diretas no contexto de elaboração dos instrumentos de avaliação oral de línguas adicionais.

Antes de entrar mais detalhadamente na discussão sobre validade, é preciso esclarecer alguns conceitos básicos de avaliação, a saber: critério e construto. As avaliações orais de línguas operacionalizam construtos sobre o que é saber falar uma língua. O teste, ou o instrumento, coleta evidências por meio do desempenho oral nas respostas às tarefas, ou itens, que serão corrigidas por professores-especialistas, que interpretam o desempenho linguístico oral em uma determinada língua, ou seja, inferem o que o examinando provavelmente é capaz de fazer oralmente no mundo real. Para gerar a inferência de forma a garantir validade e confiabilidade, é preciso que os elaboradores definam o que é saber interagir oralmente em uma língua e como captar as evidências do desempenho oral no contexto do teste. A definição sobre o que é saber falar uma língua se fundamentará nas discussões teóricas da área de linguística aplicada sobre o desenvolvimento da proficiência oral em língua adicional. Há várias discussões sobre o que é fluência (CHAMBERS, 1997; NATION, 1989; FULCHER, 2003) e competência interacional (HYMES, 1972), por exemplo, e o papel dessas habilidades no desenvolvimento da proficiência oral que podem fundamentar a elaboração de instrumentos. Tais discussões teóricas da área fundamentam a maneira como os testes estão estruturados e, principalmente, como serão as tarefas e como as respostas serão corrigidas. Tais escolhas teóricas que fundamentam o instrumento estão relacionadas com a validade de construto. A validade de construto tem a ver com o quão bem o teste operacionaliza o modelo teórico ou conceitos teóricos utilizados na sua fundamentação em todas as etapas do processo: coleta, inferência e uso dos resultados.

A avaliação oral é um grande desafio para linguistas (FULCHER, 2003; BYGATE, 2009) e psicometristas (ECKES, 2015) e um dos motivos é o fato de não termos ainda um modelo teórico que explique os estágios e como um falante desenvolve sua proficiência. O que normalmente se faz é compor

o construto da proficiência oral a partir de conceitos teóricos de diversas perspectivas teóricas como fluência oral, pronúncia, adequação lexical, dentre outros. O desenho do teste deve ser eficiente e coerente com as escolhas teóricas que foram feitas para fundamentar o desenho do instrumento. Dizendo de outra forma, as questões ou a tarefa do teste devem capturar as evidências da proficiência oral que foram teoricamente definidas pelos avaliadores. Se a fluência, pronúncia e adequação lexical são os construtos ou conceitos teóricos que interessam para avaliar a proficiência oral, o desenho do teste deve coletar evidências destes construtos.

As grades de avaliação oral dos exames de português (Celpe-Bras) e de espanhol (CELU), por exemplo, contemplam tanto itens mais fortemente relacionados aos aspectos linguísticos formais das línguas, tais como: *pronúncia*, *gramática* e *léxico*; quanto aqueles itens mais ancorados nos contextos de interações orais, a saber: *objetivo y interacción* (CELU), *competência interacional*, *fluência (*Celpe-Bras). No exame CELU, '*objetivo y interacción'* está diretamente relacionado ao desenho da tarefa, uma vez que diz respeito ao quanto o examinando cumpriu com as tarefas propostas durante a interação face a face, especialmente a parte de simular uma situação real com papéis definidos pelo comando da tarefa, do tipo *juego de roles*.

Fulcher (2003), ao historicizar os testes orais, comenta que na Universidade de Cambridge havia um exame Certificado de Proficiência em inglês de 1913 que estava organizado de forma que o examinando era submetido a meia hora de ditado, seguida de meia hora de atividades que envolviam leitura e uma interação face a face. Porém, o que era avaliado no instrumento era a pronúncia apenas. Neste caso, embora esteja previsto no desenho do instrumento uma interação face a face, o construto de proficiência oral avaliado era apenas o da pronúncia, ou seja, saber falar significa saber pronunciar as palavras, no contexto deste exame. Se o objetivo da avaliação é a pronúncia, a leitura em voz alta, por exemplo, é uma tarefa que permite avaliar o construto de proficiência oral definido no exame. No entanto, se saber falar, envolve fluência, competência interacional etc., é preciso que a tarefa capture evidências destes construtos e, neste caso, uma interação face a face seria um desenho mais coerente.

McNamara resume a relação entre teste, construto e critério no quadro abaixo:

Quadro 01: Teste, construto e critério

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

 DOI: 2318-7115.2020v41i5a2

teóricos

Fonte: McNamara, 2004, 765p. (tradução da autora)

Cabe ressaltar que critério é um termo técnico da área de avaliação relacionado ao significado da inferência que se faz a partir dos resultados do exame. Vamos imaginar que queremos saber se o estudante sabe ler na língua adicional para seguir seus estudos na pós-graduação, que envolve muita leitura em língua estrangeira de textos de uma determinada área para fins de realização da pesquisa. Neste contexto, o critério é leitura em língua estrangeira de textos técnicos-acadêmicos de uma área do conhecimento específica. Os avaliadores elaborariam um teste para observar evidências sobre o quanto o examinando sabe ler na língua adicional, no contexto de uma pós-graduação. Para tanto, os especialistas se apoiariam em teorias de leitura para elaborar o instrumento. Suponhamos que os especialistas se apoiassem numa Abordagem de Leitura, baseando-se nos Modelos Interativos (GRABE e STOLLER, 2002) em que interessa avaliar as estratégias de leitura de leitores experientes, como saber ou não fazer uma inferência, a partir de uma informação disponível no texto. Neste caso, elaborariam itens de avaliação com o objetivo de saber se o examinando é capaz ou não de fazer inferências de leitura autorizadas pelo texto. O critério, neste exemplo, é a leitura em língua estrangeira/adicional. A forma como o critério foi operacionalizado em um teste está relacionado com as escolhas teóricas dos especialistas, que, neste caso, a partir das teorias de leitura baseadas no modelo interativo, prevê que saber ler é saber fazer inferências autorizadas pelo texto. Ou seja, saber fazer inferências faz parte do construto de leitura em língua estrangeira/adicional que foi operacionalizado no instrumento. Neste caso, poderíamos editar a tabela de McNamara (2004) para o nosso teste hipotético da seguinte forma:

Quadro 02: Teste, construto e critério de um instrumento de leitura em língua estrangeira/adicional

teste		construto		critério
Itens de avaliação sobre leitura em língua adicional que envolve saber fazer uma inferência observável	desenho do teste inferências	Modelo interativo de leitura, estratégias de leitura, inferências autorizadas pelo texto via modelos teóricos	sobre	Saber ler textos técnicos e acadêmicos de alguma área do conhecimento específica não-observável

Fonte: elaboração nossa a partir de McNamara, 2004, 765p.

Elaborar um instrumento de avaliação, envolve fazer pesquisa sobre o uso futuro da língua para se chegar a um critério. Estabelecido o critério, é preciso fazer escolhas teóricas sobre o construto que se pretende operacionalizar por meio do desenho das questões. Nos casos dos testes de língua, os elaboradores precisam lançar mão de um estudo sobre como as pessoas aprendem línguas e quais são as

evidências deste aprendizado que podem ser identificadas por meio de itens de avaliação. A leitura, por exemplo, é um tema bastante estudado na área, o desenvolvimento da proficiência oral, no entanto, é um tema que apenas recentemente começou a ser estudado, por isso, como já comentado anteriormente, fazer provas orais é um grande desafio, independente da língua avaliada.

A validade está diretamente relacionada ao critério e ao construto. De maneira geral, a validade está relacionada à eficiência do instrumento de gerar informações pertinentes para o cumprimento do propósito do exame, ou seja, a coleta das evidências e a inferência sobre o que o estudante é capaz de fazer devem estar relacionadas com o propósito do exame. Por exemplo, se o instrumento tem a função de nivelar ou dispensar os estudantes no contexto de um programa de línguas que seja formado por cursos que exigem uma progressão da proficiência, é preciso que o exame esteja organizado de forma a contemplar tal progressão curricular de maneira que ofereça níveis limiares de corte que sejam coerentes com os níveis dos cursos. As evidências geradas na coleta do instrumento de nivelamento devem ser interpretadas a fim de cumprir o propósito, ou seja, de alocar o estudante no curso adequado para seu nível de proficiência. Neste caso específico, uma possibilidade de elaboração do exame seria prever no instrumento um determinado desempenho no teste que se assemelha a dos egressos de um determinado curso para determinar com mais precisão o critério do teste e também o construto, que neste caso, poderia estar relacionado com as discussões teóricas sobre letramento acadêmico. Por exemplo, se o curso tem como objetivo desenvolver a escrita acadêmica, o critério deveria ser a escrita acadêmica e, poderia ser possível pedir aos examinandos que pretendem ser dispensado no curso que escrevam naquela língua em situações acadêmicas, por exemplo, por meio da redação de resumos acadêmicos, resumos de planejamento do trabalho acadêmico (*abstract*), resenhas, etc.

Em geral, a relação entre critério e construto e os instrumentos pode estar definida nas especificações, nos manuais, na página dos organizadores do exame, nas fichas de avaliação, etc. A validade está também relacionada com o julgamento sobre a pertinência entre critério, construto e instrumento de avaliação, mas, especialmente, no contexto das avaliações orais, a discussão sobre validade vai além da análise documental das diretrizes dos testes e dos instrumentos em si (MESSICK, 1987; 1996).

McNamara (2004) e Fulcher (2003), especialistas em avaliação da língua inglesa, afirmam 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. A história do desenvolvimento dos testes de línguas orais de larga escala refletem, em vários momentos, um conjunto de ideologias linguísticas, ou seja, crenças sobre

ensino e aprendizagem de línguas, ao operacionalizar construtos de língua, bem como sobre o papel dos testes nas políticas imigratórias anti-humanistas dos Estados Unidos (FULCHER, 2003), na Austrália e em Israel (CONSOLO, 2005), nas políticas educacionais inglesas (FULCHER, 2003) e nas políticas de cooperação de Brasil e Argentina (SCHLATTER *et al*, 2008). Os testes orais foram desenvolvidos inicialmente no contexto do mercado linguístico para ensino de língua inglesa com diferentes finalidades. A partir da década de 90, impulsionados pelo Tratado de Assunção, os testes de larga escala de português e espanhol começaram a ser desenvolvidos por especialistas brasileiras e argentinas (SCHLATTER *et al*, 2008).

Em linhas gerais, as avaliações de desempenho são organizadas de forma a permitir que o examinando demonstre o que sabe fazer na língua (MCNAMARA, 2004). Exames internacionais de línguas como Celpe-Bras e o CELU são exemplos de avaliação de larga escala de desempenho. No contexto de promoção da integração e cooperação entre países do Mercado Comum do Sul (Mercosul), os exames foram elaborados por especialistas da área com intensa colaboração entre as responsáveis pelo exame brasileiro e pelo argentino. Schlatter *et al.* (2008) descrevem o contexto de elaboração dos exames e assumem que tanto o Celpe-bras quanto o CELU partilham de critérios, construtos e semelhanças em seus desenhos de avaliação. O critério da avaliação destes exames é o uso da língua em situações reais e os construtos estão relacionados, de forma geral, às discussões em torno da abordagem comunicativa do ensino de línguas (DELL'ISOLA *et al.*, 2003; SCARAMUCCI, 2001) bem como as teorias de gêneros (SCHOFFEN, 2009), que embasam fortemente a etapa escrita dos exames.

A partir das discussões sobre os métodos comunicativos, os linguistas aplicados tendem a conceber as avaliações de desempenho de línguas de forma a simular situações reais de uso daquela língua em determinado contexto, em geral, as institucionais. Segundo Fulcher (2003), uma primeira tentativa de definir uma avaliação oral foi proposta por Kaulfers em um artigo publicado 1944. 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 para atribuir notas às interações 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.

Se por um lado, os exames de desempenho, como o descrito por Kaulfers (1944 *apud* FULCHER, 2003), procuram simular ações reais da língua em situações de prova para ter acesso a evidências de performance mais próximas das de interações autênticas, por outro a correção de questões abertas implica

lançar mão de estratégias para que a nota seja confiável. Um dos aspectos da confiabilidade dos testes é a garantia que o resultado ou a nota do examinando não varia por questões alheias a sua proficiência. Explico: o examinando não pode ter notas diferentes em duas edições ou aplicações do exame que se propõe medir o mesmo construto, por exemplo. Se um examinando faz duas entrevistas de proficiência oral do mesmo exame, uma após a outra, em que se varia os entrevistadores e avaliadores; espera-se que a nota nas duas aplicações seja próxima. Se as notas forem muito diferentes, há um problema de confiabilidade que precisa ser estudado, identificado e corrigido, para que o examinando não seja prejudicado. Em testes de larga escala, é desejável que a nota seja o mais estável possível, ou seja, é preciso garantir condições para que um mesmo texto ou mesma interação oral, por exemplo, seja avaliado por mais de um avaliador e que as notas atribuídas, por avaliadores diferentes e de forma independente, sejam semelhantes. Por este motivo, as provas são corrigidas por pares e deve haver controle de discrepâncias de notas, para consertar falhas pontuais no processo de correção. Para garantir que todos envolvidos na correção estejam interpretando as evidências de performance de forma similar, é preciso que todos envolvidos no processo de correção compreendam o construto do exame e os itens avaliados de forma mais ou menos consensual, por isso os avaliadores fazem treinamentos específicos para atuarem nas correções das respostas abertas. Os protocolos de refino e controle da correção de textos escritos já é uma prática recorrente no processo de avaliação externa de línguas, mas, e quando a prova é oral?

Fulcher (2003) explica que as críticas aos primeiros testes orais focavam muito mais os itens avaliados, os descritores e a escala do que a adequação da situação de avaliação da proficiência oral em si. A problematização sobre a validade das situações de avaliação, especialmente, a entrevista de proficiência oral (EPO) para avaliar as habilidades de interação oral em língua adicional veio posteriormente.

Da perspectiva dos estudos da validade e da confiabilidade, quanto menos controlado é o processo de coleta de informações, menos válido e confiável poderiam ser seus resultados (MESSICK, 1987; BACHMAN, 1990). Por exemplo, a situação de ditado para avaliar a pronúncia pode não ser tão autêntico quanto a simulação de uma interação oral, mas é um processo de coleta de informações mais controlado do que uma interação oral, especialmente se, por algum motivo, o objetivo da avaliação for o da pronúncia de algumas palavras específicas. Instaura-se, então, uma tensão entre duas características salutares para o desenho dos testes de língua: a autenticidade do desenho do instrumento e a validade.

Apresento, a seguir, uma breve discussão sobre validade.

3. A validade segundo psicometristas e linguisticas aplicados

Antes de começar o debate sobre encontros e desencontros na definição de validade para psicometristas e linguistas aplicados, é preciso ter em mente que estamos falando de especialistas de duas áreas diferentes que apresentam uma definição do mesmo termo: a validade. Psicometristas e linguistas aplicados precisam cooperar e trabalhar interdisciplinarmente na construção de avaliações de larga escala. Por isso, é normal que os linguistas aplicados dêem mais ênfase em alguns aspectos da validade que para os psicometristas são irrelevantes. Neste ponto, os psicometristas são os especialistas em como medir; e os linguistas aplicados, em o quê medir. Cabem aos linguistas definir o critério e o construto dos testes. A forma como o critério e o construto de língua/linguagem serão operacionalizadas no instrumento deve ser orientada pelas técnicas de mensuração, desenvolvidas no âmbito da psicometria. As experiências de mediação das decisões sobre o como e o quê medir no momento da construção dos instrumentos fomentam o debate sobre a validade dos instrumentos de avaliação linguística. De maneira geral, a validade está relacionada à eficiência do instrumento de gerar informações pertinentes com o critério do exame, ou seja, a coleta das evidências e a inferência sobre o que o estudante é capaz de fazer devem ser coerentes com o propósito do exame.

No documento da AERA (2014), *Standards for Educational and Psychological Testing*, afirma-se que a validade se refere ao grau em que evidências e teorias embasam ou justificam o uso dos resultados dos testes para seus propósitos definidos. Há vários aspectos da avaliação que podem ser analisados quanto à sua validade. O processo de validação é um processo em construção constante que carece de evidências teóricas e empíricas na formulação de argumentos a favor e contra a validade dos testes (MESSICK, 1987, 1996; FULCHER, 2003, BACHMAN, 1990; MCNAMARA, 2004).

Messick (1987), ao historicizar a definição de validade nos *Standards*, afirma que no documento dividia-se o conceito de validade em vários, tais como: validade substancial, validade preditiva, dentre outros, a depender do tipo de evidência ou argumento utilizado no estudo de validação. A última versão do documento de 2014, assim como na Teoria da Validade, escrita pelo psicometrista Messick em 1987, que impactou consideravelmente os estudos de validade; convergem no sentido de definir a validade como um conceito unitário, porém a depender da evidência apresentada, um ou outro aspecto da validade será ressaltado. Trata-se, portanto, de um conceito único, porém multifacetado.

Nos estudos de avaliação em línguas, os linguistas tendem a apresentar o conceito de validade, apoiando-se nos psicometristas, porém ao discorrer sobre os aspectos da validade não há uma convergência na área sobre o significado dos termos. Por exemplo, na Teoria de Validade, Messick (1987) afirma que todos os aspectos da validade dizem respeito à validade de construto, porque o que interessa é saber o quão bem o construto (ou outros construtos irrelevantes) estão sendo representados nas diferentes etapas do processo de avaliação. Se após uma análise empírica ou teórica há evidências de que o teste está

medindo construtos irrelevantes, isso significa que pode ser que o teste não esteja medindo o que deveria medir, porque está medindo um construto irrelevante para seus propósitos. Para o autor, a validade de construto é sinônimo de validade porque a validade envolve analisar o quanto os construtos teóricos relevantes ou irrelevantes estão sendo operacionalizados no instrumento em si, nos processos de julgamento da performance, nas escolhas sobre os pontos de corte das faixas de proficiência, bem como as consequências sociais e individuais do resultado da avaliação. Ou seja, o construto que se quer avaliar deve estar representado em todas as fases da avaliação, por isso a validade é, sobretudo, uma investigação ou argumentação teórica e ou empírica sobre como o construto do teste está sendo operacionalizado.

Neste trabalho, entendo a validade como um conceito unitário (AERA, 2014; MESSICK,1987 ,1996) e vale lembrar que alguns linguistas aplicados como Fulcher (2003) e McNamara (2004) também definem validade na esteira das ideias dos psicometristas, porém nem todos da mesma forma.

Embora Hughes (1989), Bachman e Palmer (1996) e Brown e Abeywickrama (2004) tendem a enfatizar a validade de construto como sendo um termo geral ou mais importante, ao fragmentar os aspectos da validade, principalmente, em validade de face, validade de conteúdo, autenticidade ou formato direto ou indireto dos itens avaliados etc.; os linguistas podem gerar compreensões diversas sobre a relação entre o construto e sua relação com os desenhos dos instrumentos de avaliação. Alguns linguistas aplicados relacionam os desenhos dos testes, sua autenticidade ou o fato de serem diretos ou indiretos com a validade de face ou de imagem, validade de conteúdo ou validade de critério, gerando uma grande confusão sobre o significado dos termos e, consequentemente, sobre o conceito de validade.

Em *Test for Language Teachers*, Hughes (1989) afirma que validade de construto é um termo geral para validade no caso de testes, no entanto, a define como a relação entre nota e construto, ao mesmo tempo em que define validade de face como sendo um julgamento se o teste mede o que tem que medir. Hughes (1989) não deixa claro como as evidências de um julgamento do desenho do teste por especialistas iriam fundamentar as medidas geradas pelo instrumento sem avaliar as respostas aos itens. Avaliar se o teste mede o que tem que medir é muito mais complexo do que julgar o desenho do teste. Messick (1987) organiza a definição de forma diferente, aproximando a validade de face da validade de conteúdo. Em um texto de 1996, o psicometrista entre no debate específico sobre os termos autenticidade, avaliações diretas e indiretas e sua relação com a validade no contexto dos exames de línguas. A validade de conteúdo é ponto de divergência entre os psicometristas e alguns linguistas aplicados. Um ponto central para o debate é o fato que os linguistas tentem a desmembrar a validade de conteúdo em autenticidade e o formato direto ou indireto da situação de uso da língua avaliada no teste. Cabe lembrar que, especialmente após o movimento comunicativo no ensino de línguas, linguistas e professores passaram a elaborar aulas e

avaliações que refletem usos reais da língua, por isso ao operacionalizar de construtos e critérios em exames de larga escala, há uma atenção dispensada à eficiência dos itens ao simular usos reais da língua e, consequentemente, a validade passa a estar relacionada também à autenticidade das tarefas.

Messick (1987) explica que tradicionalmente a validade de conteúdo diz respeito à relação do conteúdo do teste com as situações da sala de aula ou o assunto estudado. Da mesma forma, os linguistas afirmam que a validade de conteúdo está relacionada ao programa do curso, aos objetivos de ensino. Cabe uma ressalva aqui, pois os testes de línguas de larga escala se baseiam, normalmente, na definição de proficiência ou de critério (*criterion-referenced*) definida no próprio contexto da avaliação. O propósito dos testes de proficiência não está atrelado a um processo prévio de ensino, mas a uma expectativa de interação bem sucedida em algum contexto especificado pelo exame. Validade de conteúdo, para a maioria dos linguistas, é um aspecto da validade restrito à avaliação de sala de aula. Em Messick (1996) e no documento da AERA (2014) tanto a validade de conteúdo quanto a validade de face estão relacionadas à escolha e pertinência dos itens que serão elaborados no instrumento para capturar as evidências necessárias para gerar o resultado. Dessa forma, entendo que a validade de conteúdo e de face são conceitos similares e que estão relacionados ao formato, à tarefa ou situação de avaliação, no caso da prova oral.

No contexto de testes de línguas, a validade de conteúdo ou validade de face diz respeito à relação do conteúdo da prova com as situações da sala de aula estudadas ou com as situações de uso da linguagem as 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 representatividade de cada uma das tarefas ou atividades do teste fornece evidências que fundamentam a sua relevância para gerar informações pertinentes. A finalidade desse julgamento é avaliar se o teste 'parece' que mede o que tem que medir, trata-se de um julgamento do desenho do teste e não da avaliação como um todo, que envolve recolher outras evidências sobre adequação do processo de correção, do peso dos itens na composição da nota, se todos os examinandos têm as mesmas chances de demonstrarem o que sabem, etc. Os linguistas aplicados, especialmente Hughes (1989), tendem a valorizar este aspecto da validade, como se o julgamento de especialistas fosse suficiente para atestar a validade de um instrumento como um todo.

O psicometrista Messick (1987), por outro lado, adverte que tais julgamentos não necessariamente geram evidências que fundamentam as inferências feitas a partir da nota. Isto ocorre porque eles não contemplam na análise do aspecto do teste, como 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 avaliações sobre a relevância e a representatividade da tarefa feitas pelos especialistas influenciam a natureza da nota, Messick (1987) considera que a validade de conteúdo e de face não é qualificada como

um estudo sobre a validade propriamente dito, uma vez que determinar o que está sendo medido exige outros tipos de análises.

Neste ponto, o recente documento da AERA (2014) diverge da Teoria de Validade de Messick (1987), ao afirmar que a etapa de elaboração de teste que normalmente resulta na redação de uma especificação da forma e dos conteúdos do teste, cuidadosamente selecionada por um grupo de especialistas, configura também como uma etapa de validação, pois assume-se que a análise de juízes ou especialistas podem fornecer evidências de validação. Fulcher (2003), inclusive, incita os elaboradores de testes a registrarem, sempre quando possível, as tomadas de decisão sobre o processo de elaboração e pilotagem dos testes para fins de justificar e engrossar o argumento da validade do exame. 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 a habilidade ou proficiência avaliada, ou seja, os especialistas podem julgar a representatividade de cada uma das tarefas ou itens avaliados. O documento da AERA, assim como Messick (1987), também aponta para a possibilidade de investigar a nota para avaliar aspectos da relação entre conteúdo e habilidade, proficiência ou conhecimento avaliados. Os linguistas aplicados, especialistas em avaliação de línguas, tendem a corroborar o *Standards* (2014) e Fulcher (2003), ao afirmarem que a representatividade das tarefas de um teste é um fator de avaliação da validade do exame, ainda que limitado.

Hughes em 1989 afirma que "o teste é dito válido do ponto de vista do conteúdo apenas se for constituído por amostras representativas da habilidade linguística, estruturas, etc. com as quais o exame está relacionado (…) E o que for considerado como estrutura relevante vai depender, claro, do propósito do exame" (HUGHES, 1989, p.22 tradução da autora) [5] Na perspectiva de Brown e Abeywickrama (2004), a validade de conteúdo tem a ver com os objetivos do curso e com o fato do teste avaliar as habilidades de forma direta, definem os autores: "validade de conteúdo está relacionada ao quanto o teste avalia os objetivos reais do curso e garante uma avaliação direta" (BROWN; ABEYWICKRAMA, 2004, p.32 tradução da autora)[6] Interessante ressaltar que os autores relacionam validade de conteúdo tanto ao programa de algum curso e como também à forma como a habilidade linguística é avaliada, que pode ser direta ou indiretamente. Testes diretos são, por exemplo, uma conversa para avaliar a habilidade oral. Em um teste indireto, poderia ser requisitado que o examinando escrevesse um diálogo, ou seja, escrever uma

[5] A test is said to have content validity IF its content constitutes a representative sample of the language skills, structures, etc. with it must be concerned. (…) Just what are the relevant structure will depend, of course, upon the purpose of the test. "(HUGHES, 1989, p.22).

[6] The extent to which the tests assess real course objectives and by ensuring that they test directly (BROWN; ABEYWICKRAMA, 2004, p.32).

interação que normalmente acontece na modalidade oral e não escrita, como pede este exemplo hipotético. Ao pedir para escrever diálogos, indiretamente, ou seja, por meio da escrita de diálogos, o instrumento avalia a habilidade oral. Por outro lado, uma entrevista de proficiência oral é direta, ou seja, o modo oral está coerente com a habilidade avaliada, por isso muitos autores equivocadamente já a classificam como válidas. No entanto, outros aspectos da validade de conteúdo devem ser levados em conta. A coerência entre a habilidade avaliada, construto, e seu modo de coleta de informações é um aspecto relevante, mas outros também devem ser levados em consideração.

Autenticidade ou o aspecto direto ou indireto das questões do teste estão relacionados com o critério ou, no caso das avaliações de línguas, com o quanto o teste simula de forma realística as situações de uso da língua. Messick (1996) discute a relação entre autenticidade e validade em contextos de elaboração e validação de testes de línguas. O autor questiona o fato de alguns linguistas aplicados, como Hughes (1989) ou Brown e Abeywickrama (2004), afirmarem que o fato do teste ser direto e autêntico garantiria a validade do teste. O processo de validação do teste é extremamente complexo e envolve não só avaliação do desenho do teste, mas uma imensa quantidade de evidências tanto teóricas quanto empíricas que fundamenta a coerência entre o significado da nota e suas consequências individuais e sociais. Messick (1996) é cirúrgico ao afirmar que o termo 'avaliações diretas' é um péssimo termo para qualificar as avaliações porque promete o impossível, uma vez que as medidas são sempre indiretas. O psicometrista explica que "as medidas sempre envolvem, ainda que tacitamente, processo de julgamento, comparação e inferência."[7] (p.244). A questão para nós, linguistas e professores, é saber que ao propor uma avaliação que simula a interação próxima do uso da língua na vida real, seja escrita ou oral, não significa que a nota gerada por meio do instrumento esteja avaliando com eficiência o que o examinando sabe fazer na língua. Dizendo de outra forma, fazer uma avaliação autêntica e direta, com materiais de insumo autênticos e questões ou tarefas próximas das que os examinandos desempenham (ou precisam desempenhar na língua) são interessantes do ponto de vista do construto teórico baseado em abordagens comunicativas, por exemplo, mas o fato dos itens simularem bem situações de uso da língua não garante que o instrumento está avaliando ou medindo o que precisa ser avaliado. É preciso atentar-nos a outros aspectos da avaliação como a forma como estamos corrigindo as respostas, a forma como estamos organizando as faixas de proficiência, dentre outros aspectos.

Bachman e Palmer (1996) elencam características de testes de línguas que deveriam nortear estudos e elaboração de testes úteis e de qualidade. Na lista, os autores colocam a validade de construto e a autenticidade. A forma como se define a validade de construto é próxima a de Messick (1987), para os

[7] Measurements always involves, even if only tacily, intervening processes of judgement, comparison, or inference. (MESSICK, 1996, p. 244)

autores "validade de construto diz respeito à adequação e significância das interpretações que se faz a partir das notas"[8] (BACHMAN, PALMER, 1996, p.21) Na definição fica claro que os autores entendem a validade a partir das notas, em outro momento do texto, eles tornam a afirmar que para demonstrar e justificar a validade das interpretações geradas é preciso levar em conta a nota e não apenas afirmar que o teste é válido, sem considerar a nota, ou seja, os autores, em um primeiro momento, parecem desconsiderar que o desenho da prova por si só poderia ser uma evidência de validade. Os autores entendem o conceito de validade assim como Messick (1987), uma vez que partem de uma perspectiva das análises psicométricas ou das teorias das medidas para validar ou não testes e interpretações geradas por eles. Embora tenham deixado de lado as questões de representatividade da tarefa, por meio do juízo de especialistas na definição da validade, os autores criam uma outra categoria para enquadrar a seleção de conteúdo e ou formato do instrumento como um indicador de qualidade dos testes: a autenticidade. A autenticidade, para os autores, está relacionada ao quanto as tarefas ou as atividades correspondem ao uso real da língua nos domínios especificados pelos exames. Os autores defendem a importância de os testes serem autênticos, porque a autenticidade dos testes interfere na interpretação que fazemos a partir das notas. Importante ressaltar que, embora os autores se alinhem à Messick (1987) ao fechar o conceito de validade, excluindo as evidências de representatividade da tarefa para fundamentar uma validação, os autores parecem retomar o mesmo conceito de validade de conteúdo, substituindo-o pelo termo 'autenticidade', uma vez que para eles "a autenticidade oferece meios para investigação da medida em que as interpretações das notas podem ser generalizadas para além do desempenho no teste (..) Este é o link entre autenticidade e validade de construto, porque investigar o quanto a nota pode ser generalizada é uma parte importante para a validade de construto."[9] (BACHMAN, PALMER, 1996, p.24). No fim das contas, para todos os autores resenhados até então a escolha de conteúdo, tarefas e atividades é algo a ser levado em conta pelos elaboradores do exame, porque interfere na validade dos instrumentos.

McNamara (2000) adverte que avaliar a validade do conteúdo ou da tarefa é algo complexo, porque está relacionado ao uso futuro da língua que o teste tenta prever. Especialmente em testes de larga escala, este tipo de validação se torna ainda mais complicado, uma vez que os testes são usados para outros fins. É comum, por exemplo, que um teste de conhecimento de gramática seja usado para certificar proficiência linguística em um contexto de ingresso em pós-graduação, em que, via de regra, a habilidade de leitura

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

[9] Authenticity thus provides a means for investigations the extent to which score interpretations generalize beyond performance (...) This links authenticity to construct validity, since investigating the generalizability of score interpretations is an important part of construct validation.(BACHMAN, PALMER, 1996, p.21)

em língua estrangeira será a mais exigida. Outra problematização sobre a validade de conteúdo levantada por McNamara (2000) diz respeito ao desenho do teste, até que ponto o formato ou comando das questões do teste faz com que os examinandos acertem ou errem?

Sobre a discussão de validade e sua implicação para o desenho das tarefas orais, retomo a ideia defendida por Fulcher (2003) para quem a validade não é uma questão de tudo ou nada, pois trata-se de ir engrossando o argumento a favor do teste ao coletar diversas evidências, inclusive sobre conteúdo e formato do teste.

A seguir, apresento uma breve revisão bibliográfica sobre as discussões acerca da autenticidade, validade de conteúdo e validade de face no contexto das entrevistas de proficiência oral.

4. Avaliação da proficiência oral: uma discussão sobre sua validade

Bygate (2010) ao resenhar diversos estudos sobre ensino e avaliação da língua oral, conclui que, tanto os procedimentos para ensino quanto avaliação da oralidade deveriam ser holísticos, de forma que abarquem tanto os conhecimentos declarativos como os de natureza linguística, como a pronúncia e a prosódia, por exemplo, envolvidos na interação. 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 interagirem 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"[10] (HUGHES, 1989, p.101). 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.

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.

A metodologia da entrevista de proficiência oral (EPO) desenvolvida pelo *Foreign 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

[10] The tasks should elicit behavior which truly represents the candidates' ability, and which can be scored validity and reliably." (HUGHES, 1989, p.101)

sido objeto de estudo de muitas pesquisas que forneceram insumos para compreensão dos construtos de avaliação da proficiência oral, entre outros aspectos. Brown (2005) sugere que a partir das críticas ao teste oral do FSI, especialmente sobre o fato do desenho das tarefas não contemplarem a competência interacional, 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.

Brown (2005), ao defender a entrevista de proficiência oral como forma de avaliação oral, argumenta que a validade do modo de avaliação 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"[11] (BROWN, 2005, p.1). 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. O argumento de Brown (2005) remete à confiabilidade do teste no que diz respeito à estabilidade da nota. Neste sentido, é desejável que o teste ofereça mais ou menos as mesmas condições para todos os examinandos demonstrarem o que sabem, por isso a importância da situação de interação face a face ter um roteiro, ainda que haja espaço para alguma flexibilidade no gerenciamento das perguntas, dos turnos, etc.

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. Cabe ressaltar que o entrevistador apenas conduz a entrevista que é gravada, a mesma entrevista é enviada para dois avaliadores atribuirem notas independentes.

Na entrevista de proficiência oral do Celpe-Bras, poder-se-ia dizer que há duas tarefas, sendo a

[11] While performance-based speaking tasks may take a number of forms involving a range of different task-types (such as roleplay, picture description or information gap) the claim to validity for oral interviews derives from the conversational nature of interation. (BROWN, 2005, p.1)

primeira, o quebra-gelo; e a segunda, uma interação partir do tema de três Elementos Provocadores[12]. Na entrevista de proficiência oral do CELU, há três tarefas: o quebra-gelo e uma interação a partir do tema de uma lâmina e um *juego de roles*. Os Elementos Provocadores ou lâminas são elaborados a partir de recortes de reportagens para subsidiar a interação. Na prova do CELU, um *juego de roles* envolve, geralmente, dois personagens que estão em desacordo sobre tomar ou não uma atitude, por isso é esperado que o examinador-interlocutor e examinando se coloquem em papéis diferentes na interação para cumprir a tarefa.

Quadro 03: Desenho da parte oral no Celpe-Bras e no Celu

	Celpe-Bras		Celu		
tarefa	quebra-gelo	Interação a partir de 3 EPs	quebra-gelo	Interação a partir de 1 lâmina	*juego de roles* (a partir do tema da lâmina)
tempo	5 minutos	5 minutos cada EP (15 min. total)	3 minutos	5 a 7 minutos	5 minutos

Fonte: elaboração da autora

As discussões sobre o formato de entrevista de proficiência oral em língua inglesa dizem respeito também aos exames de Celpe-Bras e Celu, uma vez que argumentamos acima que os três instrumentos compartilham desenhos similares, ou seja, todos propõe uma situação de entrevista de proficiência oral, porém com tarefas distintas.

Há pesquisadores 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 do exame IELTS, concluiu que, devido à assimetria entre os interagentes – interlocutor e examinando –, a conversa mais se assemelharia a uma entrevista sociolinguística. A autora estudou esta situação de avaliação oral 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). Na entrevista de proficiência oral é comum que a primeira parte da prova se assemelhe a um quebra-gelo em que o examinando já está sendo avaliado. Na segunda parte da prova, as mudanças de tópicos ou tarefas são frequentes. Segundo Johnson (2001), o que faz com que a entrevista tenha o foco adequado ao tópico é o gerenciamento de suas perguntas e o controle do interlocutor

[12] Para um debate mais completo sobre as tarefas dos exames orais do Celpe-Bras e Celu, consulte Ferreira (2019).

sobre a interação.

Brown (2005) pondera que, embora haja argumentos como os apresentados por Johnson (2001), que questionam a validade da entrevista ao afirmar 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 por Johnson (2001) nos remete aos limites da autenticidade da situação da avaliação e do desenho de tarefa. Limito-me, neste texto, a discutir a autenticidade da situação de entrevista de proficiência oral.

A autenticidade se refere à simulação mais próxima possível dos usos da língua. 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 tratar de contextos de avaliação. Todo o debate sobre a natureza das entrevistas de proficiência oral está relacionado ao argumento da validade. Julgar a validade de uma situação de prova, analisando a situação da interação, está associado à validade de face ou conteúdo, discutidos acima.

A falácia do argumento relacionado à validade de conteúdo das entrevistas de proficiência oral 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 (1989) e Brown e Abeywickrama (2004), ao passo que o exame indireto não seria considerado válido. Fulcher (2003) refuta os argumentos de Hughes (1989) e Brown e Abeywickrama (2004), 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 construtos de avaliação muito frequentemente não fornecem definições operacionais do construto. Fulcher (2003) problematiza a questão do uso da validade de conteúdo 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 válida confunde a compreensão e o debate sobre a manifestação do comportamento que se quer medir e que refletem os construtos teóricos do exame. A interação verbal é um comportamento a ser medido por meio de testes que são indicadores indiretos da proficiência a ser avaliada, por isso é preciso considerar o desenho do instrumento como um todo. Essa ideia de Fulcher (2003) remete ao texto de Messick de 1996 em que o psicometrista chama atenção para o fato de que nenhum teste mede diretamente a habilidade ou a proficiência, porque as medidas são sempre uma representação indireta do que se quer avaliar. 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, ou seja, da proficiência oral, ao passo que minimiza o efeito do desenho do instrumento nas notas. O efeito do desenho do instrumento ocorre quando a nota está sendo influenciada pela situação como um todo ou por algum de seus aspectos. Por exemplo, no Celpe-Bras, o Elemento Provocador para fomentar a interação face a face pode impactar no desempenho do examinando e na sua interpretação pelo avaliador. É desejável, para fins de avaliação de larga escala, que o efeito do desenho do instrumento seja minimizado. Se o Elemento Provocador for inadequado, por exemplo, ao exigir conhecimentos prévios que a maioria dos examinandos provavelmente não têm, o examinando sairá prejudicado, pois não conseguirá conversar sobre o assunto da entrevista ou terá de pedir ajuda para o avaliador-interlocutor para conseguir entender o tema da conversa, ocasionando talvez variação na sua nota.

Sobre a natureza da tarefa, o que se põe à prova, segundo Messick (1987), 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 determinado exame poderia ser generalizado para outros contextos em que se demanda a proficiência oral. 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 uma situação de uso da língua por meio da qual inferências sobre a capacidade do examinando de usar essa língua em contextos fora do teste serão feitas. Nesse sentido, a validade de conteúdo, que está relacionado ao desenho da prova, poderia ser mais bem compreendida se fossem analisadas não só a situação da entrevista oral a partir da análise das interações, como fez Johnson (2001) e Fulcher (2003), mas também a partir do estudo da nota que também operacionalizam os construtos de avaliação. Cabe ressaltar ainda que 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). 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; FERREIRA, 2018; NEVES, 2018). Para além do debate da validade de conteúdo das situações das provas orais, é preciso também investir na análise de outros fatores que interagem com o desenho dos instrumentos.

Considerações Finais

A avaliação do desenvolvimento da oralidade em línguas adicionais é algo relativamente recente na agenda de pesquisa dos linguistas aplicados. Embora a oralidade seja central em alguns contextos de ensino e aprendizagem de línguas, o debate sobre como avaliar a língua oral está distante de um consenso. Várias áreas contribuem para a compreensão dos processos de avaliação. No caso da avaliação de línguas, não só as teorias sobre ensino e aprendizagem de línguas estão no centro do debate, mas também conceitos da área de avaliação educacional que vem da psicometria como a validade e a confiabilidade.

Neste trabalho, explorei as definições de validade, relacionando com as ideias sobre validade apresentadas pelos linguistas aplicados. O aspecto do conteúdo das avaliações ou autenticidade da tarefa é um tema relevante e intensamente discutido pelos pesquisadores da área de línguas, porém há algumas questões que os especialistas de línguas divergem. No entanto, é possível afirmar que há um consenso, a partir das obras de referência em avaliação em línguas resenhadas neste texto, de que o desenho do instrumento está relacionado a sua validade. As avaliações orais, por serem as menos estudadas, quando comparadas às provas escritas, foram o foco de debate deste texto e busquei exemplificar o debate a partir dos dois maiores exames de larga escala da América do Sul: o Celpe-Bras e o Celu. Por meio da revisão bibliográfica de vários autores que escrevem sobre avaliações de línguas foi possível retomar os principais argumentos e debates sobre a validade de entrevistas orais de proficiência.

Em geral, os linguistas aplicados concordam que as situações de avaliações orais devem ser semelhantes às situações reais de uso da língua oral. Os linguistas aplicados concordam ao dizer que as avaliações deveriam ser diretas, comunicativas e o mais autêntica possível, porém o debate está em aberto sobre como estas avaliações deveriam ser. Sobre os argumentos levantados que dizem respeito à validade da situação de entrevistas orais, pudemos perceber que há uma lacuna de pesquisa no que diz respeito à necessidade de levantar evidências empíricas sobre como o desenho da prova interage com outros aspectos do instrumento.

Referências

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

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

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

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

BROWN, H. Douglas; ABEYWICKRAMA, Priyanvada. 2010 *Language Assessment:* principles and

classroom practice. Nova Iorque: Longman Pearson.
BROWN, Annie. 2005 *Interviewer variability in oral proficiency interviews.* Frankfurt: PeterLang.
BYGATE, Martin. 2011 Teaching and testing speaking. In: LONG, Michael H.; DOUGHTY, Catherine J. *The handbook of language teaching.* Chichester: Wiley-Blackwell. p.411-440
CHAMBERS, Francine. 1997. What do we mean by fluency? *System,* v.25, n.4, p.535-544.
CONSOLO, Douglas A. 2005 *Postura sobre avaliação de proficiência oral do professor de língua estrangeira: implicações para o cenário brasileiro.* IN: FREIRE, M.M.; VIERA-ABRAHÃO, M.H.; BARCELOS, A.M.F. Linguística Aplicada e contemporaneidade. São Paulo: ALAB, Pontes Editores.
DELL'ISOLA, Regina L. Péret; SCARAMUCCI, Matilde V.R.; SCHLATTER, Margarete; JÚDICE, Norimar. 2003 A avaliação de proficiência em português língua estrangeira: o exame CELPE-Bras. *Revista Brasileira de Linguística Aplicada***,** Belo Horizonte, v.3, n.1, p.153-184.
ECKES, Thomas. 2015 *Introduction to many-facet rasch measurement:* analyzing and evaluating rater-mediated assessments. Frankfurt: PeterLang.
ELSE, Consorcio Español como Lengua Segunda y Extranjera. 2019 Dispõe informações sobre: exame CELU, *Certificado de Español Lengua y Uso.* Disponível em: https://www.celu.edu.ar/es/content/actividades-del-examen- Acesso em: 11 jun. 2019.
FERREIRA, Laura Márcia Luiza. Avaliação da proficiência oral: uma análise fatorial e de discriminação de itens do exame Celpe-Bras. Tese de Doutorado. 244f. Belo Horizonte: CEFET-MG, 2018.
_____. Validade de conteúdo e especificação das propostas de interação oral dos exames Celpe-Bras e CELU. IN: BARROSO, Silvina; DANDREA, Fabio. X Coloquio CELU: enseñanza y evaluación ELSE: aportes para una politica orientada a la integración regional. Río Cuarto: UniRío Editora (Universidad Nacional de Río Cuarto), 2019. p.23-37
FULCHER, Glenn. *Testing second language speaking.* Londres: Routledge, 2003.
FULCHER, Gleenn; DAVIDSON, Fred. 2007 *Language testing and assessment:* an advanced resource book. Routledge: Nova Iorque. p.91-114
GRABE, William; STOLLER, Fredricka L. 2002 *Teaching and researching reading.* Harlow, England: Pearson Education, Longman.
HYMES, D. 1972 *On Communicative Competence***.** In PRIDE, J. B. e HOLMES, J. Sociolinguistics. England: Penguin Books. 381 p.
HUGHES, Arthur. 1989 *Testing for language teachers.* Cambridge: Cambridge University Press.
JOHNSON, Marysia. 2001 *The art of non-conversation:* a reexamination of the validity of the oral proficiency interview. Yale Haven & London: Yale University Press.
LUCKESI, Cipriano Carlos. 2011 *Avaliação da aprendizagem escolar:* estudos e proposições. 22ed. São Paulo: Editora Cortez.
NATION, Paul. 1989 Improving speaking fluency. *System,* v.17, n.3, p.377-384.
McNAMARA, Tim. 2000 *Language Testing.* Oxford: Oxford University Press.
McNAMARA, Tim. 2004 Language Testing. In: DAVIES, Alan; ELDER, Catherine. *The handbook of applied linguistics.*
MCNAMARA,Tim. 2008 *Language assessment as social practice.* IN: IV COLOQUIO CELU, Universidad de San Martín, Buenos Aires. Anais eletrônicos...Buenos aires: Universidad de San Martín, 2008. Disponível em: https://www.celu.edu.ar/sites/www.celu.edu.ar/files/images/stories/pdf/McNamara_conferencia_.pdf Acesso em: 01 de mai. 2019
MESSICK, Samuel. 1987 *Validity.* Nova Jersey: Educational Testing Service Princeton.
MESSICK, Samuel. 1996 Validity and Washback in Language Testing. *Language Testing,* v. 13, n.3, p.241–256., nov., p.241–256. Doi:10.1177/026553229601300302
NEVES, Liliane Oliveira. 2018 *Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras: um estudo longitudinal.* Tese de Doutorado. Belo Horizonte: CEFET-MG.

SCARAMUCCI, Matilde. V. R.. 2001 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. p. 77-90

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

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

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