Bibliografia comentada sobre validade e confiabilidade de exames de proficiência em línguas

Melissa Santos Fortes
Simone Paula Kunrath

CHAPELLE, Carol A. Validity in language assessment. *Annual Review of Applied Linguistics*, v. 19, p. 254-272, Jan. 1999. Disponível em: <https://www.cambridge.org/core/journals/annual-review-of-applied-linguistics/article/validity-in-language-assessment/1B08C26848F89D0D2E148129AA788918>.

Com base na perspectiva histórica dos estudos sobre a validade da avaliação de língua adicional, o artigo mostra como os testes de desempenho no uso da linguagem desafiaram a validade e a confiabilidade vigentes até a década de 1970. A grande contribuição ao tema ocorreu no final da década de 1980, quando a área de avaliação em educação e em psicologia publicou um conjunto de parâmetros que substituíam as três validades dos anos 1970 por uma visão unificada. Nessa nova visão, a validade de construto passou a ocupar o lugar central, e a validade de conteúdo, assim como a validade correlacionada, foi apresentada como método para a investigação da validade de construto. Outro marco foi a publicação do artigo de Samuel Messick na terceira edição do *Handbook of Educational Measurement*, em 1989, no qual propõe um modelo de validade segundo a concepção de que não são os testes que são válidos, mas as inferências que fazemos e as decisões que tomamos com base no resultado dos testes, com um objetivo específico, é que são válidas ou não. A validade de construto assume o lugar central, e a confiabilidade é entendida como uma das evidências para conferir validade à avaliação.

DIAS, Ana Luiza Krüger; PINTO, Joana Plaza. Ideologias linguísticas e regimes de testes de língua para migrantes no Brasil. *Revista Brasileira de Linguística Aplicada* [*online*], v. 17, n. 1, p. 61-81, 2017. Disponível em: <http://www.scielo.br/pdf/rbla/v17n1/1984-6398-rbla-17-01-00061.pdf>.

Três testes de língua no contexto de migração transnacional para o Brasil foram analisados com o objetivo de verificar sua relação com ideologias linguísticas hegemônicas na construção de sistemas de diferenciação corporal: 1) Certificação de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras); 2) avaliação de língua portuguesa para o Programa Mais Médicos; e 3) exigência de conhecimento de língua portuguesa nos processos de naturalização. A análise mostrou discrepâncias entre o estado da arte sobre testes linguísticos em contexto migratório e a realidade brasileira, pois as articulações entre seu aspecto linguístico e seu aspecto de barreira são contraditórias. Os documentos sobre testes no Brasil indicam uma autonomia no gerenciamento da língua portuguesa e a comoditização do ensino do português para estrangeiros como estratégia de mercado e vitrine da cultura brasileira globalizada, numa construção ideológica de correspondência estática e naturalizada entre língua oficial e nação, produzindo, assim, hierarquizações entre identidades migrantes "desejáveis" e "indesejáveis".

FERREIRA, Laura Márcia Luiza. *Avaliação da proficiência oral: análise fatorial e de discriminação dos itens do exame Celpe-Bras*. 2018. Tese (doutorado em Estudos de Linguagens) – Centro Federal de Educação Tecnológica de Minas Gerais (Cefet-MG), 2018. Disponível em: <https://sig.cefetmg.br/sigaa/>.

As duas escalas de avaliação da prova oral do exame de Certificação de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) foram analisadas para coletar evidências da validade interna. Quanto à dimensionalidade dos itens, é apresentada uma análise fatorial exploratória. Quanto ao ajuste e à quantidade de informação de cada item, é apresentada uma análise de discriminação de itens, por meio do modelo Rasch básico na extensão Partial Credit Model. A escala do avaliador-intelocutor contém um item e a escala do avaliador-observador, seis itens: compreensão, competência interacional, fluência, adequação lexical, adequação gramatical e pronúncia. O conjunto de dados é composto por notas atribuídas para mil participantes que se submeteram ao exame na primeira edição de 2016. O resultado da análise fatorial sugere que a nota da prova oral seja uma medida unidimensional. Ao considerar intervalos entre os valores de *threshold* dos sete itens das escalas, conclui-se que uma mesma faixa de nota em cada um dos itens pode não discriminar da mesma forma o mesmo perfil de examinandos. Com base nas análises, propõe-se a mudança de peso dos itens na composição da nota oral, especialmente quanto ao item compreensão, bem como a necessidade de investimento na revisão dos descritores das escalas, da tarefa e da situação de entrevista de proficiência oral.

FORTES, Melissa Santos. *Uma compreensão etnometodológica do trabalho de fazer ser membro na fala-em-interação de entrevista de proficiência oral em português como língua adicional.* 2010. 329 p. Tese (Doutorado em Estudos da Linguagem) – Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010. Disponível em: <https://lume.ufrgs.br/handle/10183/26736>.

A compreensão da proficiência oral em língua adicional é descrita segundo a perspectiva dos participantes da entrevista do exame para o Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras), propondo um entendimento êmico e uma reflexão sobre parâmetros de avaliação dessa proficiência, válidos e confiáveis para o construto *uso da linguagem para ação no mundo*. A metodologia consistiu na análise das entrevistas do exame, com um *corpus* de 10 horas de gravação em áudio em um posto aplicador na Região Sul do Brasil. A descrição realizada teve como implicações a redefinição do conceito de proficiência oral, a reconfiguração do conceito de confiabilidade e a elaboração de novos parâmetros na área de avaliação de língua adicional de modo que a avaliação seja válida e confiável para o construto.

FULCHER, Glenn; DAVIDSON, Fred (Ed.). *The Routledge handbook of language testing*. New York: Routldege, 2013. Disponível em: <http://english2.zanjansadra.ir/ufiles/47685381518426262.pdf>.

Introdução às dimensões teórico-práticas da avaliação da linguagem conforme especialistas internacionais, em nove seções temáticas: 1) conceito de validade; 2) avaliação de rendimento em contextos de sala de aula e o efeito retroativo dela; 3) dimensão social dos usos das avaliações da linguagem, como a avaliação para diagnóstico de transtornos de comunicação ou a avaliação para cidadania, imigração e asilo político; 4) avaliações de desempenho; 5) escrita de itens de avaliação; 6) elaboração de itens-piloto e testes; 7) Teoria de Resposta ao Item, conceito de confiabilidade e questões relacionadas à generalização de resultados; 8) aplicação das avaliações e formação dos avaliadores, com textos sobre a confiabilidade entre avaliadores e o uso da tecnologia em avaliações da linguagem; 9) ética e as políticas linguísticas em avaliações de larga escala. É um livro de leitura obrigatória para quem inicia os seus estudos na área de avaliação da linguagem.

HE, Agnes Weiyun. Answering questions in language proficiency interviews: a case study. In: YOUNG, Richard; HE, Agnes Weiyun (Eds.). *Talking and testing: discourse approaches to the assessment of oral proficiency*. Amsterdam: Benjamins, 1998. p. 101-115. Sumário do livro disponível em: <https://benjamins.com/catalog/sibil.14>.

A entrevista é uma realização coordenada e conjunta na fala-em-interação. Com base na análise da conversa, a autora afirma que as entrevistas de proficiência

oral são produções colaborativas que, por essa característica constitutiva, devem ser estudadas como resultado de um trabalho interacional entre os participantes mediante o uso da linguagem, porque "a competência gramatical e a competência discursiva são inseparáveis uma da outra" (p. 103). Entretanto, na análise dos dados de entrevista de avaliação de proficiência oral de falantes de inglês como língua adicional candidatos à vaga de professor assistente em uma universidade norte-americana, ela considera os recursos linguísticos, a despeito das ações realizadas pelos participantes, produzindo uma descrição da perspectiva da analista e não dos entendimentos entre os participantes na construção coordenada e conjunta de suas ações nas entrevistas. O livro é importante por considerar as diversas perspectivas discursivas na pesquisa sobre a avaliação de proficiência oral em língua adicional. No entanto, é necessário que o leitor esteja atento a eventuais interpretações particulares das teorias nas quais se fundamentam as pesquisas nele apresentadas.

HUGHES, Arthur. *Testing for language teachers*. Cambridge Cambridge University Press, 1989. Disponível em: <https://epdf.pub/testing-for-language-teachers.html>.

O livro tem como interlocutores professores de línguas adicionais preocupados em qualificar suas práticas avaliativas de ensino-aprendizagem. Inicialmente, apresenta os conceitos de efeito retroativo e de confiabilidade para mostrar os impactos positivos da avaliação no ensino. Em seguida, o capítulo 3 comenta os tipos de avaliação, quando e como cada tipo deve ser utilizado em relação aos objetivos de ensino e de aprendizagem. Os capítulos 4, 5 e 6 apresentam os conceitos de validade, de confiabilidade e de efeito retroativo. Os capítulos 7 e 8 abordam procedimentos para a elaboração de uma avaliação (com análise de exemplos e com proposta de atividades para disscussão) e abordam também técnicas de avaliação de mais de uma habilidade, com ênfase nos testes de múltipla escolha e nos testes Cloze. Os capítulos 9 a 13 são dedicados à avaliação de habilidades, de gramática e de vocabulário, com o objetivo de o leitor refletir sobre os instrumentos de avaliação, e se eles são válidos e confiáveis para o que se deseja avaliar. Por último, o capítulo 14 apresenta um conjunto de procedimentos a serem adotados para uma aplicação confiável de avaliações em larga escala. Pelo seu grau de detalhamento e abrangência, o livro é bastante adequado para a formação de professores de línguas adicionais cujo foco sejam os princípios e as práticas de avaliação da linguagem.

IN'NAMI, Yo; KOIZUMI, Rie. Task and rater effects in L2 speaking and writing: a synthesis of generalizability studies. *Language Testing*, Thousand Oaks, v. 33, n. 3, p. 341-366, 2016. Disponível em: <https://journals.sagepub.com/doi/abs/10.1177/0265532215587390>

Trata-se de uma síntese dos estudos de generabilidade sobre os efeitos relativos das tarefas, dos avaliadores e de suas interações na fala e na escrita em

segunda língua (L2). Os autores abordam os estudos de Deville e Chalhoub-Deville (2006), Schoonen (2012) e Xi e Mollaun (2006) conforme a teoria da generabilidade, considerando as características contextuais relacionadas às interações entre o examinando e as tarefas (*to person-by-task interactions*) por meio de dois caminhos. Primeiro, os autores sintetizam quantitativamente os estudos de generabilidade para determinar a porcentagem de variação no desempenho da fala e da escrita em L2, que foi contabilizado nas tarefas, nos avaliadores e em suas interações. Depois, examinam as relações entre as interações dos examinandos e as tarefas com as variáveis dos avaliadores. Foram utilizados 28 conjuntos de dados de 21 estudos para fala em L2 e 22 conjuntos de dados de 17 estudos para escrita em L2. Os resultados indicam que os contextos, os métodos e os critérios de pontuação escolhidos podem levar a um desempenho diferente do proposto pelas tarefas. Nesse sentido, os autores destacam a importância de definirmos bem a relação entre o construto e os objetivos de testagem.

McNAMARA, Tim; ROEVER, Carsten. *Language testing: the social dimension*. Maiden, MA: Blackwell Publishing, 2006. Sumário do livro disponível em: <https://www.wiley.com/en-us/Language+Testing%3A+The+Social+Dimension-p-9781405155434>.

As avaliações de uso da língua de orientação psicométrica obscurecem o papel e o efeito dos testes de língua adicional na sociedade pelo seu caráter cognitivo – e, portanto, com foco no processamento mental do indivíduo. Entre os impactos desses testes na sociedade, destaca-se seu uso como instrumento de permissão ou de impedimento a empregos, educação ou cidadania, por exemplo, por falantes de línguas adicionais. Com base no entendimento dos testes como práticas sociais, os autores afirmam que os testes e seus usos são fruto de processos políticos e de decisões ideológicas sobre: 1) quem deve ser autorizado a imigrar; 2) qual deve ser o parâmetro de proficiência exigido para a concessão da cidadania ou para exercer uma profissão em outro país; e 3) quais são os valores políticos e sociais que estão em jogo na implementação de políticas por meio dos testes.

MENDOZA RAMOS, Arturo. La validez de los exámenes de alto impacto: un enfoque desde la lógica argumentativa. *Perfiles Educativos*, Mexico, v. 37, n. 149, p. 169-186, 2015. Disponível em: <http://www.scielo.org.mx/pdf/peredu/v37n149/v37n149a10.pdf>.

A validade na avaliação educacional de alto impacto representa um desafio para as instituições que se dedicam a projetar e validar exames. O autor destaca a importância da validade a partir da segunda metade do século 20, as dificuldades latentes na hora de validar os exames no final do século passado e uma nova abordagem de validade, que consiste em todo um processo de inferências baseadas na evidência. Esse modelo está sendo seguido em várias áreas educacionais

internacionalmente; no entanto, ganhou relevância especial na avaliação de segundas línguas. A aplicabilidade desse enfoque é apresentada por meio da descrição das seis inferências adaptadas ao exame de espanhol destinado a candidatos não hispano-falantes que pretendem cursar graduação ou pós-graduação na Universidad Nacional Autónoma de México (Unam). As inferências/os argumentos são: 1) descrição de domínio; 2) avaliação; 3) generalização; 4) explicação; 5) extrapolação; 6) utilização.

NEVES, Liliane de Oliveira. *Confiabilidade e comportamento avaliativo na prova oral do exame Celpe-Bras: um estudo longitudinal*. 2018. Tese (doutorado em Estudos de Linguagens) – Centro Federal de Educação Tecnológica de Minas Gerais (Cefet-MG), 2018. Disponível em: <https://sig.cefetmg.br/sigaa/>.

A prova oral do Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras) é uma interação entre o examinando, o avaliador-interlocutor e o avaliador-observador. A avaliação é feita em primeira instância – imediatamente após a aplicação da prova – e, havendo discrepância significativa entre as notas atribuídas pelos dois avaliadores, a interação é reavaliada em segunda e/ou terceira instância. Foi empregada uma metodologia quantitativa, que levou em conta dados de sete edições consecutivas do Celpe-Bras, envolvendo notas de 29.831 examinandos. A análise dos níveis de proficiência atribuídos aos examinandos e de informações estatísticas das notas, como medidas de tendência central e de dispersão, revelou variabilidade de comportamento avaliativo. Constatou-se que: 1) a escala de avaliação é unidimensional, ou seja, avalia um único construto, na primeira instância; na segunda instância, ela é bidimensional; 2) as sete edições apresentam valores altos do coeficiente de confiabilidade na avaliação em primeira instância, o que significa que os itens da escala possuem elevada consistência interna; já na avaliação realizada em segunda instância, a confiabilidade revela-se moderada; e 3) as sete edições, na avaliação em primeira instância, apresentam valores satisfatórios de concordância entre os avaliadores, ainda que baixos; na avaliação em segunda instância, apresentam valor pobre de concordância. Isso significa que a segunda instância, instituída para dirimir os problemas avaliativos que surgem na primeira, é marcada por comportamento diferenciado dos sujeitos avaliadores, diminuindo, portanto, a confiabilidade dos resultados.

PILEGGI, Maria Gabriela S. Integração de habilidades: perspectiva histórico-teórica e operacionalização no exame Celpe-Bras. *Estudos Linguísticos*, São Paulo, v. 46, n. 2, p. 577-592, 2017. Disponível em: <https://revistas.gel.org.br/estudos-linguisticos/article/view/1677>.

Na década de 1960, a avaliação em línguas se desenvolveu como área de estudos. Para a teoria estruturalista, a linguagem era um sistema de hábitos de comunicação que envolviam questões de forma, sentido e níveis de estrutura (oração,

frase, palavra, morfema e fonema). Os testes avaliavam os componentes linguísticos isoladamente e sem contexto. No final da década de 1970, com a ideia de competência comunicativa, a ênfase passou a ser o uso da língua, os testes deviam ser práticos e medir habilidades linguísticas. Na década de 1990, surgem os testes de desempenho e desenvolve-se o conceito de integração de habilidades como resultado da distinção entre tarefas integrativas e integradas, segundo exigissem mais de uma ou apenas uma habilidade. A integração se opõe à avaliação de aspectos isolados, e a proficiência é medida como um todo unitário. As vantagens da abordagem integrada são o aumento da validade do teste e a racionalização dos procedimentos, pois é possível avaliar mais de uma habilidade ao mesmo tempo. Essa abordagem também provocou controvérsias sobre a validade, a complexidade das tarefas, a contaminação na avaliação e o estabelecimento de níveis de desempenho. Um exemplo de operacionalização da integração de habilidades é apresentado por meio da análise de duas tarefas das edições 1 e 2 de 2014 do exame para a obtenção do Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). Conclui que o construto do exame Celpe-Bras busca avaliar a proficiência no uso da língua em situações realistas de comunicação e que as tarefas analisadas são coerentes com esse construto e integram duas habilidades: leitura e produção escrita.

SCARAMUCCI, Matilde Virgínia Riccardi. Proficiência em LE: considerações terminológicas e conceituais. *Trabalhos em Linguística Aplicada*, Campinas, v. 36, p. 11-22, jul./dez. 2000. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/tla/article/view/8639310>

A compreensão do conceito de proficiência na área de avaliação de língua adicional é negativamente afetada por: 1) abrangência do conceito; 2) confusões terminológicas criadas pelo uso técnico e não técnico do termo; e 3) divergências teóricas sobre o que seja saber outra língua. Com relação à abrangência, o conceito de proficiência é de interesse tanto dos construtores de testes quanto de professores, administradores, elaboradores de currículos, pais, alunos e pesquisadores na área dos estudos da linguagem. Sobre as confusões terminológicas, o uso não técnico é ancorado em julgamentos impressionistas e holísticos do desempenho oral, não sendo baseados em resultados de exames ou testes e, por isso, sem a explicitação dos critérios que levam a esses julgamentos. Já o uso técnico considera a proficiência um conceito relativo e múltiplo, em níveis definidos a partir da especificidade da situação de uso da linguagem, em uma gradação de proficiência, segundo a qual todos que possuem certo domínio são considerados proficientes, porém em níveis diferentes. Por fim, as divergentes compreensões teóricas do que seja saber ou dominar uma língua são apontadas como uma das causas para a dificuldade de entendimento do conceito de proficiência em língua adicional.

SHOHAMY, Elana. Assessing multilingual competencies: adopting construct valid assessment policies. *The Modern Language Journal*, Madison, v. 95, n. 3, p. 418-429, out. 2011. Disponível em: <https://www.researchgate.net/publication/230316113_Assessing_Multilingual_Competencies_Adopting_Construct_Valid_Assessment_Policies>.

Para que os testes tenham validade de construto, precisam estar baseados em um construto que siga o entendimento das teorias de linguagem vigentes. Considerando que o ensino, a aprendizagem e a compreensão de línguas atuais seguem as abordagens multilíngues, é preciso destacar a importância de as políticas de testagens de línguas, os procedimentos e as tarefas dos testes estarem em consonância com essas abordagens. Assim, a autora critica as atuais abordagens monolíngues de avaliação argumentando que são instrumentos ideológicos para a criação de uma identidade coletiva e nacional e ignoram a realidade de como as línguas estão sendo usadas. Para corroborar sua posição, apresenta dados empíricos dos custos desse viés político adotado pelos elaboradores de testes, que contribui para perpetuar uma visão limitada e restrita da linguagem. Contrapondo-se a essa visão, apresenta propostas de como os testes podem refletir as abordagens multilíngues atuais em diferentes contextos e destaca o desafio da área de testagem de línguas para desenvolver e inventar testes e rubricas baseados em um construto mais amplo.

TOFFOLI, Sônia F. L.; ANDRADE, Dalton F.; BORNIA, Antonio C.; QUEVEDO-CAMARGO, Gladys. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. *Educação e Pesquisa,* São Paulo, v. 42, n, 2, p. 343-358, abr./jun. 2016. Disponível em: <http://www.scielo.br/pdf/ep/v42n2/1517-9702-ep-42-2-0343.pdf>.

O conceito de validade vem sendo proposto e modificado desde os anos 1920. Aplica-se às interpretações e ações sobre os resultados dos testes para saber se são justificadas, tanto com base nas evidências científicas quanto nas consequências sociais e éticas da utilização do teste. Confiabilidade refere-se à consistência dos escores de avaliação, ou seja, é esperado que um indivíduo alcance o mesmo resultado independentemente da ocasião em que ele respondeu ao teste. A comparabilidade diz respeito à validade das inferências sobre comparações que são feitas com base em resultados de avaliações. Três abordagens são utilizadas: 1) de desempenho, considera apenas as características do teste; 2) de normas estatísticas, leva em conta o desempenho dos examinandos de uma população; e 3) do desempenho em relação ao construto comum, isto é, notas provenientes de diferentes exames podem ser interpretadas como indicando o desempenho do participante em relação ao mesmo traço latente ou construto. As questões sobre justiça estão relacionadas à possibilidade de garantir oportunidades iguais aos participantes, e, para isso, os instrumentos

devem ser apropriados para os vários grupos que serão testados. Uma avaliação de qualidade deve permitir às pessoas oportunidades de respostas que assegurem inferências corretas sobre seu desempenho em relação ao construto medido.

WEIR, Cyril J. Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, Thousand Oaks, v. 22, n. 3, p. 281-300, 2005. Disponível em: <https://journals.sagepub.com/doi/10.1191/0265532205lt309oa>.

A análise das limitações do Quadro Comum Europeu de Referência (QCER) para o desenvolvimento comparável de exames e testes se concentra no contexto de uso da língua, na validade de construto e como eles impactam o desenvolvimento e a comparabilidade de testes. Embora contenha informações valiosas sobre proficiência linguística e aconselhamento para profissionais, o QECR, na sua forma atual, não é abrangente, coerente ou transparente para uso não crítico em testes de língua. A falha em explicar o construto a ser testado dificulta o uso do QECR para desenvolver testes comparáveis entre línguas em níveis equivalentes e impede vincular diferentes sistemas de avaliação. Os elaboradores de exames de língua precisam saber o que e como os aprendizes "são capazes de fazer" em cada nível de uma escala de proficiência, e em que condições os desempenhos ocorrem e com que qualidade. A crítica é de que o QECR precisa atender melhor às demandas de validade de construto, de contexto de uso da língua e de pontuação.

Melissa Santos Fortes, doutora em Estudos da Linguagem pela Universidade Federal do Rio Grande do Sul (UFRGS), é professora adjunta de língua inglesa da Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA). É coordenadora do posto aplicador Celpe-Bras UFCSPA, sendo também aplicadora, corretora e elaboradora desse exame desde 2003. Integra o grupo de pesquisa Interação Social e Etnografia e também o de Ciências da Linguagem, ambos certificados pelo CNPq.

melfortes73@gmail.com

Simone Paula Kunrath, doutoranda em Avaliação de Proficiência em Língua Adicional no Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul (UFRGS), é professora de Português como Língua Adicional (PLA) e coordenadora pedagógica na Escola de Português Bem Brasil, em Porto Alegre.

simone.kunrath@gmail.com

Recebido em 4 de janeiro de 2019
Aprovado em 7 de março de 2019