DESENVOLVIMENTO DO LETRAMENTO EM AVALIAÇÃO DE LÍNGUAS A PARTIR DE UM PROTOCOLO DE REFINO DE CORREÇÕES

Laura Márcia Luiza Ferreira[*]

RESUMO: A avaliação em línguas é indissociável do ensino. A constante expansão dos exames de larga escala gera informações sobre avaliação que podem ser apropriadas por professores de línguas para potencializar sua prática docente. Neste texto, discuto a noção de letramento em avaliação de línguas, bem como os conceitos-chave validade e confiabilidade. Em seguida, apresento um protocolo de refino de correção e argumento como esta prática pode desenvolver a conhecimento por parte dos avaliadores sobre seu comportamento de atribuição de notas às respostas abertas, uma vez que a partir a análise das notas e o refino das correções podem ser uma grande oportunidade para tomada de consciência dos avaliadores sobre suas crenças, seus preconceitos, seus conceitos de ensino-aprendizagem de língua, dentre outros, que são mobilizados no momento de atribuição de notas.

PALAVRAS-CHAVE: avaliação de línguas; formação de professores; letramento em avaliação de línguas; validade.

Letramento em avaliação de línguas

Taylor (2013) afirma que a noção de letramento em avaliação de línguas surge em um momento de expansão do conceito de letramento como habilidade de ler e escrever para diversos domínios e com propósitos distintos. De acordo com a autora, os letramentos

[*] Doutora em Estudos de Linguagem pelo Centro Federal de Educação Tecnológica de Minas Gerais (Cefet-MG). Professora Adjunta da Universidade Federal da Integração Latino-Americana (Unila).

estão relacionados aos aspectos tanto socioculturais como funcionais. O letramento acadêmico, ou seja, relacionado ao contexto da educação superior prevê um conjunto de habilidades complexas que vão além de ler e escrever e do conhecimento cultural que são fundamentais para o sucesso em comunidades acadêmicas. O letramento em avaliação de línguas, segundo Taylor (2013), seria mais um na lista dos diversos domínios de uso da linguagem no âmbito acadêmico e surge em um momento de crescente trabalho na área da avaliação em línguas que pressupõe um contingente de pessoas envolvidas nos processos de elaboração, aplicação e pesquisa relacionadas aos testes.

Scarino (2013) sugere que a formação professional dos professores deva trabalhar a avaliação simultaneamente tanto como prática que transforma avaliação em um benefício para o processo de ensino aprendizagem quanto como para desenvolver nos professores uma auto-compreensão e concientização da natureza do próprio fenômeno de avaliação, seu papel e suas práticas enquanto professores-avaliadores. Interrelacionar tais objetivos implica lidar com preconceitos, crenças, compreensões e visões de mundo sobre avaliação que os professores-avaliadores trazem do seu fazer profissional e de sua formação. De acordo com a autora, o principal desafio ao desenvolver o letramento em avaliação de línguas seria compreender como os professores integram os conhecimentos de diversas disciplinas nas suas práticas e no seu repertório teórico. Scarino (2013) considera dois aspectos como principais nesse processso: (1) a identificação dos domínios relevantes que fazem parte do conhecimento de base; (2) desenvolvimento do letramento em avaliação de línguas. O conhecimento de base, segundo a autora, emerge de questões sobre a natureza do conhecimento humano, como ele se desenvolve e é usado na prática e, por isso, seria preciso também levar em consideração as visões distintas sobre a natureza do conhecimento docente uma vez que o conhecimento sobre o fazer pedagógico é dinâmico. Quanto ao domínio do conhecimento de base, é preciso considerá-lo como uma interseção, ou seja, vários domínios interligados. Nesse sentido, o conhecimento de base diz respeito não só aos paradigmas de avaliação, teorias, propósitos e instrumentos, mas também às teorias e práticas sobre ensino aprendizagem. Nos currículos de cursos de formação de professores,

a avaliação não deveria ser separada das teorias e práticas sobre linguagem, segundo Scarino (2013).

Quanto ao desenvolvimento do letramento em avaliação, a autora basea-se na proposta de Inbar-Lourie. A proposta leva em consideração os seguintes questões: o contexto social da avaliação, a definição e descrição da proficiência, a elaboração e análise de avaliações de larga e pequena escala. De acordo com Scarino (2013), Inbar-Lourie inclue na discussão os contextos sociais, institucionais e educacionais da avaliação assim como as teorias e conceitos relacionados à avaliação tais como validade, confiabilidade, etc. Além disso, Scarino (2013) concorda com a ênfase que Inbar-Lourie atribui aos conceitos específicos do ensino de línguas e sua relação com o letramento em avaliação de línguas tais como a relação entre aprendizagem de primeira e segunda língua e entre língua e cultura. Sobre a questão dos testes de larga escala e os testes relacionados aos currículos, Scarino (2013) se apoia em Inbar-Lourie ao questionar a separação entre análise e desenvolvimento de testes padronizados dos testes relacionados aos currículos dos cursos, pois os dois estariam relacionados e os professores em formação devem saber os pressupostos e as tradições de ambos. A autora ainda concorda com Inbar-Lourie ao incluir no processo de desenvolvimento do letramento em avaliação a exploração de pesquisas sobre avaliação, uma vez que tais conhecimentos devem fazer parte também do repertório do fazer docente e não apenas dos especialistas em avaliação.

Com objetivo de reunir a opinião de um grupo de professores de línguas sobre o que deveria conter em um material didático voltado para formação de docentes com a finalidade de desenvolver o letramento em avaliação de línguas, Fulcher (2012) apresenta o resultado de uma pesquisa exploratória por meio da qual um questionário online foi respondido por 278 professores de línguas de diversos países. Após a análise das respostas fechadas, o autor concluiu que o material deveria tratar das seguintes questões: desenvolvimento e elaboração de teste, exames padronizados de larga escala, avaliações para sala de aula e seu efeito retroativo, validade e confiabilidade. A partir das respostas fechadas, a pesquisa aponta que é também de interesse dos professores a compreensão de conceitos

estatísticos aplicados à avaliação e também de questões práticas como, por exemplo, a verificação da confiabilidade e da validade dos testes ao longo do processo de criação das avaliações. Quanto à discussão sobre validade e confiabilidade no contexto de avaliações para sala de aula e avaliações de larga escala, de acordo com Fulcher (2012), os professores entendem que os conceitos devem ser tratados de forma diferente a depender do contexto. O autor comenta que conceitos como validade e confiabilidade em diferentes contextos de avaliação são discutidos na literatura específica, mas produziram pouco impacto nos materiais pedagógicos destinados à formação do professor. Na última etapa da pesquisa, foi pedido que os professores avaliassem os materiais didáticos que se encontram no mercado para o desenvolvimento do letramento em avaliação. Ao final das avaliações dos livros sobre testes, o autor sintetiza as informações ao resumir que um bom material deveria conter aspectos teóricos explicados de maneira clara, principalmente se tratar de estatística; um guia prático no estilo 'como-fazer', mas que não seja prescritivo; exemplos diversificados de avaliações tanto de sala de aula quanto testes padronizados de larga escala e atividades que relacionam o que foi abordado no material e com o fazer pedagógico do professor. A partir da pesquisa, o autor esboça um conceito de letramento em avaliação que leva em conta os dados levantados. Para Fulcher (2012), o letramento em avaliação poderia ser definido como:

> conhecimento, capacidade e habilidade para elaborar, desenvolver, gerir ou avaliar tanto testes padronizados de larga escala quanto avaliações de sala de aula; trata-se também de estar familiarizado com os processos de avaliação bem como com princípios e conceitos que subzagem à prática, incluindo a ética e os códigos relacionados à tarefa de avaliar. Habilidade de mobilizar conhecimentos, habilidades, processos, princípios e conceitos situados em um contexto histórico político e filosófico com o objetivo de compreender os motivos pelos quais uma determinada prática de avaliação assume determinadas características e de avaliar o papel e impacto da avaliação

> para a sociedade, para a instituição e para os indivíduos. (FULCHER, 2012, p. 114: tradução nossa)[1]

Assim como Scarino (2013) e Taylor (2013), a pesquisa de Fulcher (2012) contribue para a compreensão do conceito de letramento em avaliação de línguas. Os pesquisadores almejam debater uma maneira de desenvolver no âmbito da formação dos professores as habilidades e capacidades relacionadas às práticas de avaliação. Cabe ressaltar que os autores não se limitam a debater o letramento em avaliações como um conjunto de informações e conhecimentos relacionados à avaliação. Os autores compreendem que o desenvolvimento do letramento em avaliação no âmbito da formação dos professores é complexo e vai além da transmissão do conhecimento teórico. Poder-se-ia dizer que é mais ou menos consenso entre os professores a necessidade de contextualizar as avaliações e seus impactos, ampliar o público das pesquisas em avaliação para os professores em formação, debater as questões relacionadas às avaliações tanto no que diz respeito aos testes padronizados quanto para os testes de sala de aula e propor práticas com vistas a desenvolver habilidades e capacidades específicas do letramento em avaliação de línguas estrangeiras.

Abaixo discutiremos dois conceitos centrais relacionados à elaboração e análise de testes para, em seguida, apresentar uma proposta de como operacionalizá-los em um contexto de teste.

Confiabilidade e validade como conceitos centrais para o letramento em avaliação de línguas estrangeiras

[1] The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (FULCHER, 2012, p. 114)

Validade e confiabilidade são conceitos centrais para elaboração e análise tanto de exames de larga escala e como para os voltados para o contexto de sala de aula. Bachman e Palmer (1996) afirmam que a validade e confiabilidade fazem parte de um conjunto de critérios qualitativos que definem a relevância de um teste. Além dessas duas qualidades, os autores ressaltam que a pertinência de um teste pode ser avaliada por meio não só da análise da validade e confiabilidade como também da autenticidade das tarefas e dos textos que compõem o exame, a interatividade entre tarefas e examinandos que o teste pressupõe assim como questões relacionadas ao impacto e praticalidade da aplicação e correção do teste. O professor avaliador deve estar ciente dos fatores ao propor um processo de avaliação e, segundo Bachman e Palmer (1996) tais qualidades devem ser analisadas levando em consideração suas interações umas com as outras.

Bachman (1990) situa os conceitos de validade e confiabilidade nas extremidades de um contínuo, como se ambos fossem faces de uma mesma moeda, porém cada qual com suas especificidades. O autor afirma que a confiabilidade diz respeito à identificação de fontes potenciais de falhas em um determinado processo de atribuição de notas e à criação de estratégias para minimizar os efeitos desses erros nos escores ao passo que a validade está relacionada à falha da medida e instrumento ou de outros fatores externos à habilidade linguística que se quer avaliar. Bachman (1990) se apoia no texto de Messick de 1987 para defender a ideia de que a validade é um conceito único que abarca aspectos relacionados à validade de construto, conteúdo e critério. A validade de construto está relacionada aos aspectos teóricos operacionalizados no instrumento, a de conteúdo aos aspectos relacionados às tarefa ou itens e a de critério aos documentos que estabelecem os processos de atribuição de notas tais como as grades, a granulação dos descritores por nível de proficiência e a própria atuação do avaliador. O conceito de validade seria "o julgamento integrado e avaliativo baseado em estudos empíricos e teóricos que demonstram o quanto

as inferências e ações são adequadas e apropriadas a partir das notas do teste."[2] (MESSICK, 1987. p. 6: tradução nossa).

Quanto aos trabalhos com o foco na confiabilidade, Bachman (1990) afirma que os estudos devem envolver análise lógica e pesquisa empírica de maneira a primeiro identificar as fontes de falhas e estimar o quanto essas falhas influenciam nas notas finais. De acordo com o autor, "estritamente falando, a confiabilidade se refere à nota atribuída e não ao teste em si."[3] (BACHMAN, 1990, p. 171: tradução nossa). O autor agrupa alguns fatores potenciais que interferem na mensuração dos escores em três categorias: facetas do método de avaliação, atributos pessoais, fatores aleatórios e, obviamente, a habilidade comunicativa da língua. Os métodos são sistemáticos e o efeito da uniformização do processo de avaliação na nota pode ser comum a todos os examinandos ou individualizado, por isso a estatística é usada como ferramenta para investigar esses efeitos.

Letramento em avaliação de línguas e a psicometria

Como apontado pelo estudo de Fulcher (2012), a verificação da confiabilidade e da validade dos testes ao longo do processo de criação das avaliações assim como os modelos estatísticos que dão conta da análise das qualidades psicométricas dos testes são temas que interessam aos professores-avaliadores em formação. McNamara e Knoch (2012) problematizam o fato dos pesquisadores da área de avaliação terem uma formação em linguística aplicada com pouco conhecimento em estatística. Entre os especialistas de avaliação de larga escala, os modelos matemáticos para a análise da validade e confiabilidade foram sendo incorporados aos poucos pela comunidade científica. O principal argumento contra a psicometria levantado pelos linguistas aplicados era o fato da natureza da proficiência linguística ser complexa e, por isso, não poder ser mensurada levando em conta uma única

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

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

dimensão estabelecida nos modelos estatísticos. Em geral, os linguistas aplicados ao fazerem a validação dos testes apoiavam se na argumentação sobre o construto do exame. Neste trabalho, defendemos que a validade é uma questão que envolve não só a argumentação sobre o construto teórico que subjaz aos testes como outros aspectos. O processo de validação de testes consiste em levantar evidências que vão além da argumentação sobre as questões teóricas. A análise das notas por meio de procedimentos estatísticos são instrumentos importantes na captura das evidências para discussão da validade de um exame.

Os modelos estatísticos para análise de validade e confiabilidade de testes são fundamentados pela *latent trait theory* ou *iten response theory*. Tais teorias formam o paradigma ainda hoje vigente na área. *Latent trait theory* ou *iten response theory* partem do pressuposto de que a probabilidade de alguém acertar a resposta de um determinado item pode ser calculada por meio de uma função matemática que leva em conta a dificuldade do item e a habilidade do examinando. Um modelo estatístico muito popular entre os especialistas da área de avaliação de larga escala é o *Multi Facet Rasch Measurement* (MFRM). O MFRM compartilha dos mesmos pressupostos teóricos da teoria de resposta ao item, ou seja, envolve também a análise da probabilidade de acertos e erros e os relaciona com cada um dos itens que compõem os testes.

McNamara e Knoch (2012) afirmam que o crescente movimento comunicativo de ensino e aprendizagem de línguas estrangeiras da década de 90 resultou na encorporação de tarefas de produção oral e escrita que simulavam contextos de utilização da língua na vida real foi o contexto em que o Rasch se tornou mais frequente nos estudos sobre avaliação. Os testes também aderiram à perspectiva do ensino comunicativo ao propor avaliações baseadas em desempenho, afastando-se dos itens de múltipla escolha em que as notas eram mensuradas dicotomicamente. Como os testes de desempenho previam o julgamento da performance por um avaliador, o processo de atribuição de escores passou a ser mais complexo. Normalmente, é previsto no processo de atribuição de notas de testes de desempenho que a performance do examinando seja analisada por mais de um avaliador que atribui uma nota independente. A nota é comparada com a nota de um outro avaliador e,

em caso de discrepância, o examinando é avaliado por um terceiro avaliador, que, em geral, é mais experiente. As notas são atribuídas com base no comando das tarefas, em caso de textos escritos, e na grade de critérios. Os critérios refletem o construto da língua, ou seja, a perspectiva teórica sobre língua e linguagem que está sendo operacionalizada no teste. A granulação dos descritores dos critérios por nível de proficiência explicita o que se espera do examinando. Quando o teste é de desempenho, ou seja, elaborado a partir de tarefas que simulam situações reais de uso de linguagem do dia a dia, o avaliador, a tarefa, o texto com o qual a tarefa foi elaborada, assim como a grade de critérios, são considerados facetas do método de avaliação que devem ser estudadas e analisadas. De acordo com Bachman (1990) as facetas do método de avaliação são fatores potenciais que interferem na mensuração dos escores e, por sua vez, na confiabilidade e validade do exame.

Avaliação é um processo de coleta sistemática de informações. Trata-se de uma investigação que levanta inferências sobre alguma coisa. Entender os aspectos desta coleta de informações e utilizá-los para potencializar a relação entre avaliação e ensino, no contexto das línguas, é fundamental para o desenvolvimento do letramento em avaliação de línguas.

Abaixo, trataremos de um dos aspectos relacionados à validade e confiabilidade dos testes em situação de correção de prova de desempenho de línguas.

O julgamento da performance

São diversas as maneiras de elaborar avaliações. O desenho das avaliações é condicionado a diversos fatores tais como: recursos humanos e materiais disponíveis, propósito, construto de língua, quantidade de examinandos, dentre outros. A depender das condições, o desenho da avaliação vai variar quanto às possibilidades de respostas aos itens. Em avaliações de larga escala de múltipla escolha, por exemplo, as possibilidades de resposta são reduzidas para simplificar o processo de correção, ao passo que em provas de desempenho nas quais há questões abertas, a possibilidade de resposta é mais heterogênea, acarretando um maior investimento de tempo e recursos humanos na etapa de correção.

Questões abertas em avaliações de desempenho são interessantes para o contexto de ensino de línguas. Muitas vezes, o construto de língua está relacionado ao uso da linguagem em situações próximas às interações reais, por isso as provas são elaboradas de forma que há inúmeras possibilidades de respostas. Se do ponto de vista do construto as questões abertas são interessantes, pela perspectiva do controle da situação de prova, a correção de textos ou conversas impõem alguns desafios quanto à estabilidade das notas. Os psicometristas, neste ponto, sugerem que para garantir que as notas sejam confiáveis, ou seja, que sejam estáveis quando se altera o avaliador ou a edição da prova, por exemplo, é preciso investir em correção por pares, controle de discrepâncias e elaboração de parâmetros de avaliação que sejam compartilhados por todos.

Como o desempenho na tarefa é avaliado por mais de um avaliador, é importante considerar duas situações que podem comprometer os resultados da avaliação. Uma delas é a consistência da atribuição de notas em comparação com outros avaliadores (*inter-rater*) e a outra é quando se compara o avaliador com ele mesmo (*intra-rater*). A inconsistência entre avaliadores distintos (*inter-rater*) pode ser causada pela compreensão distinta das grades de avaliação o que acarreta em uma interpretação não consensual dos parâmetros de correção. É possível explicar a inconsistência de atribuição de notas entre avaliadores a partir dos parâmetros de avaliação. Pode ser que os parâmetros ou a grade não esteja organizada de forma consistente, ou seja, o problema pode ser o critério em si ou a interpretação do desempenho do examinando tendo com base nos critérios. Examinar a consistência entre avaliadores é também uma maneira de estudar as hipóteses que são relevantes para a validade no contexto de avaliações de larga escala, mas o que pretendemos argumentar aqui é que é possível fazer esta análise em contextos menores de avaliação com uma equipe reduzida de professores-avaliadores. Antes de apresentar uma metodologia para esta análise, trataremos brevemente de como este tipo de análise é feita em situações de testes maiores.

Segundo McNamara e Knoch (2012), *Multi-facet Rasch measurement* é uma ferramenta potencial para analisar o efeito do avaliador na nota final. Os autores justificam

as potencialidades do modelo ao afirmar que

> as características dos avaliadores, que foram abertas para a pesquisa detalhada usando o método Rasch, incluem relativa leniência ou severidade, grau de consistência nas notas, influência do treinamento de avaliadores, influência do background profissional e consistência da atribuição de notas com o passar do tempo [...] É como se os pesquisadores da área tivessem em mãos um poderoso microscópio para examinar a complexidade do processo de atribuição de notas[4] (MCNAMARA E KNOCH, 2012, p. 568: tradução nossa)

Os trabalhos típicos que empregam a ferramenta MFRM tem como foco a análise do julgamento do desempenho pelos avaliadores a partir de dados coletados de testes comunicativos tanto para analisar as qualidades psicométricas do teste quanto para discutir questões relacionadas a sua validade. Mais recentemente, os autores ressaltam que a ferramenta Rasch tem sido usada inclusive para fornecer um retorno aos avaliadores, no que diz respeito, ao seu comportamento de atribuição de notas em relação a outros avaliadores. Como o modelo fornece análise sobre o comportamento de avaliadores em comparação com outros, tais informações podem ser usadas para a reflexão sobre a prática de julgar o desempenho tendo ou não como base os parâmetros preestabelecidos nas grades de avaliação.

O estudo sobre interpretação do desempenho de examinandos baseado em critérios de testes de larga escala ou previamente elaborados por um grupo de professores pode ser um exercício para desenvolver o letramento em avaliação de línguas. O manuseio de softwares estatísticos como os que operacionalizam o MFRM ou outras teorias de medida requer conhecimento especializado e investimento financeiro, pois são pacotes pagos. Porém, é possível, em um contexto menor, utilizar do princípio do controle de discrepância para potencializar a consciência dos avaliadores ao atribuírem suas notas. A discussão das

[4] Rater characteristics which were now open for detailed research using Rasch methods included relative severity or leniency; degree of consistency in scoring; the influence of rater training; the influence of professional back-ground; and consistency over time. […] It is as if researchers in this field had been handed a very powerful microscope to examine the complexity of the rating process. (McNAMARA E KNOCH, 2012,568p.)

discrepâncias entre avaliadores que compõem uma determinada equipe de avaliação pode ser usada como uma instância de formação de futuros professores de forma a desenvolver o letramento em avaliação de línguas. Como os parâmetros de avaliação que compõem as grades dos exames refletem construtos teóricos, ao participarem de uma análise sobre as notas entre avaliadores, os professores seriam convidados a refletirem sobre as teorias de ensino de línguas e os construtos sobre línguas para operacionalizá-las em instrumentos que norteiem o julgamento das performances. Trabalhar com grades de critérios de exames de proficiência é tornar possível a identificação das teorias que fundamentam o exame por parte dos avaliadores. Por meio do exercício de explicitar a complexidade do processo de atribuição de notas em exames de desempenho, os professores-avaliadores poderiam lidar com os preconceitos, as crenças e as compreensões e visões de mundo sobre avaliação que emergiriam ao julgar o desempenho dos examinandos.

Tendo argumentado sobre as potencialidades do exercício de avaliação por pares e controle de discrepâncias, descrevo abaixo um protocolo de correção.

Protocolo de refino de correção de questões abertas

A construção de avaliações é um processo complexo, neste texto, me concentrarei na descrição e discussão de um protocolo de refino de correção baseado no documento *Calibration protocol for scoring student work,* do Departamento de Educação de Rhode Island, nos Estados Unidos.

De maneira geral, o objetivo da correção é refinar parâmetros de avaliação de desempenho de estudantes para gerar insumo para repensar as práticas pedagógicas ou a própria a avaliação. Vale ressaltar que esta proposta está relacionada à avaliações internas em contextos de ensino e avaliações de línguas, podendo ser adaptados para outras disciplinas que envolvem questões abertas.

Estima-se um tempo de preparação e planejamento para esta tarefa de 8 a 10 horas, podendo ser uma equipe de professores formada por 4 a 8 integrantes. Os materiais necessários para o refino seriam a tarefa ou questões, as respostas dos estudantes, a grade

de avaliação com os descritores e uma folha de registro de avaliação. Inicialmente todos os integrantes juntos fazem a leitura da tarefa e uma análise das grades de avaliação. Neste momento, são sanadas e discutidas as dúvidas, troca-se impressões sobre as tarefas, sobre os critérios, etc. Trata-se de uma compreensão geral do que foi pedido e de como corrigir, segundo as grades de avaliação. Neste momento, quem elaborou a avaliação pode explicar aspectos da tarefa, da grade, dentre outros aspectos relacionados à elaboração e aplicação do exame. Em seguida, os professores-avaliadores leem ou escutam as respostas dos estudantes com o objetivo de procurar provas típicas, ou seja, exemplos de produções textuais ou orais que se encaixam nas faixas de notas previamente estipuladas na grade de avaliação. Cada integrante deve fornecer uma justificativa da nota atribuída a partir da grade. Neste momento, haverá uma discussão da avaliação de provas típicas e refino dos parâmetros. Em seguida, haverá uma análise de atribuições discrepantes, muita discussão, eventualmente a edição da grade e/ou o estabelecimento de consenso. O mais importante nesta etapa é que todos avaliadores compreendam mais ou menos da mesma forma a grade de avaliação. Em seguida, cada avaliador atribui de forma individual e independente uma nota às respostas. Em situações em que se exige um controle mais rígido dos resultados, pode ser que um integrante fique responsável por comparar as notas de todos integrantes e resolver posteriores atribuições de notas que estejam divergentes. Nesta comparação das notas, será possível perceber aqueles avaliadores mais lenientes e os mais severos, ou seja, os que tem uma tendência para atribuir notas mais altas ou mais baixas. Estudar os dados planilhados, quer dizer, as tabelas das notas atribuídas a uma mesma resposta por avaliadores diferentes pode ser uma grande oportunidade para os avaliadores podem tomar consciência do seu comportamento de correção. Neste ponto, é possível também investigar o que motiva certos comportamentos de avaliação: será que algum critério específico da grade foi supervalorizado ou subvalorizado por algum dos avaliadores?

Após a etapa de refino de correção e da correção de todas as repostas, os professores-avaliadores podem refletir sobre outros aspectos, a saber: o desempenho dos estudantes, ou seja, o que se pode perceber a partir das respostas às tarefas e do uso da grade

de avaliação; os próximos passos para o ensino, ou seja, as informações geradas na avaliação direcionam para quais ações na sala de aula ou para o contexto do ensino de língua; a necessidade ou não de revisão da tarefa ou da grade de avaliação.

Considerações finais

A avaliação em línguas é um campo de atuação profissional dos egressos dos cursos de licenciatura em constante expansão devido à demanda de um contingente de pessoas capacitadas para atuar em contextos de exames de proficiência linguística de larga escala. No contexto de formação de professores, o tema da avaliação é indissociável do ensino. Por estes motivos, discutir estratégias de desenvolvimento do letramento em avaliação de línguas se faz necessário.

A avaliação é um tema interdisciplinar e colaboram para a questão diversas áreas do conhecimento. Noções como as de validade e confiabilidade são conceitos-chave para compreensão de como se estrutura um instrumento de avaliação. Por este motivo, a partir das noções de validade e confiabilidade apresentei um exemplo prático de possibilidade de desenvolvimento do letramento de avaliação de línguas que pode ser utilizado tanto em contexto de formação de professores como por professores em serviço.

O protocolo de refino de correções é uma constante em situações de avaliação de larga escala para garantir a estabilidade ou confiabilidade dos resultados das avaliações. Neste trabalho, argumentei que o protocolo de refino de correções pode ser também uma prática que os professores possam se apropriar e a partir da qual potenciais discussões sobre crenças e compreensão sobre ensinar e aprender línguas possam ser compartilhadas e discutidas.

LANGUAGE ASSESSMENT LITERACY IMPROVEMENT AND CALIBRATION PROTOCOL FOR SCORING STUDENTS WORK

ABSTRACT: Language assessment and teaching are inseparable. The intense demand for large scale language exams results in more information about language assessment. Language teachers should use these information to improve their practice in classroom. In this article, I discuss key concepts

such as language assessement literacy, validity and confiability. Then, I discuss how a calibration protocol to score students work can develop raters awareness of their behavior while assigning scores. I argue that this very simple activity can be used to improve teacher's language assessement literacy since rating activities involves beliefs, prejudices, language teaching constructs and so on.

KEYWORDS: language assessment; language assessment literacy; teacher training; validity.

REFERÊNCIAS

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

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

FULCHER, Glenn. Assessment literacy for the language classroom. *Language Assessment Quarterly,* n.2, v.9, p. 113-132. 2012

MCNAMARA, Tim; KNOCH, Ute. The Rasch wars: the emergence of Rasch measurement in language learning. *Language Testing.* n.24. v.4, p. 555-576, 2012.

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

RHODE ISLAND. Departamento de Educação. *Calibration protocol for scoring student work*. Estados Unidos. Disponível em: http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Calibration_Protocol_for_Scoring_Student_Work.pdf Acesso em 28 mai. 2019.

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

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

Recebido em: 29/05/2019.

Aprovado em: 25/07/2019.