A utilização do modelo multifacetas de Rasch na análise das influências dos avaliadores sobre as avaliações com itens abertos[1]

Sônia Ferreira Lopes Toffoli [a]
Cristina Valeria Bulhões Simon [b]

Resumo

Este trabalho analisa a qualidade da pontuação em avaliações com itens abertos por meio do modelo multifacetas de Rasch (MFR). São utilizadas as pontuações atribuídas às redações de participantes do Processo Seletivo Vestibular da Universidade Estadual de Londrina, de 2015. O modelo MFR pode proporcionar estudos, tanto no nível de grupo quanto no nível individual, possibilitando a identificação de avaliadores portadores de comportamentos tendenciosos, conhecidos por causarem erros importantes nas pontuações de tarefas escritas. As análises no nível de grupo mostraram que a avaliação foi eficiente e que os dados, de modo geral, são adequados às expectativas de medição dos modelos de Rasch e, por meio das análises no nível individual, foi possível detectar avaliadores que pontuaram diferentemente do modo como os outros avaliadores, em média, atribuíram suas pontuações. O modelo MFR mostrou-se uma ferramenta adequada e eficiente para o monitoramento da qualidade das pontuações atribuídas às tarefas de escrita.

Palavras-chave: Modelo multifacetas de Rasch. Itens abertos. Provas de redação. Avaliação em larga escala. Tendências do avaliador.

1 Introdução

Atualmente, no setor educacional, há uma forte tendência para que as avaliações estejam mais direcionadas à avaliação da aprendizagem, o que tem intensificado o interesse pelas avaliações com itens abertos.

[1] Apoio: COPS/UEL: Coordenadoria de Processos Seletivos da Universidade Estadual de Londrina.

[a] Universidade Estadual de Londrina. Departamento de Matemática. Londrina, Paraná, Brasil.

[b] Universidade Estadual de Londrina. Departamento de Letras Vernáculas e Clássicas. Londrina, Paraná, Brasil

Recebido em: 22 set. 2016
Aceito em: 19 out. 2017

Nessas avaliações, muitos fatores podem afetar a medida do desempenho das pessoas. Em primeiro lugar, está a habilidade do examinando. Entretanto, a pontuação que receberá no exame não depende apenas da sua capacidade ou do conhecimento sobre o traço que está sendo medido; depende também da severidade do avaliador, da dificuldade das tarefas, do tema abordado e de outras variáveis que podem interferir em cada evento de avaliação em particular (JONSSON; SVINGBY, 2007).

As avaliações com itens abertos possuem algumas questões críticas. Entre elas está a diferença na maneira com que os diversos avaliadores da equipe de correção pontuam as tarefas. A utilização de critérios bem estabelecidos, a experiência e o treinamento dos avaliadores são fatores importantes na obtenção de bons índices de confiabilidade, mas também é necessário considerar as tendências dos avaliadores em julgamentos sistemáticos dos desempenhos avaliados. Essas tendências são frequentemente citadas nas pesquisas e consideradas componentes geradores de erros importantes na pontuação de testes com itens abertos (TOFFOLI, 2015; ECKES, 2011; MYFORD; WOLFE, 2004).

As influências de julgamentos tendenciosos mais populares são: a) severidade/complacência: tendência em avaliar de maneira muito exigente ou muito branda as tarefas elaboradas pelos examinandos; esses avaliadores atribuem pontuações que são, em média, inferiores ou superiores às pontuações atribuídas pelos demais avaliadores do grupo; b) tendência central: propensão a classificações iguais ou perto do ponto médio da escala de classificação, evitando classificações nos extremos da escala; esse avaliador pode ser incapaz de julgar com precisão níveis de desempenho nos extremos, ou de fazer distinções entre quaisquer categorias atribuindo pontuações no meio da escala; c) halo: tendência em atribuir pontuações semelhantes a um mesmo item, para muitos examinandos, isto é, os examinandos recebem pontuações semelhantes, mesmo que os seus desempenhos tenham sido muito diferentes; d) aleatoriedade: tendência de se aplicar uma ou mais categorias da escala de maneira inconsistente com o modo com que os outros avaliadores, em média, aplicam a mesma escala.

As avaliações, de modo geral, têm consequências para as pessoas avaliadas, no ambiente escolar ou em outras esferas. As avaliações devem ser consistentes, focadas principalmente na confiabilidade da medição, com julgamentos honestos e baseados em evidências (TOFFOLI; ANDRADE; BORNIA, 2015). A pontuação atribuída ao respondente deve ser independente do avaliador, e as pontuações devem ser semelhantes quando se avalia um mesmo desempenho, mesmo que a tarefa tenha sido cumprida em outra ocasião (JONSSON; SVINGBY, 2007; STEMLER, 2004). Assim, o esforço deve ser no sentido de garantir dois tipos de confiabilidade:

interavaliador, no qual os avaliadores concordam uns com os outros em suas notas, e intra-avaliador, isto é, cada avaliador atribui a mesma pontuação para um determinado desempenho avaliado em ocasiões distintas (STEMLER, 2004).

As instituições promotoras de avaliações em larga escala, frequentemente, citam o treinamento dos avaliadores como condição necessária para garantir a qualidade das pontuações (MYFORD; WOLFE, 2004; STEMLER, 2004). Nesse caso, exige-se que as pontuações provenientes de dois avaliadores que julgaram a mesma tarefa tenham um determinado grau de concordância, gerando, assim, os índices de confiabilidade da pontuação.

Outra abordagem para garantir a qualidade da pontuação, geralmente denominada de estimativas de medição, consiste em utilizar toda a informação disponível, a partir de todos os avaliadores, inclusive as notas discrepantes, para fornecer um indicador mais robusto do grau de concordância dos avaliadores ao atribuir as pontuações. Cada avaliador fornece algumas informações exclusivas, úteis para a geração da pontuação. Assim, não é necessário que dois avaliadores concordem perfeitamente no modo de aplicar os critérios de pontuação, basta que eles mantenham o mesmo padrão em suas próprias pontuações. Desse modo, as diferenças entre as pontuações dos avaliadores podem ser estimadas e compensadas na nota final de cada participante (ECKES, 2011; STEMLER, 2004).

Para calcular a confiabilidade da pontuação dentro dessa categoria de estimativas de medição, vem sendo cada vez mais utilizado o modelo multifacetas de Rasch (MFR). Esse método é uma extensão do modelo da TRI de um parâmetro (modelo de Rasch) desenvolvido por Linacre em 1989.

Por permitir a inclusão de outros parâmetros, além da dificuldade dos itens, o modelo MFR possibilita aos pesquisadores análises para os efeitos individuais causados pelos elementos que fazem parte da avaliação, ou seja, cada examinando, cada avaliador, cada uma das tarefas etc. Essa possibilidade de obter informações que possam servir de diagnóstico, em nível individual, sobre o funcionamento de cada elemento é considerada valiosa e torna a utilização do modelo MFR muito vantajosa.

O objetivo desse estudo é verificar como o modelo MFR pode ser utilizado para monitorar a qualidade da pontuação de avaliações com itens abertos. Para isso, é feita uma investigação da pontuação atribuída à questão de redação que faz parte da edição de 2015 do Processo Seletivo Vestibular da Universidade Estadual

de Londrina (UEL). As análises no nível de grupo possibilitaram verificar a qualidade da pontuação de modo geral. No nível individual, para exemplificar, são identificados alguns avaliadores portadores dos efeitos severidade/complacência, tendência central, aleatoriedade e halo.

2 Metodologia

2.1 Contexto do estudo

O Vestibular da UEL atualmente conta com duas fases. Na primeira, os candidatos respondem a uma prova objetiva. Na segunda fase, dividida em até três dias[1], os candidatos aprovados na primeira fase respondem, no primeiro dia, a questões de múltipla escolha de língua portuguesa e literaturas de língua portuguesa e de língua estrangeira, bem como produzem, no mínimo, duas e, no máximo, quatro redações; já no segundo dia, os candidatos respondem a questões discursivas (UNIVERSIDADE ESTADUAL DE LONDRINA, 2014).

Na prova de redação, os candidatos deverão produzir textos em prosa, de acordo com o enunciado de cada uma das propostas. No Vestibular 2015, a prova de redação contou com três tarefas: em cada uma delas, o candidato deveria realizar um texto, utilizando de 8 a 12 linhas, conforme as especificações. As propostas de redação deste exame podem ser conferidas em UEL (2015).

Os três itens de respostas abertas na prova de redação do Vestibular 2015 foram concebidos para avaliar a capacidade de expressão escrita dos candidatos. Os aspectos desse construto, previstos no Manual do Candidato para serem avaliados foram: observância à norma padrão do português brasileiro; as atividades de analisar, resumir, comentar, comparar, criticar, completar etc.; aspectos discursivos, textuais, estruturais e normativos (UEL, 2014).

2.2 Avaliadores

A equipe de avaliadores das redações do Vestibular da UEL (2015) foi formada por 10 indivíduos pós-graduados em Letras, a maioria deles com experiência na pontuação de provas de redação dos vestibulares dos anos anteriores e, também, de outros concursos. O treinamento dos avaliadores para realizar a pontuação dos textos foi feito pela abordagem de grupo hierárquico de coordenação, no qual o avaliador coordenador decide como os critérios de pontuação e as normas devem ser interpretados (TOFFOLI, 2015).

[1] O terceiro dia de provas destina-se a avaliar os candidatos aos cursos que exigem uma Prova de Habilidades Específicas.

2.3 Processo de pontuação

Os textos dos candidatos foram analisados segundo a pontuação holística, que se baseia na ideia de que a construção da escrita é uma entidade única e que pode ser capturada por uma única escala que integra as qualidades inerentes do texto (TOFFOLI, 2015). Neste tipo de pontuação, após treinamento dos avaliadores, no qual são discutidos os critérios e o modo como a escala de pontuação deve ser entendida, cada avaliador resume a sua pontuação para a tarefa em uma única nota. As notas são constantemente monitoradas e, quando há discrepância, em mais de um ponto, entre as pontuações, a tarefa é submetida a uma terceira correção, e os respectivos avaliadores são submetidos a novo treinamento. Cada uma das tarefas foi avaliada em uma escala que varia de 1 a 6 pontos.

2.4 Dados

Os dados utilizados neste estudo provêm da correção das provas de redação de 8.955 candidatos participantes da segunda fase do Vestibular da UEL, de 2015. Ao todo, este estudo contou 55.480 dados, sendo esses resultantes de, no mínimo, duas correções de cada uma das três tarefas de escrita. Entre esses dados estão incluídas, inclusive, as notas discrepantes.

Para as análises dos dados, foi utilizado o *software Facets* versão 3.71.4 (LINACRE, 2014a) com os critérios de convergência padrão do programa, ou seja, o procedimento de estimação é o "máxima verossimilhança incondicional". Para a descrição desse método, ver Toffoli (2015) e Linacre (1989). O tamanho da maior pontuação residual marginal é de 0,5, e a diferença máxima entre as mudanças em qualquer uma das medidas é de 0,01 *logito*. Mais detalhes sobre os processos de estimação podem ser verificados em Linacre (2014b). O processo de estimação terminou automaticamente após 131 iterações para o modelo de escala gradual, e após 195, para o modelo de crédito parcial.

3 O Modelo multifacetas de Rasch

O modelo MFR é utilizado para examinar a qualidade psicométrica de uma variedade de avaliações que necessitam de julgamento de avaliadores. A maioria desses estudos é aplicada às avaliações da linguagem (ENGELHARD Jr.; WIND, 2013; LINACRE; WRIGHT, 2002). Aos poucos, porém, a utilização do modelo MFR ganha espaço para análises de avaliações em outras áreas, como avaliação para acesso a programa de residência médica (SEBOK et al. 2015), análises sobre a qualidade da pontuação (ILHAN, 2016), estudos sobre a criatividade (LONG; PANG, 2015), entre outros inúmeros exemplos.

A utilização do modelo MFR é indicada quando o teste é composto por variáveis que podem contribuir para a ocorrência de erros nas medidas. No contexto do MFR, essas variáveis são denominadas "facetas", que podem ser incluídas ou retiradas do modelo conforme as necessidades do estudo. Na área educacional, em testes com itens abertos, as facetas normalmente consistem em variações de examinandos, tarefas, itens, critérios, avaliadores e sessões de pontuação.

O modelo de Rasch básico é utilizado quando há apenas duas categorias de respostas, correta/incorreta. Várias extensões do modelo de Rasch foram desenvolvidas para itens de respostas politômicas, que se baseiam, por exemplo, em escalas Likert, ou, então, itens nos quais as respostas são construídas pelos examinandos, e as notas são atribuídas com base em uma escala gradual.

Duas extensões do modelo de Rasch para itens politômicos com categorias de respostas ordenadas são de grande importância para a definição do modelo multifacetas de Rasch. O primeiro é o modelo de escala gradual (ANDRICH, 1978[2] *apud* ECKES, 2011), adequado para itens com categorias de respostas ordenadas igualmente espaçadas; o segundo é o modelo de crédito parcial (MASTERS, 1982[3] *apud* ECKES, 2011), adequado quando cada item possui uma estrutura de categorias de respostas própria.

Neste estudo, o modelo utilizado é constituído de três facetas: (1) habilidade dos examinandos, (2) severidade dos avaliadores e (3) dificuldade dos itens. São implementados tanto o modelo de escala gradual quanto o modelo de crédito parcial, cada um possibilitando análises específicas.

O modelo MFR para escala gradual (LINACRE; WRIGHT, 2002) é dado pela equação:

$$ln\left(\frac{P_{jihk}}{P_{jih(k-1)}}\right) = \theta_j - b_i - c_h - d_k \qquad (1)$$

P_{jihk}: probabilidade de o indivíduo *j* ser classificado na categoria *k* do item *i* pelo avaliador *h.*

$P_{jih(k-1)}$: probabilidade de o indivíduo *j* ser classificado na categoria *k-1* do item *i* pelo avaliador *h*.

[2] ANDRICH, D. A rating formulation for ordered response categories. *Psychometrika*, v. 43, n. 4, p. 561–573, 1978.

[3] MASTERS, G. N. A Rasch model for partial credit scoring. *Psychometrika*, v. 47, n. 2, p. 149–174, 1982.

θ_j: **habilidade** do indivíduo *j*.

b_i: **dificuldade** do item *i*.

c_h: **severidade** do avaliador *h*.

d_k: **tamanho do passo** *k*. É o parâmetro da dificuldade. Regula a probabilidade de ser atribuída ao indivíduo a categoria *k* em relação à categoria *k-1* do item *i*.

Nesse modelo, cada item do teste é caracterizado por uma dificuldade b_i, cada examinando, pela capacidade θ_j, e cada avaliador, por um nível de severidade c_h. A equação 1 coloca todos esses parâmetros em uma escala comum na unidade *log-odds* ou *logitos*.

Cada categoria sucessiva representa "um passo" de melhoria de desempenho, em relação à categoria anterior no traço avaliado. A função desse parâmetro (d_k) é indicar como os dados devem ser tratados em relação à escala de classificação. Desse modo, para se obter o modelo MFR, que permite que cada avaliador utilize a sua própria estrutura de escala de classificação para cada item, é necessário trocar o termo d_k na equação (1) por d_{hk}. Esse modelo é, então, denominado Modelo MFR para escala de crédito parcial. A sua equação é dada por:

$$ln\left(\frac{P_{jihk}}{P_{jih(k-1)}}\right) = \theta_j - b_i - c_h - d_{hk} \quad (2)$$

O termo d_{hk} significa a dificuldade da categoria *k,* relativa à categoria *k-1* do item *i,* como esta foi utilizada pelo avaliador *h*. Os outros parâmetros são definidos igualmente na equação (1).

Neste estudo, primeiramente é utilizado o modelo MFR para escala gradual, isto é, as localizações das categorias não variam entre os itens (equação 1), possibilitando uma visão geral do comportamento das facetas. Na sequência, é implementado o modelo MFR para escala de crédito parcial (equação 2), permitindo análises individuais do comportamento dos avaliadores.

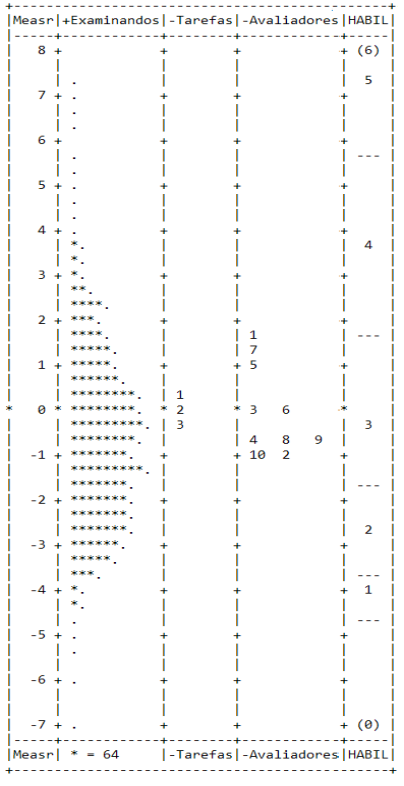
As medidas de ajuste e as estatísticas de separação são utilizadas com o objetivo de avaliar a adequação dos dados aos modelos e à qualidade das pontuações. Algumas interpretações dessas medidas serão feitas no decorrer deste trabalho, entretanto as formulações matemáticas e outras interpretações são encontradas em Eckes (2011), Toffoli (2015) e Toffoli, Andrade e Bornia (2015).

4 Resultados

4.1 Estudos no nível de grupo

O programa *Facets* calibra as medidas de todas as facetas inseridas no modelo com a mesma unidade de medida, *log-odds* ou *logitos*, e exibe todas, simultaneamente, no mapa das variáveis (Figura). Assim, é possível visualizar as distribuições de todos os elementos participantes de cada faceta.

A primeira coluna exibe a escala *logitos*. Todos os elementos das facetas são exibidos de acordo com essa escala, possibilitando comparações entre os elementos. A segunda coluna refere-se à faceta "Examinandos" e apresenta as medidas estimadas da habilidade dos examinandos no exame. Cada asterisco representa 64 examinandos, o de maior habilidade está localizado na parte superior da coluna.



Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).
Figura. Mapa das variáveis.

A terceira coluna trata da faceta "Tarefas": as que aparecem na parte de cima da coluna foram consideradas mais difíceis pelos examinandos. A quarta coluna refere-se à faceta "Avaliadores" e traz a distribuição dos avaliadores conforme seus níveis de severidade: na parte de cima, estão os mais severos.

A quinta coluna exibe a escala de habilidades, de sete pontos, 0 a 6, na qual os avaliadores pontuaram as tarefas avaliadas nas redações. O modelo MFR utilizado é o modelo de escala gradual. Nele, as distâncias entre cada par de categorias adjacentes são as mesmas para todos os itens. Os intervalos entre as categorias não são igualmente espaçados, significando que cada categoria corresponde a diferentes "quantidades" do traço latente.

Neste estudo, aproximadamente 5,6% das respostas válidas tiveram os resíduos padronizados (valores discrepantes entre os dados observados e o que seria esperado com base no modelo), em valores absolutos, iguais ou maiores que 2, e, entre elas, menos de 0,49% iguais ou superiores a 3. Esses resultados indicam um ajuste satisfatório dos dados ao modelo (LINACRE, 2014b).

A Tabela 1 traz a forma como os avaliadores utilizaram cada uma das categorias da escala. Observa-se que, como grupo, os avaliadores não mostraram tendências de severidade nem de complacência, pois não há uso excessivo das categorias dos extremos da escala. As categorias que receberam maior número de observações foram as categorias do centro da escala.

Quando a maior parte dos avaliadores apresenta efeito de tendência central, ocorre uma falta de variação entre a pontuação atribuída para os desempenhos avaliados com essas pontuações, acumuladas nos pontos centrais da escala (MYFORD; WOLFE, 2004). Nota-se, na Tabela 1, que os avaliadores, no nível de grupo, apresentam comportamento que pode caracterizar o efeito da tendência central. As pontuações se acumulam nas categorias 2, 3 e 4, sendo que apenas 4% do total se situa nas categorias 5, 6 e 0. A análise de outros índices, no decorrer deste trabalho, poderá esclarecer se os avaliadores apresentam ou não essa tendência no nível de grupo.

Tabela 1. Resumo da utilização das categorias da escala de avaliação.

Categorias	0	1	2	3	4	5	6
Utilização	3%	5%	23%	50%	19%	1%	0%

Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).

A Tabela 2 traz um resumo dos resultados no nível de grupo para cada uma das facetas.

As medidas de ajuste são centrais para a avaliação da qualidade dos dados utilizados para a construção das medidas e são calculadas para cada uma das facetas especificadas no modelo. A expectativa para as medidas de ajuste MQ-*infit* e MQ-*outfit* é para valores próximos de 1 e não existem regras rígidas para o estabelecimento de limites superiores e inferiores para seus valores, no entanto Wright e Linacre (1994) sugerem que, se os valores destas medidas estiverem no intervalo entre 0,5 e 1,5, os dados em estudo são produtivos para a construção das medidas de Rasch.

Observa-se, na Tabela 2, que as medidas de ajuste variam entre 0,96 e 1,07, portanto dentro do intervalo sugerido por Wright e Linacre (1994) para a adequação dos dados ao modelo.

Em avaliações, para se medirem a dificuldade de itens e, consequentemente, a habilidade das pessoas, é necessário que seja possível a comparação entre a dificuldade dos itens para a localização deles e as habilidades das pessoas em uma

Tabela 2. Resumo das análises estatísticas.

	Examinandos	Avaliadores	Tarefas
Medidas básicas			
Média	-1,54	0	0
Desvio-Padrão	0,69	0,50	0,18
Número	8.955	10	3
Medida de ajuste (MQ-*infit*)			
Média	0,96	0,98	0,99
Desvio-Padrão	0,82	0,35	0,14
Medida de ajuste (MQ-*outfit*)			
Média	0,96	1,07	0,99
Desvio-Padrão	0,82	0,22	0,11
Estatísticas de separação			
Taxa de separação	2,50	11,49	15,26
Confiabilidade separação	0,86	0,99	1,00
Estrato	3,67	15,65	20,68
Qui-quadrado (χ^2)	76.618,50 (p = 0,00)	5.693,10 (p = 0,00)	699,00 (p = 0,00)
Graus de liberdade	8.954	9	2

Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).

escala (MYFORD; WOLFE, 2004). Se os itens (ou as pessoas) se encontram muito próximos ou mesmo muito afastados, uns dos outros, ao longo da escala, pode não ser possível uma medição útil (WRIGHT; STONE, 1999). Para as análises das localizações dos elementos envolvidos no modelo MFR, são utilizadas quatro estatísticas de separação com foco em cada uma das facetas. São elas: 1) o índice de homogeneidade, que indica que as medidas de, pelo menos, dois elementos da faceta são significativamente diferentes; 2) a taxa de separação, que fornece a propagação das medidas dos elementos da faceta em relação à precisão dessas medidas; 3) o estrato, que é o número de níveis estatisticamente diferentes das medidas dos elementos da faceta; 4) a confiabilidade do índice de separação, que é calculada como a razão da variância verdadeira das medidas pela variância observada dessas medidas (MYFORD; WOLFE, 2004).

Na Tabela 2, para a faceta "avaliadores", a taxa de separação em 11,49 logitos significa que as diferenças entre os níveis de severidade dos avaliadores são mais de onze vezes maiores que o erro dessas medidas, não sugerindo, pois, um efeito de tendência central no nível de grupo para esses avaliadores.

O índice estrato representa a variação "verdadeira" em unidades da variância do erro. Seu valor é de 15,65, sugerindo que há mais de quinze estratos estatisticamente diferentes de níveis de severidade entre os avaliadores do grupo, o que também não sugere um efeito de tendência central no nível de grupo para esses avaliadores.

A confiabilidade do índice de separação dos avaliadores fornece informações sobre como os avaliadores são separados, quanto aos seus níveis de severidade, e reflete as variações indesejadas entre os níveis de severidade dos avaliadores. O valor deste índice em 0,99 sugere que, em média, os avaliadores exercem níveis de severidade muito diferentes. O ideal é que os valores desse índice sejam pequenos, perto de zero, sugerindo que os avaliadores podem ser intercambiáveis (MYFORD; WOLFE, 2004; ENGELHARD Jr.; MYFORD, 2003).

O teste do qui-quadrado, com hipótese nula de que as medidas de severidade dos avaliadores não são significativamente diferentes, indica resultados estatisticamente significativos com o valor do qui-quadrado em 5.693,10 com 9 graus de liberdade e $p < 0,05$. Isso significa que as medidas da severidade de, pelo menos, dois dos avaliadores do grupo são significativamente diferentes.

Para a faceta "examinandos", o índice estrato é de 3,67 logitos, o que sugere que há apenas um pouco mais de três estratos estatisticamente diferentes para o desempenho dos examinandos. Como são esperadas diferenças significativas

entre os níveis de desempenho dos examinandos, este fato pode indicar efeito de tendência central em termos de grupo para os avaliadores.

A confiabilidade do índice de separação para os examinandos indica a confiabilidade na qual a avaliação separa as pessoas da amostra, em relação aos seus desempenhos, mostrando o grau com que os avaliadores foram capazes de distinguir, de forma segura, os padrões de desempenho. A confiabilidade de separação dos examinandos é 0,86 logito. Para um índice que assume valores entre 0 e 1, os avaliadores puderam distinguir, de forma confiável, os níveis de desempenho avaliados. Portanto, esse indicador não sugere um efeito tendência central para o grupo de avaliadores.

O teste qui-quadrado, com hipótese nula de que todos os examinandos possuem o mesmo nível de desempenho, tem valor qui-quadrado de 76.618,50 com 8.954 graus de liberdade e é uma medida estatisticamente significativa ($p < 0,05$), indicando que a habilidade dos examinandos varia entre os níveis da escala de pontuação.

Esses índices descritos para a faceta dos avaliadores, juntamente com a faceta dos examinandos, sugerem não existir um efeito de tendência central no nível de grupo para os avaliadores.

4.2 Estudos no nível individual

O modelo MFR para escala de crédito parcial (equação 2) pode fornecer índices referentes às medidas de cada elemento de cada faceta, possibilitando análises individuais. Nessa aplicação, o interesse é analisar o modo como os avaliadores utilizaram as escalas de pontuação para cada tarefa com o intuito de exemplificar como avaliadores portadores de tendências que afetam as pontuações podem ser identificados. Os resultados detalhados das medidas de cada avaliador são apresentados na Tabela 3 e são organizados de acordo com a severidade dos avaliadores em ordem crescente. Os índices expostos nesta Tabela serão utilizados nas análises no decorrer de todo o trabalho.

A primeira coluna da Tabela 3 traz a identificação de cada avaliador. A segunda coluna, o número de pontuações elaboradas por cada um dos avaliadores do grupo, número que variou entre 175 e 7.927 tarefas. Essa variação se deve a fatores como ritmo de trabalho, horários diferenciados de cada avaliador e desligamentos durante o processo de pontuação. A terceira coluna expõe as medidas da severidade dos avaliadores. Entre eles, está o avaliador mais complacente (n. 10) e o mais severo do grupo (n. 1). A quarta coluna informa a precisão com que a medida da severidade do avaliador foi estimada. Quanto maior o número de pontuações em que as medidas são baseadas, maior a precisão. Desse modo, o avaliador de

Tabela 3. Medidas dos avaliadores.

Avaliador	N.pontuação	Severidade	Precisão	MQ-*infit*	MQ-*outfit*	Média obs	Média justa
10	175	-0,99	0,12	1,40	1,35	2,81	2,83
2	7.927	-0,91	0,02	0,83	0,81	3,13	3,27
8	4.953	-0,74	0,02	0,87	0,85	3,02	3,09
4	7.241	-0,67	0,02	0,97	0,98	2,79	2,95
9	6.022	-0,50	0,02	0,92	0,92	2,75	2,80
6	3.061	-0,08	0,03	1,16	1,18	2,97	3,08
3	6.481	-0,04	0,02	1,17	1,17	2,63	2,81
5	6.802	1,09	0,02	0,89	0,89	2,65	2,77
7	5.938	1,24	0,02	1,00	0,99	2,71	2,76
1	6.858	1,60	0,02	1,03	1,07	2,55	2,74

Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).

número 10 pontuou apenas 175 tarefas, a medida de sua severidade, -0,99 *logito*, foi estimada com a menor precisão do grupo, 0,12 *logito*. Os outros avaliadores, com números de pontuação superiores a 3.000, obtiveram medidas com precisão em torno de 0,02 *logito*.

As estatísticas de ajuste, MQ-*infit* e MQ-*outfit,* indicam o grau com que as classificações observadas estão de acordo com as classificações esperadas geradas pelo modelo MFR. As medidas MQ-*infit e* MQ-*outfit*, expostas nas colunas 5 e 6, fornecem uma estimativa da consistência com que cada avaliador, em particular, usa a escala de avaliação para examinandos e itens, resultando em uma medida sensível às classificações não esperadas.

Todos os avaliadores do grupo tiveram índices MQ no intervalo entre 0,5 e 1,5 produtivo para a construção das medidas, segundo orientações de Wright e Linacre (1994). Quando há muitas medidas MQ fora desse intervalo, o sistema pode mostrar-se instável, degradando os resultados, tornando-os não confiáveis.

O fato de o avaliador possuir medidas de ajuste muito diferentes do valor esperado (1,0 *logito*) pode indicar a presença de alguma tendência, mas, para o diagnóstico exato, são necessárias outras análises (MYFORD; WOLFE, 2004; ECKES, 2011), algumas das quais abordadas nas seções seguintes.

Como o grupo de avaliadores é pequeno, e todos eles possuem as medidas MQ razoavelmente próximas do valor esperado (1,0 *logito*), serão analisados, a título de exemplo, avaliadores com pequenas variações dessas medidas.

As colunas 7 e 8 trazem, respectivamente, a média observada e a média justa. Tais índices serão utilizados para estabelecer o efeito de tendência de severidade e complacência na próxima seção.

4.2.1 Efeito de tendência de severidade e complacência

Primeiramente, fazendo uma análise visual do mapa das variáveis (Figura), verifica-se que os avaliadores estão distribuídos de acordo com o seu grau de severidade no intervalo entre -1,0 e 1,6 *logitos*, aproximadamente.

Na Tabela 3, desconsiderando-se as medidas do avaliador de número 10, por ele ter pontuado poucas tarefas, o avaliador mais complacente do grupo, o de número 2, obteve medida -0,91, e o mais severo, o de número 1, obteve medida de 1,60, ambos com precisão 0,02 logito.

A comparação entre os níveis médios de severidade dos avaliadores pode não ser suficiente para determinar se um avaliador é mais severo ou complacente que o outro, principalmente se todos os avaliadores não pontuam o teste de todos os examinandos. Nesse caso, é difícil determinar se o avaliador é mais severo, ou se os examinandos, cujos testes ele pontuou, eram menos habilidosos e, por isso, as notas atribuídas são mais baixas e vice-versa. A média justa (coluna 8) para um avaliador, ajusta a média observada (coluna 7) para a diferença entre os níveis de proficiência da amostra de examinandos para todos os avaliadores. As médias justas separam a severidade do avaliador da proficiência do examinando.

Observa-se, na oitava coluna da Tabela 3, que a diferença entre as médias justas do avaliador mais severo e as do mais complacente é de 3,27-2,74 = 0,53. Isso sugere que o avaliador mais severo atribuiu pontuações, em média, 0,53 pontos menores que o avaliador mais complacente. Em uma escala com categorias de 0 a 6, essa diferença não foi muito acentuada. Desse modo, os avaliadores da equipe não apresentam tendência de severidade ou de complacência.

4.2.2 Efeito de tendência central

De acordo com Myford e Wolfe (2004), a tendência central frequentemente está associada a medidas de ajuste menores que 1, no entanto, algumas vezes, esses índices, para o avaliador que exibe essa tendência, poderão ser maiores que 1. Estes autores sugerem que sejam examinados os vetores com as pontuações dos avaliadores cujas medidas de ajuste estão muito acima ou muito abaixo dos valores esperados, antes de concluir que eles estão exibindo um efeito de tendência central.

Na Tabela 4 são expostos os vetores com as pontuações para cada categoria atribuídas por dois avaliadores que possuem os índices de ajuste um pouco diferentes de 1, os de números 2 e 3.

A primeira coluna traz a lista de categorias de pontuação; a segunda coluna contém o total de pontuação para cada uma das categorias estabelecidas pelo avaliador e a respectiva porcentagem. Observa-se que o avaliador de número 2 atribuiu 84% de suas pontuações às categorias 3 e 4, enquanto o avaliador de número 3 utilizou as categorias 2 e 3, 83% das vezes. A falta de variação da pontuação, atribuída aos desempenhos avaliados por toda a escala de classificação, sugere um efeito de tendência central para esses avaliadores.

As locações ou limiares das categorias também podem auxiliar na confirmação de tendência central para o avaliador. Trata-se dos pontos nos quais a probabilidade de atribuir pontuações às categorias adjacentes é igual (LINACRE, 1989; LINACRE; WRIGHT, 2002). As colunas 6 e 7 da Tabela 4 exibem as locações das categorias e as distâncias entre duas locações consecutivas.

Tabela 4. Estatísticas do uso das categorias.

Categoria	Contagem Categoria	Média Observada	Média Esperada	MQ-*outfit*	Locação	|Diferença|
			Avaliador número 2			
0	149 (2%)	-3,13	-2,92	0,70		
1	254 (3%)	-2,68	-2,47	0,60	-3,33	
2	900 (11%)	-1,93	-1,74	0,60	-3,41	0,08
3	3.764 (48%)	-0,31	-0,33	0,80	-2,54	0,87
4	2.835 (36%)	2,05	1,99	0,90	1,03	3,57
5	25 (0%)	4,76	5,06	0,90	8,25	7,22
6	0 (0%)					
			Avaliador número 3			
0	262 (4%)	-3,41	-3,32	0,80		
1	298 (5%)	-2,55	-2,75	1,30	-3,24	
2	1.604 (25%)	-1,62	-1,81	1,30	-4,01	0,77
3	3.740 (58%)	-0,18	-0,16	1,10	-1,89	2,12
4	566 (9%)	1,30	1,71	1,20	2,69	4,58
5	11 (0%)	2,28	3,27	1,60	6,45	3,76
6	0 (0%)					

Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).

Os limiares das categorias serão dispersos na escala de classificação para o avaliador que apresenta tendência central, com pouca utilização das categorias nos extremos da escala (MYFORD; WOLFE, 2004). Elas são bastante afastadas para esses avaliadores. Além disso, não utilizam demasiadamente as categorias 0, 1 e 6.

Algumas vezes, pode haver, também, a inversão na ordem das categorias para avaliadores portadores de tendência central, isto é, os limiares não aumentam monotonicamente. Esse fato pode ser constatado para os dois avaliadores que possuem uma inversão dos limiares das categorias 1 e 2.

Alguns índices expostos na Tabela 4 também serão utilizados nas análises sobre o efeito de aleatoriedade na próxima seção.

4.2.3 Efeito de aleatoriedade

Os avaliadores que mostram efeito de aleatoriedade em suas classificações normalmente possuem as medidas de ajuste maiores que 1, sugerindo que eles não foram capazes de diferenciar os desempenhos dos examinandos ao longo da escala de classificação, atribuindo pontuações aparentemente aleatórias.

Para eliminar a possibilidade de diagnóstico errado para o efeito de aleatoriedade, uma vez que outras tendências também exibem as estatísticas de ajuste maiores que 1, deve-se comparar as correlações bisserial desses avaliadores com as dos outros.

Observe novamente, na Tabela 4, os índices MQ-*outfit* do avaliador de número 3 (coluna 5). Os índices de ajuste são maiores que 1 e a correlação ponto bisserial desse avaliador é de 0,61, um pouco menor do que as dos outros avaliadores do grupo que variam entre 0,67 e 0,78. Isto ocorre porque as pontuações deste avaliador tende a ser em uma ordem diferente das pontuações dos demais avaliadores.

Na Tabela 4, pode-se verificar como esse avaliador utilizou cada uma das categorias. Observe que há um desacordo entre as médias observadas e esperadas principalmente nas categorias mais altas da escala (colunas 3 e 4), indicando alguma aleatoriedade nessas pontuações. Provavelmente, esse avaliador tem algum problema em avaliar nessas categorias da escala.

4.2.4 Efeito de halo

Para detectar os avaliadores portadores de tendência de halo, deve-se primeiramente procurar por índices de ajuste significativamente diferentes de 1,0 *logito*, para então inspecionar as pontuações observadas desses avaliadores a fim de determinar

a porcentagem na qual as pontuações são praticamente as mesmas. Isso deve ser feito para cada uma das categorias ao longo de todas as pontuações efetivadas.

Myford e Wolfe (2004) também sugerem uma análise dos vieses da interação entre os avaliadores *versus* tarefas. Essa análise é fornecida pelo programa *Facets* e indica o grau com que as pontuações elaboradas por um avaliador diferem das expectativas produzidas pelo modelo.

A maioria das medidas de viés é pequena e, estatisticamente, insignificante. Desse modo, as medidas de vieses são relatadas pelo programa *Facets* quando o índice estatística-t em valor absoluto for maior do que 2, identificando, assim, os avaliadores que apresentam alguma inconsistência em suas pontuações. As análises são as seguintes: se a estatística-t é maior do que 2, o avaliador foi mais severo que o esperado; se a estatística-t é menor que -2, o avaliador foi mais complacente que o esperado. Entretanto, para determinar se esse desajuste é resultado de tendência a efeito de halo, é necessário examinar também as pontuações observadas e esperadas para os respectivos avaliadores (LINACRE, 2014a; MYFORD; WOLFE, 2004).

A Tabela 5 traz as medidas dos vieses das interações avaliadores *versus* tarefas para o avaliador de número 7, que apresenta índices estatística-t significativamente maiores que 2 em valores absolutos (coluna 10).

Comparando-se as médias observadas e esperadas, mostradas nas colunas 3 e 4, nota-se que esse avaliador atribuiu pontuações mais baixas que as esperadas para as tarefas 2 e 3 e pontuações mais elevadas que as esperadas para a tarefa 1. As diferenças entre as médias observadas e esperadas, divididas pelo total das pontuações atribuídas (cont.) são calculadas na coluna 7. A segunda coluna traz as medidas da dificuldade das tarefas. A tarefa de número 1 foi considerada a mais difícil e, por isso, recebeu pontuações mais baixas dos avaliadores, enquanto a de número 3 foi considerada a mais fácil, por isso, recebeu pontuações mais elevadas. Esse avaliador, no entanto, atribuiu pontuações mais elevadas

Tabela 5. Vieses: Avaliadores versus Tarefas. Avaliador número 7.

Tarefa	Medida	Média Observada	Média Esperada	Cont.	G.L	(Obs-Esp)/ Cont.	Viés	Precisão	estat. t
1	0,22	5508	5231,36	1995	1994	0,14	0,34	0,04	9,54
2	0,01	5273	5355,31	1979	1978	-0,04	-0,1	0,04	-2,93
3	-0,23	5300	5494,35	1958	1957	-0,10	-0,26	0,04	-7,17

Fonte: Elaborada pelas autoras com base nos dados da pesquisa (2016).

para a tarefa de número 1, contrariando as pontuações médias dos outros avaliadores, assim como pontuações mais baixas para as tarefas de números 2 e 3, enquanto os outros avaliadores, em média, atribuíram pontuações mais elevadas. Esse avaliador tende a atribuir pontuações de modo diferente dos outros avaliadores para as mesmas características. Isso sugere que ele possui tendência a efeito de halo.

5 Discussão

Os sistemas de avaliações em larga escala devem proporcionar inferências válidas, confiáveis e justas em relação à medida obtida da habilidade dos participantes. Seria ideal que a classificação dos participantes não dependesse de cada avaliador que julgou o desempenho dos indivíduos. Apesar do objetivo deste trabalho não ter sido o de fazer uma análise profunda sobre a qualidade da pontuação das provas, pôde-se constatar que esta foi feita com bons índices de confiabilidade. Os exemplos de comportamentos tendenciosos relatados não depõem contra a boa qualidade dos resultados, uma vez que estes problemas não ocorreram de forma severa e também porque, neste estudo, estão incluídas também as notas discrepantes.

O modelo MFR pode possibilitar o acesso a informações específicas sobre cada examinando, cada avaliador, cada item e ainda permite análises da forma como os avaliadores utilizaram a escala de classificação e os critérios utilizados para o julgamento das tarefas, além do estudo das influências de outros fatores incluídos nas análises.

As vantagens na utilização do modelo MFR nas avaliações com itens abertos são inúmeras, e muitos estudos estão sendo desenvolvidos com foco nos mais variados elementos. O modelo MFR proporciona análises mediante um conjunto diverso de índices quantitativos, ilustrações gráficas, tabelas, entre outros, auxiliando na determinação de evidências para o monitoramento da qualidade das avaliações.

O grande número de candidatos nas avaliações em larga escala e o fato de que a diferença entre a pontuação de um indivíduo que alcançou a aprovação e de outro que não obteve o mesmo sucesso pode ser muito pequena (BARROS, 2014) justificam a necessidade de busca por precisão com a qual o teste pode estimar a capacidade das pessoas e acarretar, assim, uma classificação verdadeira e confiável.

Desse modo, aumenta-se a exigência por instrumentos de medida apropriados e pelo monitoramento constante da qualidade dessas avaliações.

The use of Many-Facet Rasch Model to explore the raters influence in open questions exams

Abstract

This paper analyzes the quality of scores in open questions exams through Many-Facet Rasch (MFR) model. Scores assigned to essays compiled by the participants in the college entrance examination for Londrina University, in 2015, were used. The MFR Model can provide both group and individual level studies, enabling the determination of raters with biased behavior, which are known to cause significant errors in the written assignment scores. Analyzes at group level showed that the evaluation was efficient and the data, in general, are suitable for Rasch models measurement and through the analyzes at individual level it was possible to find raters who scored differently from the average scores of others reviewers. The MFR model proved to be an appropriate and effective tool for monitoring the quality of scores assigned to writing tasks.

Keywords: *Many-Facet Rasch Model. Open questions. Essay Tests. Large scale evaluation. Rater tendency.*

La utilización del modelo multifacético de Rasch en el análisis de las influencias de loa calificadores sobre las evaluaciones con ítems abiertos

Resumen

Este estudio analiza la calidad de las puntuaciones en los exámenes con ítems abiertos a través del modelo multifacético de Rasch (MFR). Para ello se utilizan las puntuaciones atribuidas a las redacciones de participantes de la prueba de acceso de 2015 de la Universidad Estatal de Londrina. El modelo MFR puede proporcionar estudios, tanto a nivel de grupo como en el plano individual, permitiendo la identificación de evaluadores portadores de comportamientos tendenciosos, conocidos por causar errores significativos en las puntuaciones de trabajos escritos. Los análisis a nivel de grupo mostraron que la evaluación fue eficiente y que los datos, en general, cumplen con las expectativas de medición de los modelos de Rasch, y por medio de los análisis a nivel individual fue posible encontrar calificadores que obtuvieron una puntuación diferente a las puntuaciones medias de otros colaboradores. El modelo MFR demostró ser una herramienta adecuada y eficaz para el control de la calidad de las puntuaciones asignadas a las tareas de escritura.

Palabras clave: *Modelo multifacetas de Rasch. Ítems abiertos. Pruebas de redacción. Evaluación a gran escala. Tendencias del evaluador.*

Referências

BARROS, A. S. X. Vestibular e Enem: um debate contemporâneo. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 22, n. 85, p. 1057-90, dez. 2014. http://dx.doi.org/10.1590/S0104-40362014000400009

ECKES, T. *Introduction to Many-Facet Rasch Measurement*: analyzing and evaluating rater-mediated assessment. Frankfurt: Peter Lang, 2011.

ENGELHARD JR., G.; MYFORD, C. M. *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model.* New York: College Entrance Examination Board, 2003.

ENGELHARD JR, G.; WIND, S. A. *Rating quality studies using rasch measurement theory*. Educational Testing Service (ETS). [Princeton, NJ]: College Board: 2013.

ILHAN M. A comparison of the results of many-facet rasch analyses based on crossed and judge pair designs. *Educational Sciences: Theory & Practice*, v. 16, p. 579-601, 2016. https://doi.org/10.12738/estp.2016.2.0390

JONSSON, A.; SVINGBY, G. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, v. 2, n. 2, p. 130-44, 2007. https://doi.org/10.1016/j.edurev.2007.05.002

LINACRE, J. M. A user's guide to FACETS [computer program manual 3.71.4]. Chicago: *MESA Press*, 2014b.

_____. Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: *Winsteps.com*, jan. 2014a.

_____. *Many-facet Rasch measurement*. Chicago: MESA Press, 1989.

LINACRE, J. M.; WRIGHT, B. D. Construction of measures from many-facet data. *Journal of Applied Measurement*, v. 3, n. 4, p. 484-509, 2002.

LONG, H.; PANG, W. Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, v. 15, p. 13-25, 2015. https://doi.org/10.1016/j.tsc.2014.10.004

MYFORD, C. M.; WOLFE, E. W. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, v. 5, n. 2, p. 189-227, 2004.

SEBOX, S. S. et al. Examiners and content and site: Oh My! A national organization's investigation of score variation in large-scale performance assessments. *Advances in Health Sciences Education Theory and Practice*, v. 20, n. 3, p. 581-94, 2015. https://doi.org/10.1007/s10459-014-9547-z

STEMLER, S. E. A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, v. 9, n. 4, p. 1-11, 2004.

TOFFOLI, S. F. L. *Avaliações em larga escala com itens de respostas construídas no contexto do modelo Multifacetas de Rasch*. 2015. 313 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2015.

TOFFOLI, S. F. L.; ANDRADE, D. F.; BORNIA, A. C. Evaluation of open items using the many-facet Rasch model. *Journal of Applied Statistics*, v. 43, n. 2, p. 299-316, 2015. https://doi.org/10.1080/02664763.2015.1049938

UNIVERSIDADE ESTADUAL DE LONDRINA - UEL. Coordenadoria de Processos Seletivos. *Manual do candidato*, 2014. Disponível em: <http://www.cops.uel.br/vestibular/2015>. Acesso em: 14 set. 2016.

_____. Coordenadoria de Processos Seletivos. *Processo Seletivo Vestibular 2015*: provas e gabaritos da 2ª fase, 2015. Disponível em: <http://www.cops.uel.br/vestibular/2015/provas-gabaritos/fase-2/>. Acesso em: 14 set. 2016.

WRIGHT, B. D.; LINACRE, J. M. Reasonable mean-square fit values. *Rasch Measurement Transactions*, v. 8, n.3, p. 370, 1994.

WRIGHT, B. D.; STONE, M. *Measurement essentials*. 2nd ed. Wilmington, Delaware: Wide Range, 1999.

Informações das autoras

Sônia Ferreira Lopes Toffoli: Doutora em Engenharia de Produção pela Universidade Federal de Santa Catarina. Professora adjunta da Universidade Estadual de Londrina (UEL), Departamento de Matemática. Contato: sonialopes@uel.br

Cristina Valeria Bulhões Simon: Doutora em Estudos da Linguagem pela Universidade Estadual de Londrina. Professora adjunta da Universidade Estadual de Londrina (UEL), Departamento de Letras Vernáculas e Clássicas. Contato: cristinavbsimon@gmail.com