UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Sonia Ferreira Lopes Toffoli

AVALIAÇÕES EM LARGA ESCALA COM ITENS DE RESPOSTAS CONSTRUÍDAS NO CONTEXTO DO MODELO MULTIFACETAS DE RASCH

Florianópolis

2015

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Sonia Ferreira Lopes Toffoli

AVALIAÇÕES EM LARGA ESCALA COM ITENS DE RESPOSTAS CONSTRUÍDAS NO CONTEXTO DO MODELO MULTIFACETAS DE RASCH

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina para a obtenção do grau de Doutora em Engenharia de Produção.
Orientador: Prof. Dr. Dalton Francisco de Andrade
Coorientador: Prof. Dr. Antonio Cezar Bornia

Florianópolis

2015

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Toffoli, Sônia Ferreira Lopes
Avaliações em larga escala com itens de respostas construídas no contexto do modelo multifacetas de Rasch / Sônia Ferreira Lopes Toffoli ; orientador, Dalton Francisco de Andrade ; coorientador, Antonio Cezar Bornia. - Florianópolis, SC, 2015.
315 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção.

Inclui referências

1. Engenharia de Produção. 2. Avaliações com itens de respostas construídas. 3. Avaliações em larga escala. 4. Modelo multifacetas de Rasch. 5. Teoria de Resposta ao Item. I. de Andrade, Dalton Francisco. II. Bornia, Antonio Cezar. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Produção. IV. Título.

Sonia Ferreira Lopes Toffoli

AVALIAÇÕES EM LARGA ESCALA COM ITENS DE RESPOSTAS CONSTRUÍDAS NO CONTEXTO DO MODELO MULTIFACETAS DE RASCH

Esta Tese foi julgada adequada para obtenção do Título de Doutora em Engenharia de Produção e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina.
Florianópolis, 10/04/2015.

Lucila Campos, Dra.
Coordenadora do Programa

Banca Examinadora:

Dalton Francisco de Andrade, Dr.
Orientador
Universidade Federal de Santa Catarina (UFSC)

Antonio Cezar Bornia, Dr.
Coorientador
Universidade Federal de Santa Catarina (UFSC)

Adriano Ferreti Borgatto, Dr.
Membro interno
Universidade Federal de Santa Catarina (UFSC)

Antônio Sérgio Coelho, Dr.
Membro interno
Universidade Federal de Santa Catarina (UFSC)

Carlos Henrique Sancineto da Silva Nunes, Dr.
Membro interno
Universidade Federal de Santa Catarina (UFSC)

Eduardo Carvalho Sousa, Dr.
Examinador externo
Instituto Nacional de Estudos e Pesquisas (INEP)

Gladys Plens de Quevedo Pereira de Camargo, Dra.
Examinador externo
Universidade de Brasília (UnB)

AGRADECIMENTOS

A meus pais e irmãos: Laércio, Lúcia, Vânia, Ênio, Vera e Luciano, pelo apoio, paciência e compreensão pelo tempo que roubei ao nosso convívio.

À minha prima-irmã Maria Márcia, que nos momentos certos e incertos sempre esta por perto.

Ao meu marido, Cleber Toffoli, pelo amor, carinho, compreensão e por todo o apoio dado no decorrer do trabalho.

Ao professor orientador Dalton Francisco de Andrade, pela acolhida e orientação segura durante toda a caminhada.

Ao professor coorientador Antonio Cezar Bornia, pelo apoio, ensinamentos e contribuições sobre este trabalho.

À Professora Gladys Quevedo-Camargo, pela amizade, leitura e críticas pertinentes.

À Ivone Alves de Lima, pela disponibilidade, incentivo e competência na leitura e revisão dos textos.

Aos colegas do curso, pela amizade, discussões e contribuições.

A todos que, direta ou indiretamente contribuíram para a concretização deste trabalho.

RESUMO

Esta tese apresenta um estudo sobre as avaliações com itens de respostas construídas em larga escala no contexto do modelo multifacetas de Rasch (LINACRE, 1989 *apud* LINACRE, 1994). Essas avaliações necessitam de avaliadores para julgar o desempenho das pessoas quanto à habilidade que está sendo medida por meio do teste. Entre as avaliações com itens de respostas construídas mais utilizadas no âmbito educacional e de seleção estão as provas das diversas disciplinas do Ensino Médio, as provas de redação do ENEM e dos concursos vestibulares e as provas com itens abertos de concursos para provimento de vagas de trabalho.

Os resultados das avaliações com itens de respostas construídas não dependem apenas do nível de habilidade dos examinandos quanto ao construto avaliado e da dificuldade das tarefas, dependem também da severidade dos avaliadores que julgam os desempenhos e da estrutura da escala de classificação. Um dos principais problemas nessas avaliações é a pontuação de um mesmo desempenho com graus diferentes de severidade. Quando existem vários avaliadores, o ideal é que todos atribuam exatamente a mesma pontuação para os mesmos desempenhos observados, essa é a condição principal para se ter confiabilidade de pontuação. Entretanto, são muitos os fatores que podem causar variabilidade nessas pontuações.

O modelo multifacetas de Rasch vem sendo cada vez mais utilizado para aferir a qualidade das avaliações com itens de respostas construídas, por permitir a inclusão de outras variáveis aos sistemas avaliativos, além da capacidade dos indivíduos e da dificuldade das tarefas.

Algumas dessas variáveis consistem em importantes fontes geradoras de vieses nos processos avaliativos. Como exemplos têm-se as características pessoais dos avaliadores, as diferenças entre a severidade dos avaliadores, as tendências dos avaliadores em julgamentos sistemáticos, as diferenças entre as dificuldades de tarefas distintas e a variação quanto ao entendimento e utilização das categorias da escala de classificação por parte dos avaliadores. O modelo multifacetas de Rasch permite a inclusão de cada variável que pode interferir na avaliação, além de possibilitar análises para os efeitos causados por cada elemento que faz parte da avaliação individualmente, o que torna a utilização desse modelo muito vantajosa.

O objetivo deste estudo é estabelecer como o modelo multifacetas de Rasch pode contribuir para a determinação da qualidade das avaliações com itens de respostas construídas. A abordagem utilizada pelo modelo multifacetas de Rasch proporciona análises sobre a qualidade das medidas relacionadas

aos examinandos, aos avaliadores, às tarefas, aos itens e às escalas de classificação utilizadas para a pontuação das tarefas.

Este trabalho também apresenta uma aplicação do modelo multifacetas de Rasch aos dados provenientes de uma avaliação real, na qual estabelece as principais análises sobre a qualidade dessa avaliação.

Palavras-chave: Modelo multifacetas de Rasch, Avaliação com itens abertos, Avaliação em larga escala, Confiabilidade de pontuação, Severidade do avaliador, Escala de classificação.

ABSTRACT

This thesis presents a study about the large-scale construct-response item evaluations in the context of the many-facet Rasch model (LINACRE, 1989 *apud* LINACRE 1994). These evaluations require raters in order to judge the performance of the people regarding the ability that is being measured through test.

Among the evaluations with constructed-responses items most frequently used in the educational and hiring ambit are those with open questions of the disciplines of the High School, the writing test of the Brazilian High School National Exam and of the university entrance exams and the tests with open questions of contests.

The results of the construct-response item evaluations do not depend only on the ability level of the examinants regarding the evaluated construct and the difficulty of the tasks; they depend also on the severity of the raters that judge the performance and the structure of the classification scale. One of the main problems of these evaluations is the rating of a same performance with different severity degrees. When there are many raters, it would be the ideal if all would give exactly the same rating for the same performances observed, this is the main condition in order to have reliability of rating. However, many are the factors that can cause variability in these ratings.

The many-facet Rach model have been even more used to check the quality of the construct-response item evaluations, since it allows the inclusion of other variables to the evaluating systems, besides the capabilities of the individuals and the difficulty of the tasks. Some of these variables consists of important sources generator of biases in the evaluating processes. As examples are the personal characteristics of the raters, the differences between the severity of the raters, the tendencies of the raters in systematic judgements, the differences between the difficulties of the distinct tasks and the variation regarding the understanding and use of the categories of the classification scale by the raters. The many-facet Rach model allows the inclusion of each variable that can interfere in the evaluation besides allowing analyzes for the effects caused by each element that is individually part of the evaluation, which makes the use of the many-facet Rach model very advantageous.

The objective of this study is to establish how the many-facet Rach model can contribute to the determination of the quality of the evaluations with construct-response items. The approach used by the many-facet Rach model provides analyzes on the quality of the measure related to the examinees, to the the raters, to the tasks, to the questions and to the classification

scales used for the rating of the tasks.

This work also presents an application of multi-faceted Rasch model to data from a real assessment, which establishes the main analyzes of the quality of the evaluation.

Keywords: Many-Facet Rach Model, Construct-response assessment, Large-scale assessment, Rating reliability, Rater severity, Rating Scales.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Professores e pesquisadores estão constantemente em busca de mecanismos para avaliar a capacidade de escrita de seus alunos quando estes são submetidos a escrever sobre conteúdos específicos ou chamados a redigir algum texto. Profissionais da avaliação educacional consideram a avaliação escrita uma área problemática devido ao grande número de variáveis envolvidas. Tratando-se de avaliação em larga escala, o problema é ainda maior. Apesar de as pesquisas educacionais terem avançado em discussões gerais sobre o significado dessas avaliações, os estudos que explicitam claramente os conceitos envolvidos na elaboração dos instrumentos de avaliação e os critérios de pontuação para as tarefas estabelecidas aos alunos ainda são assuntos que geram controvérsias, cercados de iniciativas intuitivas baseadas nas experiências dos professores e avaliadores (BROWN; GLASSWELL; HARLAND, 2004; BONAMINO; COSCARELLI; FRANCO, 2002).

Os processos avaliativos possuem diferentes objetivos: classificação de candidatos com a finalidade de aprovação para um emprego ou vaga de escola, determinação do grau de habilidade para alguma atividade específica, avaliação do progresso ou do alastramento de uma doença, avaliação do desenvolvimento físico e psicológico de uma criança, avaliação da capacidade de aprendizagem, avaliação do desempenho escolar, avaliação do desempenho logístico, avaliação da qualidade de algum produto industrial ou serviço, entre outros inúmeros exemplos. As informações provenientes das avaliações auxiliam as decisões, sejam elas pessoais ou na esfera pública. É por essa razão que as avaliações devem ser confiáveis.

As avaliações em larga escala exercem forte influência sobre as políticas educacionais e sobre os currículos nos diversos níveis de ensino em todo o mundo. Portanto, é evidente a importância de examinar as diversas variáveis envolvidas na construção, aplicação e pontuação desses exames (BEHIZADEH; ENGELHARD, 2011; SCARAMUCCI, 2011; BECK; JEFFERY, 2007).

Existe uma gama de diferentes instrumentos de avaliação utilizados para os mais diversos objetivos. Quando se trata de avaliações educacionais em larga escala, destacam-se basicamente duas categorias: as avaliações de itens de respostas objetivas e as de itens de respostas abertas, comumente denominados também de itens de respostas descritivas ou subjetivas.

Os itens de respostas objetivas limitam ao máximo as opções dos participantes, que são obrigados a fornecer uma resposta altamente definida, por exemplo, a uma operação matemática ou a palavra que falta para completar uma frase. Os itens de múltipla escolha também se enquadram nas avaliações com itens de respostas objetivas, mas, nesse caso, há uma lista de opções.

As avaliações com itens de respostas abertas permitem certa liberdade ao participante na elaboração da sua resposta. Esses itens podem ser de respostas curtas, as quais devem ser sucintas e específicas para cada pergunta, ou de respostas estendidas, nas quais os candidatos devem desenvolver a tarefa determinada com base nas informações fornecidas e respeitando alguns critérios estabelecidos. Itens desse tipo permitem ao candidato a liberdade de construir as respostas de forma original. As redações dos vestibulares e de outros concursos são avaliações com itens abertos de respostas estendidas, e as provas de vestibulares que utilizam questões descritivas muitas vezes se encaixam nessa categoria. Neste trabalho, para maior simplicidade e uniformidade na denominação, estas serão tratadas como avaliações com itens de respostas abertas ou simplesmente avaliações com itens abertos. Algumas vezes serão referidas apenas por avaliações escritas.

As avaliações que necessitam do julgamento de avaliadores quanto ao desempenho na execução de alguma tarefa, como as avaliações com itens abertos, são frequentemente designadas na literatura por avaliações de desempenho (HAERTEL; LINN, 1996; LINACRE; WRIGHT, 2002). Nessa categoria também se encontram as entrevistas, as avaliações orais, algumas competições esportivas, entre outras.

Os testes com itens abertos são muito utilizados para avaliar a capacidade de expressão escrita das pessoas. Podem abranger conteúdos relacionados apenas com a linguagem ou também assuntos de outras áreas, como conhecimentos gerais ou conteúdos específicos que fazem parte de alguma disciplina. Entretanto, independentemente do conteúdo abordado, por necessitar de habilidades necessárias para a comunicação escrita, tais avaliações são também denominadas de avaliações escritas.

Os resultados das avaliações com testes de itens abertos não dependem apenas do nível de habilidade dos examinandos quanto ao construto avaliado e da dificuldade das tarefas dependem também da severidade dos avaliadores que julgam os desempenhos e da estrutura da escala de classificação. Por esse motivo, não são possíveis pontuações completamente objetivas.

As avaliações com testes de itens abertos têm uma longa história, uma vez que o modo consagrado através dos tempos de descobrir se uma pessoa pode ou não executar uma tarefa é fazer com que ela tente executar essa

tarefa. Os testes objetivos tiveram a maior parte de seu desenvolvimento a partir de 1950 e ganharam destaque por oferecerem uma série de vantagens práticas, especialmente nas avaliações em larga escala, nas quais o número de indivíduos avaliados é grande (YANCEY, 1999). Os indicadores fornecidos por essas avaliações, entretanto, tendem a ser resultados educacionais indiretos e parciais. Há muitas situações em que uma avaliação mais direta do desempenho é desejável (JONSSON; SVINGBY, 2007; MESSICK, 1996). O crescente reconhecimento das limitações dos testes objetivos e a preocupação com o impacto das avaliações nos sistemas educacionais e na vida das pessoas têm gerado um aumento no interesse pelas avaliações com testes de itens abertos (KANE; COOKS; COHEN, 1999).

Nessas avaliações, são muitos os fatores que podem afetar a medida do desempenho das pessoas ao executar a tarefa determinada no teste. Em primeiro lugar, está a habilidade do examinando, mas a pontuação que ele receberá no exame não depende apenas da sua capacidade ou do conhecimento sobre o construto que está sendo medido, depende também da severidade do avaliador, da dificuldade das tarefas, do formato da questão, do tema abordado, dos critérios e da escala de pontuação e de outras variáveis que podem interferir em cada evento de avaliação em particular.

Esses e outros fatores são frequentemente constatados em estudos relacionados com avaliações com itens abertos, principalmente nas avaliações da linguagem de primeira e segunda língua. Alguns exemplos podem ser obtidos nos trabalhos de Huang (2012), Rezaei e Lovorn (2010), Gyagenda e Engelhard (2009), Jonsson e Svigby (2007), Sudweeks, Reeve e Bradshaw (2005) e Weigle (1999).

Atualmente, no Brasil e também em outros países, são vários os processos de seleção ou de avaliação em larga escala que utilizam avaliações com itens abertos. Na elaboração dessas avaliações, os maiores desafios referem-se à concepção dos itens, à atribuição de pontuação precisa e à comparabilidade entre testes distintos.

Nos últimos anos, o crescimento de pesquisas relacionadas com as avaliações educacionais escritas é notório, e um dos fatores que estimulam esse crescimento é a motivação político-econômica. O uso das avaliações como instrumento político tem ocorrido em muitos países, como Estados Unidos, Austrália, Nova Zelândia, Canadá, Reino Unido, Brasil e Chile, (SCARAMUCCI, 2011; DE SOUZA; GOUVEIA, 2011; HAMP-LYONS, 2011; DE CASTRO, 2009). Outro motivo para tal expansão, particularmente nos países de língua inglesa, relaciona-se com o aumento no número de estudantes universitários estrangeiros, de língua nativa não inglesa, exigindo

o aumento de testes de inglês como segunda língua, nos quais a componente escrita é vista como essencial (HAMP-LYONS, 2011). Nas universidades norte-americanas, também são usuais os testes de colocação ou posicionamento, que se destinam a avaliar a capacidade de redação de estudantes do primeiro ano dos cursos e auxiliam na identificação dos alunos com necessidade de algum apoio acadêmico adicional (RAMINENI, 2012).

No Brasil, o número de pesquisas na área da avaliação em larga escala ainda pode ser considerado limitado, principalmente no que diz respeito a estudos sobre a qualidade dos instrumentos das principais avaliações nacionais, como, por exemplo, ENEM (Exame Nacional do Ensino Médio), ENADE (Exame Nacional de Desempenho de Estudantes), ANA (Avaliação Nacional da Alfabetização), ENCCEJA (Exame Nacional para Certificação de Competências de Jovens e Adultos), SAEB (Sistema de Avaliação da Educação Básica). O mesmo ocorre com os concursos vestibulares. As universidades não costumam divulgar estudos relacionados com a qualidade de suas provas nos exames de acesso ao Ensino Superior, nem mesmo informações como dados, gráficos, estatísticas, entre outros. Muitos estudos existentes sobre as avaliações nacionais são divulgados apenas localmente em veículos pouco expressivos, mas mesmo esses são escassos quando se trata de concurso vestibular (VICENTINI, 2011; SCARAMUCCI, 2004; VIANNA, 2003).

Por outro lado, em relação aos principais testes internacionais, é frequente a veiculação de estudos subsidiados pelos governos, centros de pesquisas ou órgãos provedores dos exames que incentivam pesquisas para melhorar a qualidade de seus instrumentos, como exemplo, o *Scholastic Aptitude Test* (SAT) (KOBRIN; DENG; SHAW, 2011; BECK; JEFFERY, 2007), o *Test of English as a Foreign Language* (TOEFL) (HUANG, 2012; BRELAND; NAJARIAN; MURAKI, 2004), o *National Assessment of Educational Progress* (NAEP) (JEFFERY, 2009), o *Educational Testing Service* (ETS) (ENGELHARD; MYFORD; CELINE, 2000; MYFORD; WOLF, 2000), do *College Board* (ENGELHARD; WIND, 2013; ENGELHARD; MYFORD, 2003) e o *Cambridge ESOL examinations* (JONES; SHAW, 2003).

A avaliação de tarefas escritas, como a de qualquer outra competência, necessita de instrumentos padronizados, válidos, fidedignos, capazes de selecionar de maneira justa, apoiar condutas para a melhoria do ensino ou a organização de programas de intervenção. Os protocolos de uma avaliação devem atender a essas especificações para assegurar a confiança na pontuação.

Métodos estatísticos clássicos são muito utilizados, mas são limitados

para fornecer informações suficientemente detalhadas, principalmente sobre a capacidade em avaliações com testes de itens abertos, por sua complexidade. A Teoria de Resposta ao Item (TRI) está sendo, gradualmente, incorporada aos procedimentos de análise dos dados desses exames por oferecer mais recursos (McNAMARA; KNOCH, 2012; HAMP-LYONS, 2011; BEHIZADEH; ENGELHARD, 2011).

A utilização da Teoria de Resposta ao Item (TRI) para auxiliar a pontuação e classificação dos respondentes ao teste, assim como a de seus avaliadores, pode trazer vantagens e credibilidade ao processo (HAMP-LYONS, 2011), uma vez que a TRI permite comparar os desempenhos de indivíduos, posicionando-os em uma escala comum. Essa possibilidade de comparação é possível mesmo que os indivíduos tenham participado de testes diferentes, proporcionando estudos mais aprofundados e garantindo uma análise melhor dos problemas, subsidiando a tomada da decisão para a adoção de uma política adequada para o seu enfrentamento (TEZZA; BORNIA; ANDRADE, 2011).

No Brasil, o SAEB já utiliza uma escala única referenciada para Língua Portuguesa e Matemática, e as avaliações realizadas pelos diversos estados brasileiros mantêm a mesma matriz de referência do SAEB, garantindo a comparabilidade de resultados entre os anos avaliados por meio da Teoria de Resposta ao Item (TRI) (LIMA *et al.*, 2008; KLEIN *et al.*, 2008; BONAMINO; COSCARELLI; FRANCO, 2002; VIDAL; FARIAS, 2008; BRASIL/MEC). Os testes do SAEB que permitem tais comparações são com itens dicotômicos, isto é, são corrigidos apenas como certos ou errados.

Aliás, a maior parte das avaliações em larga escala no Brasil utiliza itens dicotômicos. Itens politômicos, nos quais as respostas são construídas pelos alunos e as notas são atribuídas com base em uma escala gradual, além de certo ou errado, não são muito frequentes, apesar de serem amplamente utilizados em outros países, como Estados Unidos e Inglaterra. Em consequência, também são raros os exemplos no Brasil de pesquisas sobre avaliações em larga escala envolvendo itens abertos.

Até os dias de hoje no Brasil, a maioria das pesquisas envolvendo as avaliações escritas em larga escala relaciona-se com os processos seletivos para acesso aos cursos superiores das diversas universidades públicas. Essas avaliações efetivamente tiveram início a partir de 1978, com a aprovação do Decreto n°. 79.298 de 1977, exigindo que os candidatos ao vestibular fizessem uma prova ou questão de redação em língua portuguesa (BRASIL, 1997; CASTRO, 2008).

Ribeiro Netto, presidente da Fundação Carlos Chagas na década de

1980, empresa responsável pela organização dos principais exames vestibulares no estado de São Paulo naquela época, destaca em artigo apresentado em seminário sobre o vestibular que a sociedade apregoava a má qualidade da expressão escrita dos estudantes como resultado do emprego exclusivo de testes de múltipla escolha nos concursos vestibulares. Essa modalidade de teste foi utilizada desde 1964. Antes dessa data, as provas eram escritas, orais ou práticas, a critério da instituição. Com a obrigatoriedade de uma questão de redação no vestibular, voltou a ser utilizada no Brasil a modalidade de avaliação escrita (RIBEIRO NETTO, 1985).

A inclusão da redação nos exames do vestibular foi acompanhada da carência de especialistas em medidas educacionais para assessorar o exame e dar suporte às diferentes pesquisas que pudessem avaliar o impacto dessas mudanças e dar mais confiabilidade ao processo. Uma grande dificuldade residia na pontuação da prova, já que a correção da redação tem um caráter subjetivo e depende do julgamento dos avaliadores, podendo comprometer a confiabilidade da nota atribuída aos candidatos (MORAES, 1992). Desse modo, começou a surgir uma série de estudos para testes com questões abertas, principalmente relacionados com o estabelecimento de uma escala para a pontuação e métodos para se obter maior confiabilidade entre as pontuações dos avaliadores (VIANNA, 1978; MORAES, 1992; HOFFMAN, 1988).

Mais tarde, algumas das mais importantes universidades públicas do país passaram a utilizar, além da redação, provas dissertativas das várias disciplinas do núcleo comum obrigatório do Ensino Médio em seus exames de acesso, como é o caso da Universidade Estadual de Campinas (UNICAMP) e das instituições que utilizam os exames elaborados pela Fundação Universitária para o Vestibular (FUVEST). A UNICAMP tem suas provas feitas em duas fases e contou com questões totalmente discursivas desde 1986 até 2010. Em 2011, a primeira fase do exame foi reformulada e passou a contar com 48 questões de múltipla escolha e uma prova de redação. A segunda fase continua composta por itens discursivos (COMVEST, 2010; ABAURRE, 1995). A FUVEST, responsável pela organização dos vestibulares para ingresso em cinco importantes instituições públicas de ensino superior do estado de São Paulo, utiliza, desde sua fundação em 1977, questões objetivas nas provas da primeira fase e questões discursivas nas provas da segunda fase (FUVEST, 2013; PINHO FILHO, 1996).

Atualmente, as avaliações em larga escala que utilizam testes com itens abertos, no Brasil, se resumem às redações dos vestibulares, à redação do ENEM, provas das outras disciplinas de alguns vestibulares, por exemplo, o da FUVEST, o da UNICAMP e o da Universidade Estadual de Londrina

(UEL), que possuem parte de suas provas com questões abertas, e a alguns outros exemplos, como é o caso de algumas edições específicas do SAEB, da ANA que possui alguns itens abertos para avaliar o desempenho quanto à produção da escrita e concursos do setor público e privado para provimento de vagas de trabalho. Apesar de esses exames exercerem uma grande influência na sociedade e na vida das pessoas, pesquisas recentes sobre os itens abertos são ainda mais raras, com a maioria delas na área da linguística aplicada e problemas de ensino/aprendizagem relacionados com a sala de aula.

Uma avaliação brasileira em larga escala de grande importância é o ENEM. Essa avaliação é de responsabilidade do Ministério da Educação com execução do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Foi criado em 1998 com o objetivo principal de avaliar o desempenho dos alunos egressos do Ensino Médio e proporcionar uma avaliação nacional da educação. Em 2009, o ENEM passou por uma reformulação possibilitando também a utilização dos resultados individuais como mecanismo de seleção para o acesso à Educação Superior e para os programas de concessão de bolsas de estudos e financiamento estudantil do Governo Federal, como o *Programa Universidade para Todos* (ProUni) e o *Programa de Financiamento Estudantil* (Fies).

Como a nota da redação é responsável por uma parcela substancial na pontuação final do aluno, com a nova configuração do ENEM, essa pontuação passou a ser decisiva entre aprovação ou reprovação para a carreira desejada, assim como para a possibilidade de se ter auxílio financeiro para subsidiar os estudos, por isso alunos, professores e interessados questionaram a subjetividade e a pouca transparência no processo de correção da redação do ENEM. Em resposta a esses questionamentos, para as edições de 2012 e 2013, foram elaborados documentos sobre a redação do ENEM, com o objetivo de tornar o mais transparente possível a metodologia de sua correção, bem como o que se espera do participante em cada uma das competências avaliadas (BRASIL, 2012, 2013).

No caso das redações dos vestibulares, os *sites* das universidades se limitam a divulgar as informações oficiais, como as presentes nos manuais do candidato e as notícias divulgadas pela imprensa: número de candidatos inscritos, número de vagas reservadas pelas cotas, entre outras. Algumas apresentam relatórios bem completos após a aplicação e a correção das provas com destaques sobre as questões: o objetivo principal, comentários sobre os textos apresentados, a resposta esperada, e exemplos de redações elaboradas pelos candidatos, nos diversos níveis de desempenho, acompanhadas de notas explicativas sobre o que foi considerado na correção, como é o caso

da UNICAMP e da UEL (COMVEST, 2013; COPS/UEL, 2012). A UEL também apresenta na revista *Diálogos Pedagógicos* os comentários de todos os itens de seus exames vestibulares, tanto das questões descritivas como das objetivas. Mesmo assim, nos *sites* das principais universidades públicas, não existem estudos divulgados sobre a validade e a confiabilidade de seus exames vestibulares. Também há escassez de informações no que tange aos critérios de correção, como, por exemplo, as escalas e os tipos de pontuações utilizados.

1.2 APRESENTAÇÃO DO PROBLEMA DE PESQUISA

A Teoria Clássica dos Testes (TCT) começou a ser desenvolvida no início do século 20, mais precisamente em 1904, por Charles Spearman, com o reconhecimento da presença de erros nas medições e a concepção de erro como uma variável aleatória, as correlações e os posicionamentos. Posteriormente, a TCT foi sendo aperfeiçoada até atingir a forma conhecida atualmente, versão elaborada por Melvin Novick em 1966 (TRAUB, 1997). Desde então, a Teoria Clássica dos Testes tem sido utilizada nas análises dos resultados das avaliações e seu foco principal centra-se na confiabilidade dos resultados dos testes (BEHIZADEH; ENGELHARD, 2011). Na verdade, as pesquisas sobre avaliações escritas se concentraram em duas correntes durante o século 20, as teorias de medição com as pontuações e as escalas e as teorias da escrita, destacando ideia, forma e conteúdo e contexto sociocultural. Autores defendem o surgimento de uma nova disciplina denominada de *writing assessment* em inglês, responsável por estudos que agregam as teorias da medição e as teorias da escrita com o potencial de definir o cenário das avaliações escritas no século 21 (BEHIZADEH; ENGELHARD, 2011; HAMP-LYONS, 2011; YANCEY, 1999; HUOT, 1990).

Atualmente no setor educacional, há uma forte tendência para que as avaliações estejam mais direcionadas para a avaliação da aprendizagem, no lugar dos testes tradicionais de conhecimentos, o que tem intensificado o interesse pelas avaliações com itens abertos. Acredita-se que os testes com itens de respostas abertas são necessários para obter informações sobre o pensamento de ordem superior das pessoas[1] (JONSSON; SVINGBY, 2007; MESSICK, 1996) e podem, de certa forma, reproduzir atividades relaciona-

[1] Segundo Lipman (1995), as três características básicas do pensamento de ordem superior são o pensamento: (1) conceitualmente rico, (2) coerentemente organizado e (3) persistentemente investigativo (LIPMAN, 1995, p. 37)

das ao mundo real do estudante, uma vez que a aprendizagem é um produto do contexto em que ela ocorre. Assim, esse tipo de avaliação pode tentar refletir melhor a complexidade da realidade e fornecer dados mais válidos sobre a competência da pessoa que está sendo avaliada (DARLING-HAMMOND; SNYDER, 2000).

Um dos principais problemas nessas avaliações é a pontuação de um mesmo desempenho com graus diferentes de severidade. Quando existem vários avaliadores, seria ideal se todos atribuíssem exatamente a mesma pontuação para os mesmos desempenhos observados, esta é a condição principal para se ter confiabilidade de pontuação. Entretanto, são muitos os fatores que podem causar variabilidade nessas pontuações, especialmente quando se trata dos testes com itens abertos. As características pessoais dos avaliadores, tais como cultura, experiências, expectativas, estilo de correção, entre outras variáveis, podem influenciar substancialmente a pontuação das tarefas. Esses fatores podem ser tão importantes para a pontuação quanto a qualidade da resposta escrita pelo participante da avaliação (HARSCH; MARTIN, 2012; WEIGLE, 2002, 1999; COHEN, 1994).

Outra classe de problemas que interferem na obtenção de bons índices de confiabilidade é a tendência dos avaliadores em julgamentos sistemáticos dos desempenhos avaliados. Essas tendências são comportamentos frequentemente citados nas pesquisas e são consideradas componentes geradores de erros importantes na pontuação de tarefas escritas. Alguns dos efeitos mais citados são: o efeito da severidade, que é a tendência em avaliar de maneira muito exigente ou muito branda em comparação com a pontuação atribuída por outros avaliadores ou em comparação com classificações preestabelecidas como referência; o efeito halo que ocorre quando os avaliadores não conseguem distinguir entre um número de categorias conceitualmente distintas e avaliam o desempenho da pessoa com base em uma impressão geral, desse modo, diferentes desempenhos podem obter a mesma pontuação; o efeito de tendência central, que é caracterizado pela tendência em classificações perto do ponto médio da escala, evitando, desse modo, classificações nos extremos da escala; o efeito de aleatoriedade, que é a tendência que o avaliador tem de aplicar uma ou mais categorias da escala de maneira inconsistente com o modo com que os outros avaliadores aplicam a mesma escala. O avaliador que possui esta última tendência é demasiadamente inconsistente no uso da escala, apresentando uma maior variabilidade aleatória do que o esperado na avaliação (KNOCK; READ; RANDOW, 2007; MYFORD; WOLFE, 2004).

Tradicionalmente, a variabilidade causada por diferenças entre os avaliadores tem sido controlada por meio da pontuação por vários avaliadores.

Acredita-se que a confiabilidade das pontuações aumenta quando as tarefas são avaliadas por pessoas diferentes. A principal fonte para a determinação de confiabilidade das pontuações pela Teoria Clássica dos Testes (TCT) é a determinação do quanto os avaliadores concordam em suas pontuações. No entanto, a ideia de que basta a confiabilidade entre os avaliadores para garantir uma medida justa da habilidade das pessoas tem sido questionada (ENGELHARD, 1991; LINACRE, 1994).

Algumas justificativas são apontadas para essa desconfiança. Uma delas é que dois avaliadores podem concordar em suas pontuações e, mesmo assim, errar em seus julgamentos, fato preocupante, pois os dois avaliadores estariam errando na mesma direção, subestimando ou superestimando a real habilidade avaliada. Por outro lado, é possível que os avaliadores discordem em seus julgamentos, mas em sentidos opostos, e a média entre essas pontuações pode resultar em uma medida mais aproximada da habilidade real do examinando. Essas possibilidades são descartadas pela determinação de confiabilidade pela TCC. Outro ponto questionado pelos pesquisadores é a expectativa de que os avaliadores sejam igualmente severos em seus julgamentos. O treinamento rigoroso para que os avaliadores concordem em suas pontuações restringe a liberdade e pode levar a uma característica determinística nos dados produzindo uma segurança artificial e ilusória nos resultados da avaliação (LINACRE, 1994).

As avaliações com testes de itens abertos possuem outros aspectos, além do estabelecimento e pontuação das tarefas, que geram preocupações e questionamentos por parte dos especialistas e também da população em geral. Um deles é a comparação entre avaliações com itens abertos, em especial as da linguagem, fato provocado pela intensificação da utilização de matrizes comuns de referência desenvolvidas para orientar os currículos em todos os níveis de ensino em países da Europa, nos Estados Unidos, na Austrália, no Brasil, entre outros (HAMP-LYONS, 2004; NORTH, 2000).

No âmbito educacional, a comparabilidade tem um significado amplo e diz respeito a muitos aspectos relacionados à comparação entre avaliações. Abrange muitas definições, metodologias e métodos, principalmente quando se trata de comparabilidade dos padrões educacionais, incluindo comparações de sistemas e resultados educacionais em uma série de contextos diferentes.

Segundo Elliott (2011), a proliferação das terminologias utilizadas nos últimos anos para descrever diferentes aspectos da investigação sobre a comparabilidade é uma das questões que têm afligido os pesquisadores, principalmente porque as diversas denominações tornam tanto os resultados quanto os problemas difíceis de explicar para o público não especializado, incluindo

os participantes dos exames. Como existe uma variedade cada vez maior de avaliações em larga escala para as mais diversas finalidades, a questão da comunicação sobre os padrões adotados nas avaliações e nos seus resultados torna-se cada vez mais importante.

Para a comunidade científica, é primordial a veracidade das afirmações sobre a manutenção dos padrões de qualificação em episódios diferentes de uma avaliação, com a afirmação de equivalência entre elas. A comparabilidade é uma área cercada por suposições, muitas vezes mal fundamentadas, e considerada por alguns como um terreno estéril (ELLIOT, 2013, 2011; HAERTEL; LINN, 1996).

Outro assunto muito discutido na literatura atual é o grau de dificuldade e a discriminação do item. São muitos os fatores que podem afetar o grau de dificuldade de um item com respostas construídas (JEFFERY, 2009; SUDWEEKS; REEVE; BRADSHAW, 2005; BRELAND *et al.*, 2004; HAMP-LYONS; MATHIAS, 1994; POMPLUM *et al.*, 1992). Aliás, este é considerado um ponto problemático nas avaliações escritas, especialmente nas avaliações da expressão escrita, pois ainda não está totalmente estabelecido o grau de dificuldade das variadas formas das tarefas.

Os principais questionamentos são: O grau de dificuldade depende de o texto ser descritivo, narrativo ou argumentativo? O grau de dificuldade é o mesmo para todos os respondentes do teste? Deve-se oferecer aos candidatos uma tarefa única, uma escolha de tarefas ou tarefas múltiplas? (HAMP-LYONS, 2011). Uma preocupação adicional na escolha do formato do teste é que algumas características da solicitação podem tornar a tarefa mais difícil do que a estabelecida em outros testes. Além disso, deve ser evitado que a escolha da tarefa e do tipo de teste possa privilegiar determinados subgrupos, proporcionando alguma vantagem a esses na pontuação final do teste. Essas e outras questões permanecem sem uma resposta definitiva, indicando, de certa forma, que ainda há muito trabalho a ser feito no campo das avaliações com itens abertos (HAMP-LYONS, 2011; HUANG, 2008; BRIDGEMAN; MORGAN; WANG, 1997; JENNINGS *et al.* 1999).

Também devem ser definidos os critérios e a escala de pontuação que serão utilizados na correção, a experiência e o treinamento dos avaliadores, até mesmo a maneira como será apresentado o resultado ao respondente (*feedback*), sem falar nas análises estatísticas para verificar a validade e a confiabilidade, e outros estudos, como os da dimensionalidade ou generabilidade. Existem estudos que comprovam que esses e outros fatores afetam a qualidade da avaliação escrita, e o impacto desses fatores é determinante para a precisão e, consequentemente, a justiça da pontuação obtida pelos res-

pondentes (HAMP-LYONS, 2011; PASQUALI, 2010; LINACRE; WRIGHT, 2002; WRIGHT; LINACRE, 1987).

Uma preocupação comum com as avaliações diz respeito à "equidade" do teste no que se refere à justiça para com as pessoas. Para um teste bem projetado, é necessário garantir que ele será justo e apropriado para todos os participantes (ETS, 2009).

Para as avaliações com itens de respostas abertas, principalmente em relação às provas de redação, frequentemente são utilizadas abordagens da Teoria Clássica dos Testes (TCT) para o monitoramento da qualidade das pontuações. Duas dessas abordagens são as estimativas de consenso e as estimativas de consistência.

As estimativas de consenso envolvem cálculos da precisão da pontuação e são utilizadas quando os avaliadores são treinados para julgamentos baseados em critérios de pontuação em escalas contínuas que representam o desempenho do indivíduo quanto ao construto avaliado. Para esses cálculos, as estatísticas mais populares utilizadas são as porcentagens do número de acordo entre os avaliadores e a estatística kappa de Cohen (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

As estimativas de consistência baseiam-se no pressuposto de que não é realmente necessário que dois avaliadores tenham o mesmo entendimento da escala e atribuam a mesma pontuação para uma tarefa específica, desde que cada avaliador seja consistente na classificação do desempenho avaliado de acordo com sua própria definição da escala. As estatísticas mais populares utilizadas nesse caso são os coeficientes de correlação de Pearson e de Sperman, além do coeficiente alfa de Cronbach (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

Essas abordagens para a análise dos dados fornecem estatísticas apenas no nível de grupo e não de cada elemento individualmente, e por esse motivo são limitadas quando se busca o aperfeiçoamento de um sistema de avaliação complexo. Seria muito vantajosa a obtenção de informações individuais dos elementos participantes do processo, como cada avaliador, cada examinando, cada item e cada escala de avaliação utilizada (MYFORD; WOLFE, 2000).

No contexto das avaliações em larga escala, o modelo multifacetas de Rasch (MFR) pode ser utilizado como uma ferramenta eficiente para aferir a qualidade das avaliações com itens de respostas construídas. Esse método é uma extensão do modelo da TRI de um parâmetro (modelo de Rasch) (RASCH, 1960) desenvolvido por Linacre em 1989. No modelo básico de Rasch, os itens do teste e os indivíduos são avaliados e colocados em uma

mesma escala de acordo com suas capacidades (indivíduos), ou dificuldades (itens). O modelo multifacetas de Rasch, ao contrário do modelo original, que possui um único parâmetro, permite a inclusão de outros parâmetros, fontes de erros sistemáticos nas avaliações, como as diferenças entre as pontuações dos avaliadores, os erros causados por inconsistências nos julgamentos dos próprios avaliadores e as diferenças na dificuldade relativa de tarefas distintas (ENGELHARD, 2013; ECKES, 2011; SUDWEEKS; REEVE; BRADSHAW, 2005; LINACRE, 1994).

O modelo MFR permite também 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, cada critério de pontuação utilizado, etc. Essa possibilidade de obter informações que possam servir de diagnóstico, no nível individual, sobre o funcionamento de cada elemento em particular é considerada valiosa e torna a utilização do modelo multifacetas de Rasch ainda mais vantajosa (ENGELHARD, 2013; ECKES, 2011; LINACRE, 1994).

Esse modelo está sendo utilizado para analisar a pontuação em avaliações com itens abertos em diversas áreas, mas tem se tornado popular, especialmente, em avaliações da escrita (MACNAMARA; KNOCH, 2012; SUDWEKS; REEVE; BRADSHAW, 2005; MYFORD, 2002) e nas avaliações de inglês para estrangeiros (LIM, 2011; JOHNSON; LIM, 2009; MYFORD; WOLF, 2000; WEIGLE, 1999).

Desta forma, devido ao grande número e à complexidade das variáveis envolvidas nas avaliações com itens abertos em larga escala, orientações especializadas de diversas áreas são essenciais em todas as etapas da elaboração dessas avaliações: concepção inicial, elaboração dos itens, pontuação das tarefas, análises sobre a confiabilidade e a validade, entre outras.

Além disso, as avaliações em larga escala devem satisfazer padrões profissionais de qualidade. Quando são detectados aspectos da avaliação que não estão funcionando de acordo com esses padrões, eles devem ser corrigidos para a próxima edição da avaliação. Para tanto, são necessários métodos eficazes para a identificação desses pontos problemáticos.

Sendo assim, com base nas informações citadas anteriormente, elaborou-se a seguinte questão-problema:

"Qual é a contribuição que a utilização do modelo multifacetas de Rasch pode proporcionar para a análise de avaliações com itens de respostas construídas?"

1.3 OBJETIVOS

1.3.1 Objetivo principal

O objetivo principal do presente trabalho é determinar como o modelo multifacetas de Rasch pode contribuir para a determinação da qualidade das avaliações com itens de respostas construídas.

1.3.2 Objetivos específicos

1. Identificar as variáveis e as teorias envolvidas no processo da concepção, elaboração, aplicação e pontuação das avaliações em larga escala com itens de respostas construídas.
2. Determinar a qualidade das avaliações com itens de respostas construídas no que tange aos critérios de pontuação, às escalas de classificação e aos julgamentos dos avaliadores no contexto do modelo MFR.
3. Propor uma nova metodologia para a pontuação de testes com itens de respostas construídas e consequente classificação dos examinandos.
4. Analisar os dados empíricos provenientes de uma avaliação da habilidade de expressão escrita real por meio do modelo MFR.

1.4 JUSTIFICATIVA

Este trabalho pode ser justificado a partir de dois aspectos: quanto à sua relevância e quanto ao ineditismo.

1.4.1 Relevância

As avaliações em larga escala, dependendo da área na qual estão sendo aplicadas, são responsáveis por orientar decisões importantes. Servem como suporte para implementar melhorias ou para suprir eventuais problemas detectados, além de selecionar pessoas capacitadas para desempenhar alguma função. Nas avaliações educacionais, os objetivos podem estar direcionados para as diferenças individuais, avaliando o desempenho dos estudantes

em diversas situações, como também na avaliação de programas ou de projetos educacionais, subsidiando ou justificando alguma ação na esfera política. Não se pode deixar de destacar o efeito retroativo das avaliações, que considera o impacto das avaliações no ensino e que provavelmente o influenciam, servindo como guias para a instrução em sala de aula (SCARAMUCCI, 2004, 2011; QUEVEDO-CAMARGO, 2011, 2014). Deste modo, a validade das medidas e suas interpretações são de suma importância, com consequências que podem afetar a população envolvida e até mesmo a sociedade. O desenvolvimento de novas metodologias de medição e avaliação, que resultem em medidas de maior precisão, torna-se mais importante a cada dia (HAMP-LYONS, 2002, 2011).

Apesar da importância das avaliações que utilizam questões abertas no Brasil, ainda são poucas as pesquisas que analisam os processos de correção e pontuação de provas desse tipo no país. A maior parte delas data das décadas de 1980 e 1990 e discorre sobre a redação no vestibular, época em que essas avaliações passaram a ser utilizadas em maior número. Algumas delas são: Moraes (1997), Rocco (1995), Sossai *et al.*, (1995), Hoffman (1988), Bessa (1986), Vianna (1976a, 1976b, 1978, 1982, 1995). Uma grande parte das pesquisas recentes é na área da linguística aplicada, direcionadas ao ensino e a problemas de aprendizagem relacionados com a leitura e a escrita (VICENTINI, 2011; GOMES, 2009; GIMENEZ, 1999).

Vianna (2003) examina os problemas ligados às avaliações em larga escala no Brasil e critica a ausência de validação de conteúdo e de construto e a falta de preocupação com a confiabilidade dos resultados em relação ao ENEM, ao SAEB e a outros exames brasileiros da época. Destaca também a escassez de trabalhos que discutem a problemática dessas avaliações e seus impactos na sociedade, assim como a deficitária divulgação oficial dos resultados desses exames por parte dos órgãos responsáveis (VIANNA, 2003). Em 2004, a pesquisadora Matilde Scaramucci destacou que os exames em larga escala no Brasil são inseridos e descartados sem estudos sobre a sua validade, confiabilidade ou impactos que exercem no ensino e na sociedade (SCARAMUCCI, 2004). Pouca coisa mudou de lá para cá, uma vez que são poucas as pesquisas, nos últimos 10 anos, sobre os principais exames nacionais, principalmente em relação à validade e à confiabilidade dos instrumentos. Essa escassez de estudos que tratam das avaliações em larga escala no Brasil resulta na pouca transparência dos processos envolvidos na elaboração, correção e pontuação, principalmente dos exames com itens abertos, como é o caso das redações dos vestibulares (VICENTINI, 2011).

O enfoque deste estudo é a análise da qualidade de processos avaliati-

vos que utilizam testes com itens de respostas construídas. Nessas avaliações, são muitos os fatores que podem afetar a medida do desempenho das pessoas ao executar a tarefa determinada no teste. A elaboração desses exames consiste em um conjunto diverso e complexo de procedimentos que visam à medida da proficiência sobre o construto que se deseja medir. Esses testes podem variar em uma gama de diferentes formatos e sofrem interferências de variáveis que podem fazer parte ou não da situação de avaliação.

Nesse sentido, é realizada, neste trabalho, uma análise crítica de diversos estudos e pesquisas sobre cada uma das etapas que compõem as avaliações com itens de respostas construídas, especialmente com relação àqueles que resultam na determinação da qualidade dessas avaliações. A complexidade desses processos avaliativos e o impacto que eles causam na vida das pessoas e na sociedade, juntamente com a carência de estudos que estabelecem a qualidade das avaliações e, ao mesmo tempo, disponibilizam novas técnicas para análises eficientes, confirmam a relevância do trabalho.

Além disso, este trabalho possui também relevância de ordem prática. As empresas provedoras de avaliações em larga escala necessitam de mecanismos que auxiliem na construção e análises das avaliações, especialmente para as avaliações com itens de respostas construídas. Na realidade, as técnicas e os modelos existentes abordados na literatura são na maior parte desenvolvidos para cada etapa da avaliação isoladamente, não integrando os diversos procedimentos e processos demandados. Há, portanto, a necessidade da elaboração de modelos práticos que possam ser aplicados como um todo e que englobem todo o processo. Assim, este trabalho pretende contribuir no sentido de elaborar uma sistemática para a concepção e a construção de avaliações com itens de respostas construídas, explicitando claramente os conceitos envolvidos na elaboração dos instrumentos de avaliação, dos critérios de correção e de pontuação para as tarefas estabelecidas, assim como as análises estatísticas para a determinação da validade da avaliação e classificação dos candidatos.

1.4.2 Ineditismo

O ineditismo deste estudo pode ser verificado em dois aspectos principais: a) determinação da qualidade de avaliações em larga escala com itens de respostas construídas; e b) utilização do modelo multifacetas de Rasch para o acompanhamento da qualidade da avaliação no que se refere à pontuação das tarefas, à estrutura das escalas de classificação, à dificuldade dos itens e à

severidade dos avaliadores.

Em busca no Banco de Teses da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), verificou-se a não existência de pesquisas relacionadas à qualidade das avaliações em larga escala e tampouco à validade de seus instrumentos. Um único trabalho presente nesse banco de teses trata da pontuação de redações dos concursos vestibulares e o faz por meio do desenvolvimento de um modelo computacional para a avaliação automática das redações (NOBRE, 2011). As outras pesquisas sobre as provas de redação dos vestibulares e do ENEM são específicas da área de letras e ensino, a maioria relacionada com linguística aplicada, análise do discurso, gênero discursivo, ensino e aprendizagem.

Também não foram encontradas pesquisas relacionadas à construção de avaliação com itens de respostas abertas no que tange à concepção da avaliação, à elaboração das tarefas e itens, à pontuação dos testes, à validade do instrumento de avaliação, à qualidade dos critérios para a pontuação e da escala de classificação utilizada. Em relação ao modelo multifacetas de Rasch, no Banco de Teses da CAPES, não há pesquisas que correspondam a esse modelo quando se busca por seu nome como palavra-chave.

Ainda sobre pesquisas brasileiras, alguns estudos divulgados nas principais bases de pesquisa relacionam as políticas educacionais, os impactos na educação básica e as consequências sociais envolvidas nos exames em larga escala (DE CASTRO, 2009; DE SOUZA e GOUVEIA, 2011; KLEIN; FONTANIVE, 2009b; VIDAL; FARIAS, 2008), outros fazem uma associação entre o nível socioeconômico dos alunos e os resultados educacionais com dados provenientes das avaliações (ALVES; GOUVÊA; VIANA, 2012; BRANDÃO; CANEDO; XAVIER, 2012).

Existem inúmeros estudos divulgados sobre as variáveis que envolvem as avaliações de itens de respostas construídas, especialmente as avaliações da escrita, nos mais diferentes contextos. Uma parte desses estudos está focada nos aspectos mais gerais, como a psicologia educacional, o ensino e aprendizagem e as práticas de sala de aula, e também sobre as iniciativas de políticas públicas (MITCHELL; McCONNELL, 2012; JEFFERY, 2009).

As empresas ou órgãos provedores de algumas avaliações em larga escala nos Estados Unidos e também alguns órgãos do governo divulgam a metodologia utilizada na elaboração, aplicação e análises dos resultados das suas avaliações. Pode-se citar *O National Post secondary Education Cooperative* (NPEC), que é uma cooperativa com o objetivo de coletar, analisar e divulgar estatísticas e outros dados relacionados com a educação nos Estados Unidos e em outras nações. O NPEC editou um relatório técnico em 2000,

no qual apresenta uma revisão detalhada dos métodos de avaliação que foram desenvolvidos para medir as habilidades de pensamento crítico, de resolução de problemas e de comunicação, como a habilidade da expressão escrita, para estudantes de Ensino Superior nos Estados Unidos. O capítulo desse trabalho dedicado à avaliação da escrita tem início com a definição da competência demandada para escrever, com uma visão geral das abordagens existentes, e organiza em uma tabela os componentes de habilidades de escrita que são medidos pelos vários testes de nível superior existentes no país. Também são descritos os diferentes formatos de testes utilizados para avaliar a habilidade da escrita com as considerações sobre as vantagens e as desvantagens de cada método e traz os detalhes dos procedimentos metodológicos e técnicos utilizados, tais como: confiabilidade, validade, pontuação, tempo de prova, custo, fins específicos, usuários, utilidade, propriedades psicométricas, escalas e critérios. Apesar de ser um trabalho bastante amplo, seu interesse maior reside na necessidade de avaliar as habilidades que estão sendo desenvolvidas nas faculdades e universidades como um meio de comparar o que está sendo aprendido pelos estudantes nessas instituições. O *The National Center for Education Statistics – NCES* é um órgão federal norte-americano, que tem a finalidade de coletar, analisar e divulgar os dados relacionados à educação nos Estados Unidos e outros países. Organizou em 1996 um importante relatório técnico, *Technical Issues in Large-Scale Performance Assessment*, com o objetivo de contribuir para o debate, descrevendo alguns dos problemas técnicos que devem ser considerados no desenvolvimento de avaliações de desempenho em larga escala. Essas avaliações são aquelas que demandam a execução de alguma tarefa por parte dos examinandos e o julgamento das tarefas elaboradas por avaliadores especialistas. O relatório é composto por cinco capítulos, abordando: validade, generalização, comparabilidade, padrões de desempenho, equidade e justiça. Nessa época, houve um aumento significativo na utilização das avaliações de desempenho em larga escala, mas uma parte dos procedimentos ainda estava em estágio experimental. Por esse motivo, alguns resultados tratados nesse estudo, assim como alguns direcionamentos, já foram modificados.

O *The National Assessment of Educational Progress – NAEP* é uma avaliação americana, realizada periodicamente desde 1969, em leitura, matemática, ciências, história, geografia e redação. Desenvolve trabalhos no formato de relatórios técnicos para divulgar informações metodológicas e as análises dos resultados sobre as avaliações (U.S., 2003, 2005, 2010). Os relatórios desenvolvidos tratam das metodologias utilizadas, mas não discutem com profundidade e fundamentação os motivos de determinadas escolhas ou

procedimentos.

O Ministério do Trabalho dos Estados Unidos preparou o guia *Testing and assessment: an employer's guide to good practices* (2000) com a finalidade de auxiliar os gerentes e profissionais de recursos humanos nas práticas de avaliação, fornecendo os conceitos essenciais para a elaboração e a utilização dos testes. O intuito desse trabalho é fornecer informações para que os profissionais possam avaliar e selecionar ferramentas ou procedimentos de avaliação que são os mais eficientes e eficazes para suas necessidades específicas, interpretar com precisão os resultados da avaliação, além de compreender os padrões profissionais e legais que devem ser seguidos na condução de avaliação de pessoal. Esse guia aborda os aspectos críticos e os problemas envolvidos em cada etapa do processo de avaliação e discorre sobre treze princípios da avaliação explicando cada um deles de maneira bastante completa. Apesar de abordar alguns detalhes sobre avaliação com itens de respostas construídas, o faz de maneira superficial, pois não foi desenvolvido para esse objetivo específico.

No Brasil, as informações e análises feitas pelas universidades sobre os concursos vestibulares são superficiais. Limitam-se às instruções fornecidas aos candidatos nos manuais do candidato. Algumas divulgam também as provas comentadas, com comentários sobre o que foi ou não considerado na correção, como é o caso da UNICAMP (COMVEST, 2013), FUVEST (FUVEST, 2013) e UEL (COPS, 2013).

O INEP, responsável pelo ENEM, divulgou nos cadernos *A redação no ENEM 2012 – Guia do participante* e *A redação no ENEM 2013 – Guia do participante* (BRASIL, 2012, 2013) a metodologia utilizada para a correção da redação das edições do ENEM dos respectivos anos. Discorre sobre as competências avaliadas, as rubricas de pontuação e a escala, mas não discute os motivos das escolhas feitas, nem fornece informações sobre a validade e a confiabilidade da avaliação.

Outra grande quantidade de trabalhos estuda etapas específicas da elaboração da avaliação, como a escolha das tarefas, os tipos de comandos (*prompts*), a escolha dos critérios e escalas de pontuação e quais os instrumentos de medidas utilizados para avaliar a validade e a confiabilidade (KROLL; REID, 1994; BROWN *et al.*, 1991). Também existem pesquisas sobre as influências históricas das teorias de medidas sobre a prática da avaliação escrita nos Estados Unidos em períodos específicos do século 20 (SERVISS, 2012; BEHIZADEH; ENGELHARD, 2011; JUZWIK *et al.*, 2006).

As pesquisas encontradas na literatura para elaborar e validar as avaliações em larga escala o fazem, na maior parte, para cada etapa isoladamente,

embora alguns pesquisadores proponham que ao menos algumas etapas de uma avaliação devam ser desenvolvidas e validadas por meio de procedimentos integrados (ECKES, 2011; ENGELHARD, 2013; PASQUALI, 2010; RUTH; MURPHY, 1984; WEIGLE, 2012; KNOCH, 2011a). Desta forma, o diferencial deste trabalho consiste na integração de diversas etapas das avaliações com itens de respostas construídas, buscando a validade do instrumento de medição e a qualidade do sistema avaliativo como um todo, sendo os procedimentos desenvolvidos no contexto do modelo multifacetas de Rasch.

1.4.3 Aderência à Engenharia de Produção

O trabalho proposto está vinculado ao Programa de Pós-Graduação em Engenharia de Produção, na área de gestão de operações. A pesquisa aborda as várias etapas envolvidas na avaliação e seleção de pessoas, o que justifica estar também inserida na linha de pesquisa Avaliação de desempenho, que faz parte da área em questão.

O modelo multifacetas de Rasch, quando aplicado às avaliações com itens de respostas construídas, permite análises no nível individual de cada elemento participante das avaliações, mostrando-se eficaz para a detecção de erros. Desse modo, é possível que os erros sejam corrigidos resultando na melhoria dos processos avaliativos.

Este trabalho será útil também por ter alguns dos conceitos essenciais para a elaboração e a aplicação de testes explicados detalhadamente, possibilitando aos gestores e profissionais de recursos humanos (RH) selecionar pessoas com o uso de instrumentos válidos, de maneira honesta e precisa; avaliar e selecionar as ferramentas apropriadas para cada objetivo da avaliação, com a finalidade de alcançar o melhor ajuste entre as vagas de trabalho e empregados; escolher e administrar os instrumentos de avaliação mais eficientes para as suas necessidades específicas; interpretar com precisão os resultados da avaliação; compreender as variáveis que devem ser consideradas na condução da avaliação de pessoas.

1.5 DELIMITAÇÕES

Embora a proposta deste trabalho seja a de determinar a qualidade de avaliações em larga escala com itens de respostas construídas, os procedimentos e resultados limitam-se ao estudo de tarefas de escrita de textos,

como as redações dos exames de seleção e concursos vestibulares. Os resultados alcançados aqui poderão ser estendidos facilmente para outros tipos de avaliações, como por exemplo, as provas com respostas construídas de disciplinas do ensino médio, tão comuns nas salas de aulas e também em alguns concursos vestibulares de instituições importantes.

Outra limitação deste trabalho, reside no fato de as pontuações às tarefas de escrita não terem sido feitas por profissionais experientes, elas foram pontuadas por estudantes dos cursos de graduação e pós-graduação em Letras além de professores do ensino médio, da área de língua portuguesa que se interessaram pelo treinamento. Este fato pode ser responsável por gerar um maior índice de desacordos entre os avaliadores do que seria alcançado em uma avaliação real, com avaliadores profissionais.

1.6 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em seis capítulos. O primeiro capítulo apresenta uma introdução às avaliações em larga escala com itens de respostas construídas, a contextualização, a apresentação do problema de pesquisa, o objetivo principal e os objetivos específicos, a justificativa com a relevância e o ineditismo do projeto, as limitações e a estrutura do trabalho.

O segundo capítulo apresenta uma revisão de literatura sobre as avaliações com itens de respostas construídas, mais especificamente as teorias da escrita, contendo a descrição e a conceitualização de cada uma das etapas da elaboração de uma avaliação da escrita. São elas: tipos de avaliação, temas e formatos da redação, validade, confiabilidade, critérios de pontuação e tipos de escalas de classificação, treinamento de avaliadores e análises dos resultados. Nesse capítulo também é desenvolvido uma sistemática contendo as etapas demandadas para a construção de avaliações em larga escala com itens de respostas construídas.

O terceiro capítulo aborda os modelos de Rasch, o modelo multifacetas de Rasch e as principais estatísticas que fazem parte de seu contexto, assim como o método de estimação utilizado.

O quarto capítulo contém a metodologia da pesquisa, com métodos de abordagem que consistem nos procedimentos técnicos e classificação da pesquisa, os procedimentos e critérios adotados na pesquisa bibliográfica, a metodologia aplicada para o desenvolvimento da tese e para a coleta de dados.

No quinto capítulo, estão apresentados os resultados das análises. Primeiramente são apresentadas as estatísticas referentes ao ajuste global dos

dados ao modelo multifacetas de Rasch com resumos das medidas fornecidas pelos modelos de escala gradual e de crédito parcial. Na sequência são feitos estudos sobre cada uma das quatro facetas incluídas no modelo multifacetas de Rasch, sendo elas, a habilidade dos examinandos, a dificuldade das tarefas, a dificuldade dos itens, a severidade dos avaliadores e ainda sobre a estrutura da escala de avaliação utilizada para a pontuação das tarefas elaboradas pelos examinandos. As análises são feitas tanto no nível global quanto no nível individual sobre os elementos de cada uma das facetas.

No sexto capítulo, são feitas as conclusões e considerações finais, além de sugestões sobre a realização de trabalhos sobre o tema, que podem ser elaborados no futuro.

As referências utilizadas nesse estudo são apresentadas na sequência, finalizando com os apêndices e os anexos.

2 AVALIAÇÃO COM ITENS ABERTOS

As avaliações com itens de múltipla escolha foram muito utilizadas durante a maior parte do século XX pela necessidade de instrumentos de medição precisos, altamente estruturados, com propriedades específicas e replicáveis. Essas vantagens estavam longe de serem alcançadas em avaliações com itens abertos com respostas construídas, tais como as provas de redação.

Atualmente, são muitas as críticas às avaliações com itens objetivos. Os itens de múltipla escolha nem sempre medem o construto pretendido, medindo muitas vezes apenas construtos substitutos. A justificativa para essa afirmação considera que a capacidade de selecionar a resposta correta em uma pequena lista de possíveis respostas está longe de ser a capacidade de aplicar conhecimentos e habilidades em situações reais do trabalho ou do dia a dia.

Os avanços na teoria de medição, com o desenvolvimento de novas ferramentas e técnicas, passaram a permitir que as avaliações com itens abertos alcançassem a validade psicométrica nos mesmos padrões dos testes de múltipla escolha. Esses avanços podem propiciar avaliações, consideradas pelos especialistas, *avaliações autênticas* (LINACRE *et al.*, 1994).

As avaliações com itens de respostas construídas na forma de redação são muito utilizadas para avaliar a capacidade de expressão escrita das pessoas e podem abranger conteúdos relacionados apenas com a linguagem ou também assuntos de outras áreas, como conhecimentos gerais ou conteúdos específicos que fazem parte de alguma disciplina. Entretanto, independentemente do conteúdo abordado, por necessitar de habilidades necessárias para a comunicação escrita, essas avaliações são também denominadas de avaliações escritas.

As avaliações da escrita são objeto de inúmeras pesquisas. São comuns em todos os continentes e possuem as mais diversas finalidades. Entre as avaliações com itens abertos, as avaliações da escrita, sem dúvida alguma, são as mais estudadas e por este motivo possuem teorias específicas já estabelecidas. Desse modo, as teorias que são utilizadas em avaliações com itens abertos em geral foram desenvolvidas inicialmente para as avaliações da escrita e então generalizadas para as avaliações com itens abertos de outras disciplinas. Portanto, a seção seguinte tratará das teorias da avaliação da escrita e das influências sofridas por esse tipo de avaliação ao longo do século XX.

2.1 TEORIAS DA AVALIAÇÃO DA EXPRESSÃO ESCRITA

Na sociedade moderna, a palavra escrita é considerada fundamental para expressar as competências comunicativas e de alfabetização. O acesso à linguagem escrita é um "bem" que influencia grandemente o acesso a muitos outros "bens" e, segundo Hamp-Lyons (2002), isso faz com que a avaliação escrita seja um ato implicitamente político. Pesquisadores, educadores e políticos defendem que, para melhorar o desempenho da escrita de estudantes em todos os níveis de ensino, é necessária uma revolução no modo de ensinar e, para isso, devido ao seu caráter retroativo, a utilização de avaliações é imprescindível, isto é, ao tentar medir os efeitos do ensino, as avaliações influenciam a qualidade e o conteúdo do que é ensinado (BEHIZADEH; ENGELHARD, 2011; SCARAMUCCI, 2011; HAMP-LYONS, 2002; MESSICK, 1996).

As avaliações da habilidade da escrita em larga escala são elaboradas segundo os conhecimentos desenvolvidos em duas áreas principais: as teorias da escrita e as teorias de medição. As teorias da escrita, tradicionalmente, são subdivididas em três linhas de pesquisa: (1) a ideia, (2) a forma e conteúdo e (3) o contexto sociocultural. Já as teorias de medição são influenciadas por outras duas linhas de pesquisa: (1) a pontuação de testes e (2) as escalas. Ao longo da história no século XX, as avaliações da escrita sofreram influências dessas duas correntes, cada uma delas se sobressaindo à outra em determinados períodos de tempo (HAMP-LYONS, 2011; BEHIZADEH; ENGELHARD, 2011; YANCEY, 1999; HUOT, 1990).

Yancey (1999) definiu a influência dessas correntes teóricas em determinados períodos do século XX como ondas que vão e vêm e que, algumas vezes, se sobrepõem. Destacou três ondas em sua análise, cada onda identificada por um método avaliativo utilizado para medir a competência da escrita. A primeira onda ocorreu no período 1950–1970 e foi dominada pelos testes objetivos. Na segunda onda, de 1970 até 1986, a principal preocupação foi com os critérios de correção e com a pontuação dos ensaios escritos. Na terceira onda, de 1986 até o presente, os estudos têm como foco o formato das tarefas da avaliação e do conteúdo avaliado.

Além disso, Yancey caracteriza o equilíbrio entre os conceitos de validade e confiabilidade como o balanço do pêndulo da avaliação escrita. Define a validade com a pergunta: *Você está medindo o que realmente pretende medir?* E a confiabilidade com essa outra: *Você pode medir de forma consistente?* Embora as duas características sejam desejáveis em toda avaliação escrita, cada uma delas é defendida por uma corrente de adeptos, e, como o aumento de uma dessas características resulta na diminuição da outra, há, de

certa forma, um embate entre os defensores de cada uma dessas correntes, como se apenas uma delas pudesse ser favorecida.

Uma das características da primeira onda foi o favorecimento da confiabilidade devido aos tipos de avaliações empregadas na época e aos objetivos de utilizá-las. Essa onda da avaliação escrita foi dominada por uma única pergunta: "Qual é a melhor medida da escrita?", mas a resposta a essa questão levou em conta outros fatores, como as necessidades institucionais, o custo e a eficiência, resultando em uma alteração da pergunta para "Qual medida é mais eficiente e mais justa para prever com a menor quantidade de trabalho e o menor custo?"

Na segunda onda, prevaleceu a validade. Na década de 1970, os professores estavam mais preparados quanto aos processos da escrita e da composição de textos aos processos do ensino da escrita. Não fazia, portanto, muito sentido a utilização de testes cujas principais preocupações eram a confiabilidade e a eficiência. A prática de avaliação foi alterada, tendo como critério principal a validade e não a confiabilidade.

A terceira onda ficou caracterizada com o aumento de hipóteses, incluindo na avaliação outras características que poderiam ser medidas e traduzidas em um *esquema de avaliação*. As práticas avaliativas não foram abandonadas quando se passou da primeira para a segunda onda, nem mesmo dessas duas para a terceira.

Elas foram aos poucos abrindo espaço e novas características foram sendo incorporadas, trocadas ou modificadas. Assim, a tecnologia dos *portfólios* foi incluída nas avaliações da escrita. Essa ferramenta se resume em propor um conjunto de tarefas de escrita ao estudante, e foi justificada pela questão: Se um texto aumenta a validade de um teste, o que se poderia dizer de dois ou três textos? (YANCEY, 1999).

Os processos avaliativos no Brasil também foram influenciados por essas correntes teóricas. No ano de 1977, com o Decreto n°. 79.298, a capacidade da expressão escrita passou a ser avaliada com a exigência de prova ou questão de redação em língua portuguesa nos exames vestibulares (BRASIL, 1977).

Yancey, em seu estudo, não se preocupou especificamente com o país de desenvolvimento das pesquisas, mas em destacar os estudos seminais, independentemente da nação de origem, que, por sua vez, desencadearam tendências que facilmente foram se espalhando pelo mundo, fato constatado também no Brasil. Com a inclusão da prova de redação nos exames do vestibular, houve a necessidade de pesquisas que pudessem avaliar o impacto das mudanças e dar mais confiabilidade aos processos avaliativos (MORAES, 1997).

Desse modo, começou a surgir uma série de estudos relacionados com o estabelecimento de escala para a pontuação e de métodos para se obter maior confiabilidade de pontuação entre os avaliadores em testes com itens abertos da expressão escrita (MORAES, 1997; HOFFMAN, 1988; VIANNA, 1978).

Pesquisadores defendem o surgimento de uma nova disciplina denominada *Teoria da avaliação da escrita*, responsável por estudos que agregam as teorias da medição e as teorias da escrita com o potencial de definir o cenário das avaliações da escrita no século 21 (BEHIZADEH; ENGELHARD, 2011; HAMP-LYONS, 2011; YANCEY, 1999; HUOT, 1990).

Behizadeh e Engelhard (2011) fazem um traçado histórico, analisando as interações entre as teorias de medição, as teorias da escrita e as avaliações da escrita nos Estados Unidos e, a partir dos resultados desse estudo, estabelecem os impactos das pesquisas com foco nessas duas teorias, da medição e da escrita, sobre as práticas das avaliações da escrita dentro de períodos de tempo selecionados do século XX. Os autores concentraram seu estudo nas pesquisas desenvolvidas nos Estados Unidos, mas, devido à natureza internacional, são também utilizadas em outras nações, assim como muitas delas tiveram suas origens fora dos Estados Unidos.

Os Quadros 1 e 2 fazem parte do trabalho de Behizadeh e Engelhard (2011) e apresentam o desenvolvimento das principais tradições em pesquisas, tanto das teorias da escrita como das teorias da medida durante o século XX. As pesquisas são dispostas em categorias, facilitando o exame das teorias que se destacaram nos respectivos períodos de tempo.

A pontuação de testes e as escalas, durante o século XX, foram as pesquisas dominantes na teoria de medição (Quadro 1), cuja preocupação principal é referente aos erros de medidas dos resultados dos testes e à elaboração de escalas referenciadas. As pesquisas relacionadas aos resultados do teste incluem os vários modelos da Teoria da Resposta ao Item, cujo intuito principal é destacar o foco nas respostas individuais, em contraste com o foco nas respostas do grupo.

A teoria da escrita (Quadro 2) procura responder à pergunta: "*O que é escrever?*". Três respostas diferentes se destacaram durante o século XX: (1) a escrita como forma incluindo a mecânica, gramática e habilidades isoladas; (2) a escrita como ideias e conteúdo incluindo a criatividade, habilidades aplicadas a situações reais e poéticas; e (3) a escrita como um processo social e culturalmente contextualizado.

Quadro 1 – Tradições de pesquisas em Teoria de medidas

Período	*Tradição em pesquisas*	*Teoria de medida*	*Exemplos de pesquisas*	*Foco da pesquisa*
1900 – 1920	Escalas: dominante	Psicofísica	Thorndike (1904)	Criação de escala
	Pontuação de teste: emergente	Teoria Clássica dos Testes (TCT)	Spearman (1904)	Fontes de variância
1930 – 1940	Pontuação de teste: dominante	Teoria Clássica dos Testes (TCT)	Kuder; Richardson (1937)	Novos métodos para estimar a confiabilidade dos escores do teste
1950 – 1960	Pontuação de teste: dominante	Teoria da Generabilidade (G teoria)	Cronbach *et al.* (1963)	Generabilidade e confiabilidade de escore
	Escalas: emergente	Medida de Rasch	Rasch (1960/1980)	Mapas de Variáveis
		Teoria de Resposta ao Item (TRI)	Birnbaum (1968)	Novas regras de medidas
1970 – 1980	Escalas: dominante	Medida de Rasch	Wright (1977)	Teoria em prática: Solução de problemas de medição
	Pontuação de teste: emergente	Teoria de Resposta ao Item (TRI)	Lord (1980)	Estudos de validade com modelagem de equações estruturais
		Extensões da análise fatorial	Joreskog (1974)	Modelagem de equações estruturais
1990 – presente	Escalas: dominante	Modelos multifacetas de Rash	Linacre (1989)	Avaliações mediadas por avaliadores
	Pontuação de teste: reemergente	Teoria da Generabilidade	Engelhard (1992) Brennan (1992)	Fontes de variação de erro nos testes de itens abertos

Fonte: Adaptado de Behizadeh e Engelhard (2011).

Quadro 2 – Tradições de pesquisas em Teoria da escrita

Período	*Tradição em pesquisas*	*Teoria de medida*	*Exemplos de pesquisas*	*Foco da pesquisa*
1900–1920	Forma: dominante	A escrita como habilidades	Charters e Miller (1915)	Mecânica (análises de erros)
1930–1940	Forma: dominante	Utilidade social da escrita	Hatfield (1935)	Desenvolvimento de livro texto
	Ideia e conteúdo/ contexto sociocultural: emergente	Processo social da escrita	Dewey (1938; 1944)	A teoria em prática
1950–1960	Forma: dominante	Estrutura da escrita	Chomsky (1957)	Linguística
1970–1980	Ideia e conteúdo: dominante	A escrita como um processo cognitivo	Hayes e Flower (1980)	Psicologia cognitiva
	Contexto sociocultural: emergente	A escrita em um contexto social	Heath (1983)	Etnográfico
1990 – presente	Contexto sociocultural: dominante	A escrita em um contexto sociocultural	Lee (2001)	Métodos mistos

Fonte: Adaptado de Behizadeh e Engelhard (2011).

2.2 PROPÓSITOS DA AVALIAÇÃO

O objetivo das avaliações em larga escala é verificar, por meio de geração e coleta de dados para o julgamento correto, os conhecimentos e habilidades dos indivíduos. As avaliações da aprendizagem têm muitas finalidades distintas, como, por exemplo, identificar os pontos fortes e fracos, monitorar para manter padrões, fazer escolhas futuras, fornecer *feedback* ao estudante, planejar o aprendizado, planejar programas de intervenção, avaliar os níveis de desempenho. Elas podem ser aplicadas tanto no âmbito de sistemas educacionais como em salas de aula.

As avaliações são elaboradas e aplicadas conforme a finalidade a que se destinam. Os tipos de avaliações a seguir são definidos segundo Cortezão (2002).

2.2.1 Avaliação diagnóstica: identificar pontos fortes e fracos

Entre os objetivos desse tipo de avaliação está a identificação da habilidade dos alunos para exercer alguma atividade, seja ela de leitura, de produção de textos, de matemática ou de qualquer outra disciplina. Essa avaliação é muito utilizada no início dos trabalhos visando à colocação do aluno em um nível ou grupo. É comum a utilização de várias avaliações durante o período letivo para acompanhar o desenvolvimento do aluno. Os resultados da avaliação diagnóstica podem ser utilizados como *feedback* aos alunos ou para os professores e gestores na identificação de pontos fortes ou fracos. A *Provinha Brasil* é uma avaliação diagnóstica, pois visa avaliar o desenvolvimento das habilidades relativas à alfabetização e ao letramento em língua portuguesa e matemática de crianças do 2º ano do ensino fundamental das escolas públicas brasileiras (BRASIL/MEC).

2.2.2 Avaliação formativa: planejar o aprendizado

O objetivo principal desse tipo de avaliação é obter dados para uma reorientação do processo ensino-aprendizagem. Esse tipo de avaliação normalmente ocorre no ambiente escolar e fornece informações sobre o progresso do aluno, permitindo que professores e alunos identifiquem os problemas que podem afetar o processo de aprendizagem. Os resultados desse tipo de avaliação auxiliam professores e alunos a identificar os pontos que

devem ser melhorados.

A avaliação formativa é usada para identificar necessidades futuras e lacunas na aprendizagem, assim como identificar necessidades individuais de apoio.

2.2.3 Avaliação sumativa: medir o desempenho

O termo "sumativa" é uma variação da palavra "sumário", pois o objetivo desse tipo de avaliação é apresentar de forma resumida o desempenho do estudante ao final de um período letivo ou unidade de ensino, estabelecendo se o aluno está apto a ser promovido para a próxima etapa de formação no curso específico. Por esse motivo é conhecida também como avaliação promocional. A avaliação sumativa pode ser interna, conduzida pelo professor ou estabelecimento, ou externa, realizada por alguma entidade ou governo. Ela é utilizada, também, para confrontar a realização do estudante com alguma especificação particular ou padrão, por isso deve ser sistematicamente projetada e sempre ter a qualidade assegurada.

Uma avaliação em larga escala de âmbito internacional, que pode ser classificada como sumativa, é o PISA. Essa avaliação é aplicada a estudantes de 15 anos em diversos países, idade em que se pressupõe o término da escolaridade básica obrigatória. A finalidade é confrontar as proficiências dos alunos dos países participantes para produzir indicadores que contribuam para a discussão da qualidade da educação, de modo a subsidiar políticas de melhoria do ensino básico. No Brasil, pelo menos duas outras avaliações em larga escala se enquadram nesse tipo de avaliação, o ENEM e o SAEB. O ENEM avalia o desempenho do estudante ao final do Ensino Médio e fornece dados para estudos comparativos nas diversas regiões do país subsidiando ações para o enfrentamento de deficiências de ensino/aprendizagem detectadas, confirmando o caráter retroativo da avaliação sobre o ensino. Já o SAEB avalia algumas séries do Ensino Básico e utiliza os resultados para comparar e acompanhar o progresso dos estudantes (BRASIL/MEC).

2.2.4 Avaliação classificatória: seleção de candidatos

A avaliação classificatória ou de colocação tem o objetivo de classificar o participante em um nível de uma escala de aprendizagem, e suas características se confundem um pouco com a avaliação diagnóstica quando

utilizada em sala de aula. É o tipo de avaliação utilizada nos concursos vestibulares para o acesso às vagas do Ensino Superior ou para a ocupação de vagas de trabalho em empresas. O ENEM também se enquadra nesse tipo de avaliação quando é direcionado para a classificação e seleção de candidatos às vagas de instituições de Ensino Superior (BRASIL/MEC).

2.3 QUESTÕES ESSENCIAIS PARA A AVALIAÇÃO

As avaliações, acima de tudo, devem ser confiáveis e respeitar as normas atuais para proporcionar as pessoas e entidades qualificações que atendam as necessidades educativas, formativas, de diagnóstico ou classificatórias. Para isso, é necessário que as avaliações sejam desenvolvidas de acordo com procedimentos altamente técnicos que envolvem diferentes áreas de pesquisa.

Muitos professores se sentem frustrados ou desinteressados quando o assunto são as avaliações fora da sala de aula, principalmente quando se trata das avaliações com itens abertos, como as provas de redação ou provas de alguns concursos vestibulares. Essa desconfiança é compreensível dada a natureza teórica das variáveis e dos procedimentos envolvidos. Para a maioria das pessoas, as etapas demandadas para o estabelecimento de uma avaliação em larga escala são assuntos restritos àqueles com conhecimentos especializados. Consequentemente, as avaliações em larga escala são alvo de dúvidas e incertezas quanto a alguns aspectos importantes, como o de *estar medindo realmente aquilo que se deve medir*, ou que os julgamentos da habilidade que está sendo medida são feitos de maneira justa, resultando em uma classificação verdadeira e confiável.

As questões fundamentais para uma avaliação em larga escala eficiente consistem em validade, confiabilidade, comparabilidade e justiça. Segundo Messick (1996), esses conceitos não se resumem apenas a princípios de medição. São valores sociais com significado, não simples medidas, e devem ser considerados sempre que decisões de valores são tomadas com base nas avaliações.

Essas questões essenciais para elaboração e aplicação de avaliações em larga escala são descritas nas seções seguintes.

2.3.1 Validade

O conceito de validade vem sendo proposto e modificado desde a década de 1920, consistindo, juntamente com o de confiabilidade, provavelmente nos conceitos mais polêmicos e discutidos atualmente na área de avaliação. Iniciou-se com a definição proposta por Kelley: *um teste é válido se mede o que pretende medir* (KELLEY, 1927). Apesar de esta definição ter sido publicada pela primeira vez há quase um século, ela ainda é muito utilizada (HAMP-LYONS, 2011; BEHIZADEH; ENGELHARD, 2011; YANCEY, 1999). O conceito de validade assim estabelecido é centro de muitas críticas. A principal delas é que, desse modo, a validade consiste em uma característica ou qualidade do teste, não levando em consideração o significado dos escores ou mesmo as consequências sociais e políticas do uso desses resultados (SCARAMUCCI, 2011).

Mais tarde, em 1955, Cronbach e Meehl escreveram um artigo que se tornou referência, no qual a validade não se refere apenas a uma propriedade do teste, mas também às interpretações da pontuação do teste. Está centrada na questão das relações entre os dados obtidos no teste e uma base teórica e observacional em uma rede que eles denominaram de *rede nomológica* (BORSBOOM; MELLENBERG; VAN HEERDEN, 2004).

Mais de 30 anos se passaram até que Messick em 1989 propôs o conceito de validade, considerado, hoje, o modo moderno de entender a validade:

> A validade é um julgamento avaliativo integrado do grau em que as evidências empíricas e teóricas apoiam a adequação e a qualidade das inferências e ações com base nos resultados de testes ou outros meios de avaliação.[1]

Assim, o novo conceito de validade consiste em saber se as interpretações e ações sobre os resultados dos testes são justificadas, tanto com base nas evidências científicas como nas consequências sociais e éticas da utilização de teste. Desse modo, a teoria da validade, gradualmente, passou a tratar todas as questões relacionadas aos testes e, ainda, a integrar todas elas em uma única definição (BORSBOOM; MELLENBERG; VAN HEERDEN, 2004).

Messick (1989) apresentou um quadro para representar a validade de testes, denominado por ele de "matriz progressiva" (Quadro 3), no qual dis-

[1] Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (MESSICK, 1989, p. 13, tradução nossa).

tingue duas faces da validade interligadas como um conceito unitário. Uma delas é a base de evidências dos testes que suportam o significado da pontuação e a utilidade da avaliação. A outra face é a base consequencial formada pelas interpretações dos resultados e das aplicações do teste.

Quadro 3 – Matriz progressiva: faces da validade

	Interpretação do teste	**Utilização do teste**
Base Evidencial	Validade de construto	Validade de construto + Relevância/Utilidade
Base Consequencial	Validade de construto + Implicação de valores	Validade de construto + Implicação de valores + Relevância/Utilidade + Consequências sociais

Fonte: Adaptado de Messick (1989)

As colunas da matriz "*Interpretação do teste*" e "*Utilização do teste*" representam os resultados dos testes. As linhas, "*Base Evidencial*" e "*Base Consequencial*", referem-se aos tipos de argumentos que devem ser usados para justificar os resultados de testes.

Essa matriz é constantemente citada em trabalhos como uma maneira de direcionar a obtenção das evidências para a validação de testes e, ao mesmo tempo, considerar os efeitos retroativos e a dimensão social dos testes que estão implicitamente inseridos nos tópicos *Implicação de valores* e *Utilidade*, que destacam as consequências sociais e o caráter cultural inseridos nos significados dos escores do teste (SCARAMUCCI, 2011). A matriz é denominada progressiva, porque sempre é acrescentada uma característica adicional à validade de construto.

Essa preocupação com as interpretações e a utilização dos resultados dos testes levanta a questão da legitimidade da incorporação das consequências do teste para a sua validação. Uma grande controvérsia consiste na questão da responsabilidade dos desenvolvedores do teste pela sua utilização. As discussões sobre esse assunto resultaram na criação de um novo termo para a validade, a *validade consequencial*, e também houve um aumento das discus-

sões sobre o *efeito retroativo* das avaliações. Mas uma grande dúvida reside no fato de essas discussões serem realmente legítimas, se as consequências e usos dos testes são preocupações que fazem parte das avaliações educacionais ou simplesmente estão situadas no âmbito político (PASQUALLI, 2007; BORSBOOM; MELLENBERG; VAN HEERDEN, 2004; LI, 2003; ALDERSON; BANERJEE, 2002).

Esse conceito de validade passou a ser aceito e utilizado por pesquisadores (SCARAMUCCI, 2011; MOSKAL; LEYDENS, 2000; CHAPELLE, 1999), inclusive pela *American Educational Research Association* (AERA), que estabelece no documento *AERA: Standards for educational and psychological testing*, "O grau em que as evidências e a teoria apoiam as interpretações dos resultados dos testes decorrentes das utilizações propostas para o ensaio" (AERA; APA; NCME, 1999).

Mesmo assim, ainda não foi estabelecido o consenso. Alguns pesquisadores consideram que a teoria de validade, assim estabelecida, torna o processo confuso para os responsáveis pela elaboração de testes, pois resulta em uma sensação de que tudo que se refere a todos os testes é relevante, fazendo com que as pessoas fiquem desnorteadas, sem senso de direção (PASQUALLI, 2007; BORSBOOM; MELLENBERG; VAN HEERDEN, 2004).

É por esse motivo e também pela simplicidade, por não depender de complicadas redes nomológicas e consequências sociais do teste, que, ainda hoje, o conceito formulado por Kelley (1927), "um teste é válido se mede o que foi proposto a medir", é utilizado e considerado correto por muitos pesquisadores de renome como Hamp-Lyons (2011), Behizadeh e Engelhard (2011), Pasqualli (2007), Borsboom, Mellenberg e Van Heerden (2004), Yancey (1999).

Borsboom *et al.* (2004) propuseram uma concepção de validade mais simples, que defendem ser teoricamente superior às posições existentes na literatura, mas, na verdade, é muito parecida com a definição de Kelley. Os autores estabeleceram que um teste é válido para a medição de um atributo se, e somente se, satisfaz as seguintes condições: (a) o atributo que se deseja medir existe? (b) as variações no atributo causalmente produzem variações nos resultados da medida? Justificam essa definição do seguinte modo: "Se alguma coisa não existe, então não se pode medir. Se ela existe, mas causalmente não produz variações nos resultados do procedimento de medição, a medida não é eficiente ou está medindo algo completamente diferente".

Pasqualli (2007) também não concorda com o conceito definido por Messick e afirma que "a validade diz respeito ao instrumento e não ao uso

que se faz dos seus escores". Explica que "não faz sentido dizer que um teste com validade de construto é válido numa situação, e não o é em outra".

Tradicionalmente, vários tipos ou visões de validade são enumerados na literatura e não há consenso quanto aos nomes, definições e métodos utilizados para a medida da característica pretendida (SCARAMUCCI, 2011; PASQUALLI, 2007; JONSSON; SVINGBY, 2007). Alguns exemplos são: validade de construto, validade de conteúdo, validade de face, validade de critério, validade concorrente, validade generalizável, entre outros.

O termo construto ou traço latente refere-se a aspectos do comportamento ou habilidades cognitivas que não podem ser observados ou medidos diretamente (DE AYALA, 2009). Exemplos de construtos são: inteligência, nível de ansiedade, atitude, grau de depressão, intensidade da dor de cabeça, habilidade em matemática, capacidade de expressão escrita, compreensão em leitura. Segundo Messick (1989), cada capacidade cognitiva envolve modelos, esquemas ou quadros mentais, e o desempenho das pessoas quanto a essas habilidades pode ter múltiplos caminhos corretos, cada qual dependente de outros conhecimentos e habilidades. Essas habilidades são de crescimento lento, difíceis de ensinar, de aprender e de medir. Desse modo, a validade de construto refere-se ao grau em que o instrumento de avaliação é capaz de distinguir os construtos que ele foi desenvolvido para medir, no caso da avaliação da escrita, o construto é a capacidade de redação ou de expressão escrita (BACHA, 2001).

A validade de conteúdo refere-se à amostra do conteúdo abordado no teste e se esse é relevante e representativo de todo o universo de conteúdo. A validade de face ou aparente consiste em se ter os conteúdos de um teste analisados por especialistas para determinar se eles são apropriados. Para a validade generalizável, os escores do teste devem ser generalizáveis para outras populações ou para um ensaio em outra data (JONSSON; SVINGBY, 2007). A validade de critério diz respeito a um critério externo ao teste (SCARAMUCCI, 2011; PASQUALLI, 2007). A validade concorrente ocorre quando os escores obtidos de testes diferentes, mas que foram elaborados para medir as mesmas habilidades, estão correlacionados positivamente (BACHA, 2001).

Para Moskal e Leydens (2000), um instrumento é validado pelo processo de acumulação de evidências que suportam a adequação das inferências feitas das respostas dos alunos ao serem avaliados em alguma habilidade específica. Normalmente, três tipos de evidências são requeridas para apoiar a validação de um instrumento de avaliação: a validade de conteúdo, a de construto e a de critério. Para Chapelle (1999), a validade consiste em um

conceito unitário, no qual a validade de construto ocupa um lugar central, isto é, a validade de conteúdo e a validade de critério podem ser utilizadas como evidências para estabelecer a validade de construto.

Jonsson e Svingby (2007) partilham da mesma opinião. Segundo eles, a validade refere-se à construção de um conceito unificador que incorpora os diferentes aspectos de validade. Todos esses aspectos são vistos como inter-relacionados, e todos devem ser considerados ao validar as avaliações a fim de obter uma imagem mais completa da validade.

O Quadro 4, elaborado por Chapelle (1999), traz um resumo comparativo entre o conceito tradicional e o conceito moderno de entender a validade.

Quadro 4 – Conceito tradicional × conceito moderno de validade

Passado	Presente
A Validade era considerada uma *característica do teste*: na medida que mede aquilo que pretende medir	A validade é considerada um *argumento* relativo à interpretação e ao uso: na medida que as interpretações e usos de um teste podem ser justificados
A confiabilidade era considerada distinta e uma *condição necessária para a validade*	A confiabilidade pode ser considerada como uma *evidência de validade*
A Validade era frequentemente estabelecida por meio de *correlações* entre testes	Validade é argumentada com base em um número de tipos de *justificativas e evidências*, incluindo as consequências da avaliação
A validade de construto era vista como um dos *três tipos de validade* (de conteúdo, relacionada a critério e de construto)	Validade é um *conceito unitário*, em que a validade de construto ocupa uma posição central (a validade de conteúdo e a relativa a critério podem ser usadas como evidência da validade de construto)
O estabelecimento da validade era uma tarefa de responsabilidade da avaliação, responsáveis pelo desenvolvimento de testes de larga escala e relevância	A justificativa de validade de um teste é de responsabilidade de *todos os usuários desse teste*

Fonte: Adaptado de Chapelle (1999)

Segundo Pasqualli (2007), os instrumentos de medidas são desenvolvidos com a finalidade de avaliar traços latentes (construtos). Desse modo, a qualidade do teste deve ser dada em relação à medida obtida do construto, objetivo da sua aplicação. As respostas ao teste, isto é, o escore, não cria

ou interfere no construto, pelo contrário, é o escore do teste que depende do construto. Assim, o construto deve ser o referencial para os resultados de um teste.

2.3.2 Confiabilidade

O conceito de confiabilidade está centrado na questão *se você pode medir de forma consistente* (YANCEY, 1999). Refere-se à consistência dos escores de avaliação, isso significa que é esperado que um indivíduo alcance o mesmo resultado, independentemente da ocasião em que respondeu ao teste, e também que as pontuações atribuídas por dois avaliadores a uma mesma resposta de teste não sejam demasiadamente diferentes.

Weigle (2002) define a confiabilidade como "*a consistência de medição através de diferentes características ou facetas de uma situação de teste, tais como os comandos ou avaliadores diferentes*". Em um teste confiável, o escore do indivíduo só pode variar se causado por fatores que não estão relacionados com o objetivo da avaliação (MOSKAL; LEYDENS, 2000).

A confiabilidade de um teste, na visão de psicometristas teóricos, como Guilford (1958) e Lord e Novick (1968), é definida como uma correlação entre o escore verdadeiro e o observado. Desse modo, a confiabilidade é dependente do conceito de erro de medição, uma vez que o escore observado é o escore verdadeiro acrescido de um erro (LI, 2003). O grau de erro de medição está inversamente relacionado com o grau de confiabilidade: quanto maior for o erro, menor é a confiabilidade do teste, e vice-versa, quanto maior for a confiabilidade do teste, menor deverá ser o erro de medição.

Um teste que possui confiabilidade zero pode ser caracterizado como tendo erro de medição total, enquanto um teste com confiabilidade perfeita não pode conter erros de medição. Inferências baseadas em resultados de testes com confiabilidade pequena ou zero terão pouco valor, porque as pontuações são resultados de medição com erro muito alto ou total. Por essa razão, a validade é dependente da confiabilidade (SLOMP, 2005).

Duas formas de confiabilidade normalmente são consideradas em avaliações: interavaliador, em que 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; MOSKAL; LEYDENS, 2000).

A confiabilidade interavaliador ou confiabilidade entre examinadores independentes (sem discussão ou colaboração) é considerada a característica

mais importante da avaliação escrita tradicional atual. Mesmo assim, é uma condição necessária, mas não suficiente para a validade. Isso significa que, sem um nível suficiente de acordo entre avaliadores, um procedimento de avaliação escrita não pode ser válido (HUOT, 1996). Por esse motivo, historicamente, a confiabilidade tem dominado a literatura sobre as avaliações com itens abertos, pois só após o desenvolvimento de procedimentos e critérios de pontuação e do treinamento de avaliadores é que este tipo de avaliação tornou-se psicometricamente viável. Antes disso, apenas medidas indiretas, provenientes de testes de múltipla escolha, poderiam ser suportadas pela Teoria Clássica dos Testes (HUOT, 1996).

2.3.3 Validade *versus* Confiabilidade

É frequente que o conceito de validade esteja relacionado com o de confiabilidade, pois essas duas características são consideradas importantes para estabelecer a qualidade de um teste.

Para que um teste seja válido, ele deve ser acima de tudo confiável, isso significa que, para ser válido, são necessários que sejam atribuídos escores precisos e de forma consistente. Tradicionalmente, os pesquisadores da área de avaliações consideram que a confiabilidade é essencial para se ter validade. Esse é um princípio fundamental de medição (MISLEVY, 2004), embora alguns pesquisadores defendam que o contrário pode ocorrer, quando se define confiabilidade de modo particular, como é o caso de Pamela Moss (1994), que escreveu um artigo polêmico de título "Pode haver validade sem confiabilidade?". Pesquisadores responderam à questão de Moss com uma resposta categórica, "NÃO". Essa resposta foi justificada com o esclarecimento da terminologia ambígua utilizada por Moss, confirmando a definição clássica de confiança como a medida, proveniente de um instrumento de medição, livre de erros (MISLEVY, 2004; LI, 2003). Brian Huot (1996) também concorda com essa posição afirmando que, sem um nível suficiente de confiabilidade, um procedimento de avaliação escrita não pode ser válido.

Por outro lado, um teste pode ser confiável e não ser válido. Por exemplo, pedir a um estudante que escreva toda a tabuada de cor não avalia se ele sabe e entende os procedimentos da multiplicação, embora a confiabilidade esteja garantida, pois será muito fácil para os avaliadores concordarem com a pontuação atribuída. Huot (1996) relata, como exemplo, um teste com o objetivo de medir a capacidade de escrita do estudante contando o número de palavras em cada ensaio. Esse teste poderia alcançar a confiabilidade per-

feita, uma vez que é possível a concordância entre os avaliadores do número de palavras contadas, embora dificilmente esse teste fosse válido para medir a capacidade de escrita do aluno. Um teste com esse objetivo deve conter aspectos teóricos relevantes da língua, caso contrário, não será válido (HUOT, 1996).

Li (2003) explica que, em muitas situações, tanto o instrumento de medição quanto os critérios de pontuação são compostos. Um instrumento de medição é frequentemente constituído por itens. O domínio de conteúdo utilizado para avaliar o construto e que determina o critério de correção é geralmente amplo e multifacetado. É concebível que cada item no instrumento de medida capture uma pequena região do domínio de conteúdo de outro item. Para que um instrumento de medição alcance alta validade, é preciso que ele cubra uma grande região do domínio, e a região do domínio que pode ser coberta por um dado número de itens pode ser aumentada com a diminuição de sobreposições. A correlação entre dois itens distintos do teste é uma indicação da sobreposição entre as regiões que eles cobrem.

Assim, para a psicometria, em termos de confiabilidade e validade, quanto menores forem as correlações entre os itens do teste, maior a validade do instrumento de medição. Desse modo, para a diminuição das sobreposições entre os itens, devem-se aumentar as especificidades dos critérios de pontuação, o que dificulta o treinamento dos avaliadores e a concordância entre suas notas.

Além disso, para promover maior confiabilidade, a resposta elaborada pelo participante da avaliação deve ser desenvolvida sob condições controladas, para uma tarefa uniforme, que deve ser pontuada com critérios detalhados. Como consequência dessa limitação da liberdade, a tarefa elaborada pelo participante não poderá cobrir grande parte do domínio, diminuindo, assim, a validade do teste.

Portanto, um aumento da confiabilidade leva a uma diminuição da validade e vice-versa. Aliás, desde a década de 1950, especialistas em avaliação escrita em larga escala vêm discutindo sobre o que deve ser favorecido nessas avaliações, a confiabilidade ou a validade dos testes? Uma vez que essas medidas são inversamente proporcionais, é necessário escolher qual entre as duas deve ser privilegiada em função das características de qualidade que se pretende no teste (HAMP-LYONS, 2011; SCARAMUCCI, 2011; SLOMP; FUITE, 2005; YANCEY, 1999; HUOT, 1996).

Tanto a validade como a confiabilidade são características importantes e desejáveis em toda avaliação escrita, embora cada uma delas seja defendida por uma corrente de pesquisadores convictos, e muitas vezes os adeptos de

uma das correntes são contra os adeptos da outra, como se apenas uma delas pudesse ser favorecida.

O embate entre os pesquisadores de avaliação escrita tem gerado muita discussão, e consequentemente algumas mudanças nas últimas décadas (YANCEY, 1999). Os psicometristas estabelecem uma influência estável a favor da confiabilidade, enquanto os pesquisadores da área da escrita têm proposto métodos alternativos de testes favorecendo a validade (SLOMP; FUITE, 2005; HUOT, 1996).

São frequentes as pesquisas focadas na melhoria da confiabilidade através do treinamento de avaliadores e da normatização (JOHNSTON, 2004; NYSTRAND; COHEN; DOWLING, 1993). No entanto, outra corrente de pesquisadores em avaliação escrita tem defendido que a diversidade de perspectiva, e não o consenso, é o que deve ser defendido e buscado. Criticam a suposição de que a escrita pode ser redutível a séries de estruturas simples de marcação (COLOMBINI; McBRIDE, 2012; JONSSON; SVINGBY, 2007; BROAD, 2000; YANCEY, 1999). Em sua pesquisa sobre a avaliação escrita no século XX, Yancey (1999) cita que Brian Huot explica esse embate como um pêndulo que balança entre as exigências de confiabilidade e de validade, um vai e vem de mudanças entre esses conceitos, cada hora com um desses campos dominando o outro, e que nos últimos anos tem havido uma tendência de domínio de ambos. Nesse trabalho, aproveitando a fala de Huot, Yancey define a história da avaliação escrita como um exercício de equilíbrio (balanço do pêndulo) entre o dueto dos conceitos de validade e confiabilidade.

Jonsson e Svingby (2007) também questionam as severas restrições feitas em nome da alta confiabilidade. Muitas vezes a proposta original da avaliação é desviada e ela deixa de medir aquilo para o qual foi projetada. Nesse ponto de vista, nas avaliações com itens abertos, a confiabilidade deve ceder o seu lugar de mais importante para a validade.

Para Li (2003), a confiabilidade e a validade são os dois conceitos mais fundamentais na psicometria, no entanto, às vezes, é difícil desenvolver uma discussão frutífera desses conceitos, devido à falta de congruência de seus significados na literatura. Para o pesquisador, é necessário que a comunidade educativa esteja ciente da existência de equívocos sobre os fundamentos da psicometria, e a proposta de uma convencionalização do significado de validade e de confiabilidade pode esclarecer algumas controvérsias suscitadas no passado e impedir que novas interpretações errôneas aconteçam no futuro.

2.3.4 Comparabilidade

A igualdade de oportunidades a todos os indivíduos pertencentes à população para a qual a avaliação foi projetada é essencial para haver justiça. Desse modo, a comparação entre indivíduos participantes de testes diferentes é uma vantagem almejada por pesquisadores da área da avaliação escrita. Os benefícios de comparação regional, nacional, entre instituições, ou mesmo entre nações são importantes e estão sendo incorporados aos exames em larga escala em muitos países.

A comparação entre avaliações da linguagem consiste, atualmente, em uma das questões de maior preocupação para os especialistas. Fato este provocado pela intensificação da utilização de matrizes comuns de referência desenvolvidas para orientar os currículos em todos os níveis de ensino em países da Europa, nos Estados Unidos, na Austrália, no Brasil, entre outros (HAMP-LYONS, 2004; NORTH, 2000).

A teoria de medição propõe metodologias clássicas de comparação entre avaliações, e esse processo de alinhar diferentes métricas é comumente denominado de lincagem, *linking* em inglês. Quando o objetivo desse alinhamento de métricas é comparar as pessoas, então o processo é denominado equalização. Assim, a equalização refere-se ao processo de ajustar a estimativa da habilidade dos indivíduos por diferentes métricas, e então transpor essas estimativas para uma métrica comum. A finalidade desse procedimento é proporcionar comparações individuais, que podem ser, por exemplo, entre os parâmetros dos itens, entre as habilidades dos indivíduos, entre as habilidades de um grupo, entre a habilidade de uma pessoa (ou grupo) em períodos de tempo diferentes (DE AYALA, 2009). Os procedimentos de equalização estão estabelecidos por várias metodologias da teoria clássica, mas a TRI oferece vantagens, e a equalização consiste em um conceito central de sua teoria.

No Brasil, o Sistema de Avaliação da Educação Básica (SAEB) utiliza uma escala única referenciada, e os diversos estados brasileiros mantêm a mesma matriz de referência, viabilizando a comparação entre os desempenhos dos estudantes em todo o território nacional. Mesmo assim, o SAEB não avalia o desempenho dos estudantes em testes com itens abertos, com exceção de algumas poucas edições específicas (KLEIN; FONTANIVE, 2009; VIDAL; FARIAS, 2008; BONAMINO; COSCARELLI; FRANCO, 2002; BRASIL/MEC). O ENEM possui um banco de itens equalizados, e as vantagens desse fato já foram comprovadas em várias situações, como, por exemplo, pela necessidade de oferecer o exame em outra data às pessoas privadas de liberdade. Eles respondem a provas diferentes, mas elas são elaboradas

com itens equivalentes, que medem os mesmos construtos.

Também podem ocorrer incidentes impedindo pessoas ou grupos de participar do exame na data marcada. Nesse caso, a equivalência entre as provas é garantida pela TRI, e consequentemente também é garantida a justiça quanto ao escore e à classificação dos participantes. Mas o ENEM, apesar de garantir a equivalência das provas objetivas, não oferece essa vantagem para a prova de redação, não podendo garantir, desse modo, um resultado totalmente justo aos participantes.

Na Europa, durante a década de 1990, houve o desenvolvimento de um número significativo de quadros comuns de referência para orientar os currículos e promover uma espécie de "perfil comum" da aprendizagem, tanto no âmbito institucional, como nacional, ou mesmo internacional. Exemplos de instituições europeias que utilizam "escalas comuns" em seus exames são: *University of Cambridge Local Examinations Syndicate* (UCLES), *Eurocentres* (uma fundação para o ensino de línguas europeias onde elas são faladas), *Association of Language Testers in Europe* (ALTE). O Reino Unido produziu um currículo nacional da língua inglesa para orientar o ensino nas escolas, e os britânicos, normas nacionais para qualificações relevantes da língua inglesa para o mundo do trabalho. Em toda a Europa, houve um aumento do número de projetos que buscam, de uma forma ou de outra, a comparação em avaliações da linguagem. A natureza multilíngue da população pertencente à União Europeia e o constante intercâmbio de pessoas motivadas por trabalho ou estudos levaram ao desenvolvimento de um quadro europeu comum de referência para as línguas (*Common Reference Levels*). O objetivo foi estabelecer escalas que representassem "normas comuns" ou pontos de referência para todas as línguas pertencentes à União Europeia (HAMP-LYONS, 2011, 2004; NORTH, 2000).

Essa abordagem de "normas comuns" não é muito comum fora da Europa. Nos Estados Unidos parece não haver padrões comuns, por exemplo, para a proficiência escrita em inglês esperada ao final do Ensino Médio, pois cada estado ou distrito desenvolve e implementa a sua própria avaliação (JEFFERY, 2009; HAMP-LIONS, 2004).

Estudos recentes destacam um anseio por parte dos pesquisadores para uma maior uniformidade nas normas acadêmicas dos EUA. Jill Jeffery (2009) constatou que os resultados das avaliações obrigatórias dos estados americanos apresentam discrepâncias sérias com os resultados da avaliação nacional na determinação da proficiência da escrita, uma vez que os estados apresentam níveis altos de proficiência em suas avaliações e o Governo Federal apresenta níveis baixos de proficiência em sua avaliação nacional para

o acompanhamento educacional (*National Assessment of Educational Progress - NAEP*). Essa discrepância é resultado de diferenças na forma como a proficiência da escrita é conceituada.

Em seu estudo, Jeffery comparou os construtos das avaliações de 41 estados e a nacional, NAEP, com o objetivo de conhecer como os testes variam nos caminhos para definir e medir a proficiência da escrita e entender a natureza e as implicações de tais variações. Destaca a ênfase dada à avaliação da escrita como um produto definido para avaliar a aprendizagem dos alunos em relação ao gênero das tarefas que pode variar entre escrever um artigo de opinião, uma carta argumentativa, uma resenha crítica, entre outros, e as funções retóricas que consistem nas variantes argumentativas para a comunicação em detrimento da avaliação de aspectos importantes como os processos metacognitivos que coordenam as aptidões cognitivas envolvidas na memória, leitura, compreensão de textos, entre outros.

O PISA (Programa Internacional de Avaliação de Alunos) avalia estudantes de 15 anos de idade, em mais de 50 países, em matemática, leitura, escrita e alfabetização científica em sua própria língua, mas não avalia a proficiência da escrita utilizando itens com respostas construídas em nenhuma das línguas dos países participantes (Hamp-Lyons, 2004; BRASIL).

No mundo inteiro, quando o assunto trata das avaliações oficiais em larga escala, há uma grande preocupação quanto à manutenção dos padrões da avaliação de uma edição para outra. Coe (2010) relata que, ao anunciar os resultados oficiais de uma avaliação, sempre fica a impressão de que seus padrões de qualidade foram diminuídos. Por um lado, se os índices obtidos pelos participantes melhoraram, esse fato é interpretado como evidência de que os itens estavam mais fáceis, indicando que o padrão do exame teve uma diminuição; por outro lado, se as notas dos participantes foram mais baixas, a avaliação pode não ter sido elaborada corretamente, indicando novamente que os padrões da avaliação caíram.

A comparabilidade diz respeito à validade das inferências sobre comparações que são feitas com base em resultados de avaliações. No Reino Unido, há um esforço por parte dos pesquisadores em busca de uma definição de comparabilidade que seja aceita e utilizada em todas as situações, uma vez que a comparabilidade é uma área cercada por suposições, por disputas metodológicas e considerada por alguns um terreno estéril, fadada ao fracasso. Aliás, os pesquisadores do Reino Unido se dedicam a estudos sobre a comparabilidade em uma gama de variedades: entre as dificuldades das diversas disciplinas básicas e também de diferentes línguas, entre professores, entre escolas, entre sistemas educacionais, entre grades curriculares incluindo as enti-

dades de certificação, entre indivíduos e ao longo do tempo (ELLIOTT, 2011; COE, 2010; POLLITT; AHMED; CRISP, 2007; NEWTON, 2007, 2008).

Para Pollitt, Ahmed e Crisp (2007), não há uma definição de comparabilidade assumida universalmente, os pesquisadores normalmente fazem algumas suposições dos padrões que são esperados nos dois exames que estão sendo comparados. No entanto é razoável a existência de um certo equilíbrio entre as diferenças de demandas dos exames. Por exemplo, um dos exames pode ter sido desenvolvido com um tratamento mais profundo sobre um domínio menor de conteúdos, enquanto o outro, sobre um domínio de conteúdos maior, mas, de maneira geral, sem muita especificidade. Os avaliadores devem garantir de alguma forma um equilíbrio entre essas exigências.

O fato é que os termos "comparabilidade", "dificuldade" e "padrão", muitas vezes, são usados com objetivos diferentes, mas essas diferenças nem sempre são devidamente destacadas. No campo das avaliações, quando se fala em comparabilidade, o pensamento incide sobre os processos pelos quais os resultados dos testes são traduzidos em normas ou padrões interpretáveis. Diferentes definições de comparabilidade podem ser encontradas na literatura, assim como são várias as técnicas para a sua determinação ou acompanhamento, mas basicamente três abordagens são comumente utilizadas para julgamento do padrão especificado para os exames: (1) em termos de critérios de desempenho que considera apenas as características do teste ou; (2) em termos de normas estatísticas que leva em conta o desempenho dos examinandos de uma população e; (3) em termos do desempenho em relação ao construto comum (COE, 2010). A ideia de um estudo empírico definitivo sobre a comparabilidade é um desafio que dependerá fortemente da definição adotada para apoiar a validade da técnica particular utilizada para o seu monitoramento (COE, 2010; POLLITT; AHMED; CRISP, 2007).

Assim como o termo "comparabilidade" pode ser definido teoricamente de maneiras diferentes, os termos "dificuldade" e "padrão" também podem ter significados variados. Afirmações daqueles que consideram que a avaliação estava mais fácil do que as anteriores podem não contrapor àquelas que consideram que a avaliação está ficando mais difícil. Essas afirmações devem ser apoiadas em uma mesma base de significados, caso contrário, estarão denotando coisas completamente diferentes (COE, 2010).

Pollitt, Ahmed e Crisp (2007) consideram que a comparabilidade não é apenas uma questão da capacidade dos alunos e da dificuldade das perguntas, deve-se prestar atenção às exigências abordadas nas questões e na natureza do construto que está sendo avaliado. As análises estatísticas podem indicar que duas notas provenientes de dois exames são igualmente difíceis de al-

cançar, mas não podem garantir que essas notas são resultantes de exigências equivalentes sobre os desempenhos dos alunos.

Newton (2008) estabelece uma taxonomia de definições sobre a comparabilidade por meio de três perspectivas distintas, nas quais a comparabilidade sempre é definida em termos de um perfil de realizações do estudante que está associado com um determinado padrão ou grau.

Em primeiro lugar, está a comparabilidade sob uma perspectiva de realizações que possuem as mesmas características, propriedades ou disposições. Aqui, a comparabilidade diz respeito ao conhecimento que os alunos com classificações semelhantes têm em comum.

Em segundo lugar, existe a comparabilidade a partir de uma perspectiva de causalidade, em que as realizações possuem os mesmos antecedentes. Nesse caso, a comparabilidade diz respeito aos fatores que resultaram na realização, as oportunidades para a aprendizagem que os alunos com classificação semelhantes têm em comum.

Em terceiro lugar, a partir de uma perspectiva de previsão, as realizações são as mesmas no sentido das perspectivas que elas oferecem. A comparabilidade diz respeito ao potencial que está implícito na realização; a probabilidade de sucesso futuro que os alunos com classificações semelhantes têm em comum.

Embora existam diferentes formas e concepções de comparabilidade, como as apresentadas por Newton (2008), elas geralmente podem ser enquadradas em uma das três concepções principais para a comparabilidade citadas anteriormente. A primeira considera os critérios de desempenho, a segunda é apoiada sobre resultados estatísticos e a última considera o desempenho em relação ao construto comum.

As discussões sobre a comparabilidade estão sendo intensas atualmente, em especial nos países europeus. Existem controvérsias sobre as possíveis diferenças entre as pessoas em relação às dificuldades de exames das diversas disciplinas e também entre as disciplinas. São muitos os estudos que tentam estabelecer se as pessoas consideram a química mais difícil do que o inglês, por exemplo.

Particularmente, quanto à comparação entre a dificuldade das tarefas de escrita, um tema que tem recebido muita atenção, os estudos disponíveis mostram que as discussões têm sido altamente controversas.

As perguntas mais comuns acerca dos testes de escrita referem-se ao significado dos termos: “padrões”, “dificuldade” e “comparabilidade”, mas os pesquisadores ainda não encontraram respostas decisivas. Talvez pela grande quantidade de variáveis demandadas para o estabelecimento dos tes-

tes e que influenciam diretamente o desempenho dos estudantes. Por esse motivo, as pesquisas geralmente são restritas e consideram apenas algumas dessas variáveis ou grupos específicos de respondentes.

Hamp-Lyons e Mathias (1994) com o intuito de determinar a dificuldade de *prompts* utilizados em um teste de proficiência em inglês para estrangeiros, semelhante ao TOEFL, submeteram esses *prompts* a avaliadores especialistas, partindo da hipótese de que esses avaliadores pudessem concordar entre si com o julgamento da dificuldade dos tópicos em uma escala de 3 pontos, fácil, médio e difícil. Chegaram a um nível de concordância razoável, mas que ainda poderia ser melhorado com o treinamento dos avaliadores em um conjunto claro de diretrizes. Uma segunda hipótese consistiu em determinar a relação entre a pontuação da dificuldade atribuída pelos avaliadores aos tópicos e as pontuações reais recebidas pelos candidatos nesses tópicos, e constataram, com certa surpresa, que os tópicos julgados de maior dificuldade nem sempre correspondiam a um pior desempenho dos examinandos.

Uma experiência parecida foi realizada por Breland *et al.* (2004) para estudar a comparabilidade de *prompts* utilizados em avaliações do TOEFL baseado em computador, para determinar habilidades de escrita em inglês como língua estrangeira. Nesses testes, a análise da comparabilidade dos *prompts* é de suma importância, uma vez que cada examinando recebe apenas um único tópico, que não é o mesmo para todos os participantes da prova. Se os tópicos não são equivalentes quanto à dificuldade, os candidatos submetidos aos mais difíceis seriam desfavorecidos, ao contrário dos submetidos aos mais fáceis, pois esses seriam favorecidos. Procedimentos estatísticos foram utilizados para estimar diferenças de gênero e dificuldade de 47 *prompts* na primeira fase da pesquisa e de 87 na segunda. Todos esses *prompts* já haviam sido administrados no programa. Alguns *prompts* selecionados também foram revisados por especialistas e uma taxonomia de características foi estabelecida para relacionar as diferenças de gênero e a dificuldade dos tópicos. Algumas recomendações para procedimentos de controle para identificar *prompts* menos comparáveis também foram feitas.

Sudweeks, Reeve e Bradshaw (2005) propuseram a alunos de graduação matriculados em um curso de história duas tarefas de escrita sobre temas relevantes da história mundial, que faziam parte do programa da disciplina. Os objetivos da pesquisa consistiram em determinar o grau de confiabilidade das pontuações e detectar se a ocorrência de erros nas pontuações foi devido a inconsistências interavaliador (avaliadores distintos julgam um mesmo ensaio) ou intra-avaliador (um único avaliador julga o mesmo ensaio em ocasiões distintas), determinar a diferença entre as dificuldades das tarefas pro-

postas aos alunos, e analisar os resultados referentes às interações entre essas variáveis. Confrontaram os resultados obtidos com a utilização da teoria de generalização (G-teoria) e uma extensão do modelo de Rasch, o modelo multifacetas de Rasch, como meio de alcançar esses objetivos.

Participaram do ensaio 497 alunos de graduação, em várias sessões de um curso de história. Apesar de as duas tarefas estabelecidas aos alunos terem sido concebidas como indicadores de equivalência da dificuldade, diferenças significativas entre as dificuldades das tarefas foram detectadas pelos dois métodos de análise dos dados, isto é, as tarefas não foram consideradas igualmente difíceis pelos alunos.

Huang (2008), também fazendo uso da teoria da generalização, identificou fatores nos quais os examinandos perceberam que algumas tarefas são aparentemente mais difíceis do que outras, como o conhecimento e o interesse sobre o tema, experiências de vida relacionadas ao tema e a disponibilidade de dados. Sua pesquisa consistiu em determinar as diferenças entre a variabilidade de classificação e confiabilidade das pontuações atribuídas à escrita de estudantes de língua nativa não inglesa *versus* estudantes de língua nativa inglesa em um exame provincial em larga escala no Canadá. Foram utilizados dados de três anos consecutivos a fim de completar as análises e verificar a estabilidade dos resultados. As análises sugerem que as tarefas de escrita foram, em média, comparáveis em termos do grau de dificuldade, mas não foram consideradas igualmente difíceis para todos os estudantes de ambos os grupos linguísticos.

Pomplun *et al.* (1992) analisaram o desempenho dos alunos no teste de composição de inglês do College Board (*English Composition Achievement Test* – ECT) que é oferecido todos os anos como parte das provas de admissão a essa escola. Esse teste tem o objetivo de avaliar a capacidade de escrita do candidato. Embora os organizadores da prova se esforçassem para desenvolver tópicos semelhantes em dificuldade de um ano para outro, não há garantias de sucesso absoluto. Desse modo, o objetivo do estudo foi investigar o desempenho, ao longo do tempo, de subgrupos estabelecidos por sexo, raça, cor, nacionalidade, língua materna, entre outros, para determinar o quanto a aplicação de tópicos diferentes resulta em padrões diferentes de desempenho. A dificuldade diferencial foi explorada por meio de comparação entre os grupos de referência que realizaram tanto as provas escritas quanto as provas objetivas. As diferenças foram consistentes em todos os sete anos estudados, indicando que os desenvolvedores dos testes tiveram sucesso na tentativa de elaborar comandos e temas que não apresentem tendência de privilegiar grupos específicos, no entanto dois ensaios foram identificados por

conter características relacionadas com o desempenho diferencial dos estudantes.

Bridgeman *et al.* (2011) introduziram seis tipos de variantes diferentes, a partir do mesmo comando, em uma avaliação americana em larga escala, com a finalidade de aumentar o conjunto de estímulos disponíveis para o texto, tornando a tarefa de escrita menos previsível. Foram analisados dados provenientes de 7.573 ensaios na tentativa de respostas às seguintes questões: (1) As distribuições da pontuação, como as médias e a dispersão, são comparáveis entre os *prompts* e as variantes? (2) Existe algum tipo de variante diferencialmente difícil para determinados subgrupos, como sexo, etnia, ou para examinandos cuja melhor língua não é o inglês? A confiabilidade das correções é consistente para todos os tipos de variantes?

Os resultados foram razoavelmente semelhantes para todos os tipos de variantes, sugerindo que a estratégia das variantes pode ser utilizada para as avaliações em larga escala com questões abertas. Não foram notadas diferenças quanto à dificuldade das variantes sobre nenhum dos subgrupos analisados. Além disso, os pesquisadores sugerem que a utilização de variantes aumenta potencialmente a validade através da redução do uso de materiais pré-memorizados por parte dos candidatos, além de reduzir os custos na elaboração dos testes.

Uma tendência atual nas avaliações da escrita é a incorporação de tarefas integradas para provocar os escritores a incorporar múltiplas fontes de ideias para a criação de seus textos. A utilização dessas tarefas integradas, além de agregar muitas vantagens, oferece a possibilidade de melhorar a equidade de testes e a validade de construto devido às suas naturezas multifacetadas.

Yang (2012) examina em sua pesquisa o desempenho de estudantes da área da saúde em um teste para avaliar a capacidade de escrita em inglês como segunda língua, cujos itens exigiam leitura de gráficos e textos. Foram desenvolvidos três instrumentos para a avaliação, um inventário para o estudante estabelecer quais estratégias foram utilizadas na leitura gráfica, uma tarefa destinada a avaliar o desempenho da escrita e um gráfico de rubricas projetado para o avaliador registrar as pontuações alcançadas pelo estudante. As questões de pesquisa abordadas são em relação à natureza das estratégias utilizadas para a leitura gráfica e a relação entre o uso das estratégias e o desempenho no teste. As análises indicam que os alunos escritores estavam empenhados na utilização de estratégias para a compreensão e interpretação de gráficos, e isso teve um impacto positivo no desempenho nos testes. Algumas evidências para apoiar a validade de construto da tarefa foram obtidas,

mas não foram suficientes para determinar o uso desse tipo de tarefa.

Também existem estudos com a finalidade de avaliar a diferenciabilidade nas tarefas devido a algumas variáveis, como o comprimento das tarefas ou a utilização ou não de experiências pessoais nos ensaios. Kobrin *et al.* (2011) analisaram a relação existente entre o comprimento da redação e a pontuação no ensaio e também as pontuações alcançadas pelos alunos que utilizaram exemplos acadêmicos em comparação com aqueles que usaram experiências pessoais como exemplo. Os resultados mostraram que o comprimento do ensaio está relacionado com a pontuação, mas a correlação não é tão alta como outros estudos afirmaram, como exemplo ver Worden (2009) e Penny (2003).

Lee e Anderson (2008) examinaram a validade e a generabilidade do tema de um teste com itens abertos projetado para classificar estudantes estrangeiros em cursos apropriados de inglês como segunda língua em uma grande universidade americana. O teste consistiu em um sorteio entre três temas acadêmicos integrados, com a disponibilidade de fontes extras sobre o tema sorteado para leitura e escuta antes de a tarefa escrita ser desenvolvida. Para determinar a comparabilidade dos temas, variáveis explicativas foram identificadas nos ensaios escritos pelos alunos e também foram utilizadas as pontuações alcançadas por eles no Teste de Inglês como Língua Estrangeira (TOEFL) como controle da proficiência geral em inglês.

Os resultados indicam que a proficiência dos alunos estabelecida pelo teste TOEFL não foi relacionada aos seus desempenhos na avaliação escrita, no entanto as análises estatísticas indicam que tópicos diferentes afetam o desempenho do estudante. Quanto à validade do teste, o estudo estabeleceu argumentos indicando a não comparabilidade entre os três temas, mas os autores apoiam a generalidade de tópicos, quando esses são utilizados por examinandos pertencentes a uma ampla gama de áreas disciplinares.

Situado no contexto de um teste de classificação para estudantes não nativos da língua inglesa, Lee e Anderson (2008) exploraram a relação entre o desempenho do aluno em duas tarefas de escrita. Uma delas consistiu em escrever sobre um tema geral comumente estabelecido nos testes, e a outra, sobre um tema específico da área acadêmica do aluno. O pesquisador não encontrou em seus resultados diferenças significativamente mensuráveis entre as duas tarefas, concluiu que os examinandos não obtiveram benefícios reais em seus desempenhos na opção por uma ou outra tarefa e propôs uma discussão mais aprofundada da necessidade ou não de usar um conjunto múltiplo de tarefas no lugar de uma única solicitação em teste escrito.

Pagano *et al.* (2008) conceberam um projeto para o desenvolvimento

de um instrumento padrão para avaliação do desempenho da escrita interinstitucional, no qual várias instituições americanas de ensino superior pudessem conduzir e avaliar seus programas de composição de modo a possibilitar comparações entre os desempenhos de seus alunos e permitir a colaboração entre as instituições nas decisões para a melhoria do ensino dessa disciplina. A tarefa escolhida para a avaliação consistiu em respostas referentes a um texto, pois essa tarefa, bastante comum nas avaliações da escrita, se aplicaria de modo geral a todos os programas das instituições participantes, uma vez que a leitura que antecede a escrita de um texto, fornecendo subsídios para a elaboração de um material crítico ou analítico, é uma tarefa comum e importante no âmbito acadêmico. O termo "texto" é entendido em um contexto amplo, podendo ser uma crônica, uma reportagem, um livro, uma série de artigos, um ambiente, uma cidade, ou até mesmo uma obra de arte.

A comparabilidade tem sido presente no sistema de ensino inglês desde o início do século XX, mas nas últimas décadas têm-se buscado formas de garantir que os padrões de ensino para a educação básica e consequentemente de seleção para as universidades se mantenham constantes de ano para ano. À medida que os exames se tornaram mais competitivos, as exigências de consistência nas pontuações e de comparabilidade provocaram na Inglaterra uma série de iniciativas, como um sistema nacional de currículo e avaliação, utilização de exames em larga escala com o intuito de avaliar a eficácia dos sistemas de ensinos, como um todo ou em escolas e faculdades isoladas, regulação do sistema de ensino e exames por legislação específica, entre outros. Com o início do século XXI, a comparabilidade entre provas acadêmicas e profissionais tornou-se uma questão urgente e fundamental tanto para a elevação do nível de aprendizagem como para a avaliação do desenvolvimento de habilidades necessárias para o século XXI. A modernização do sistema de exames está sendo desenvolvida e a comparabilidade consta como um requisito essencial do sistema de exames da Inglaterra (TATTERSALL, 2007). Visando atender a essa demanda por comparabilidade, muitos trabalhos estão sendo desenvolvidos por pesquisadores ingleses e pelas instituições promotoras de avaliações na Inglaterra, visando uniformizar o significado e a terminologia de comparabilidade (ELLIOT, 2011, 2013; BAIRD, 2007; NEWTON, 2007, 2008; POLLITT, 2007), a dificuldade dos exames e métodos estatísticos (ELLIOT, 2013; COE *et al.*, 2008), normas e padrões comuns (COE, 2010; BAIRD, 2007), análises históricas das metodologias, métodos e definições sobre comparabilidade (TATTERSALL, 2007; ELLIOTT, 2011), entre outros inúmeros exemplos.

Um procedimento comumente utilizado em avaliações em larga es-

cala, que assegura a comparabilidade entre administrações distintas e torna possível os estudos longitudinais, consiste em manter alguns itens repetidos de uma avaliação para a outra. Para avaliações objetivas, esse processo é comum e possui teorias bem estabelecidas. No entanto, para as avaliações de respostas construídas, esse procedimento é mais difícil, pois normalmente os testes são compostos por itens mais complexos e em menor número do que aqueles com itens objetivos. Além disso, eles são mais fáceis de memorizar, por isso repetir alguns desses itens pode depor contra a segurança do teste (HAERTEL; LINN, 1996).

Frederiksen e Collins (1989) não têm essa preocupação com a segurança em avaliações com itens abertos. Ao contrário, defendem que essas avaliações devem ser conhecidas, que todos devem saber exatamente o que é esperado, como em algumas competições esportivas, por exemplo, as modalidades de ginástica olímpica, em que todos os interessados sabem o que será pontuado e quais são os padrões de excelência. Mesmo assim, conhecer as respostas esperadas e os critérios de pontuação em uma avaliação da escrita não isenta a necessidade de tarefas diferentes em cada administração da avaliação, pois, caso contrário, a avaliação não seria válida. Os alunos facilmente poderiam decorar ou mesmo copiar a resposta previamente escrita, inclusive por terceiros.

Outras variáveis que podem prejudicar a comparabilidade entre avaliações dizem respeito à população-alvo. O teste deve ser desenvolvido levando-se em conta fatores como idade, cultura, grau de instrução, entre outros. As informações sobre os respondentes devem influenciar as decisões sobre o conteúdo avaliado, o formato das tarefas, a apresentação do teste, o tempo destinado à resposta. Uma avaliação desenvolvida para ser aplicada a uma população específica não pode ser comparável, se aplicada a outra população, com formação ou idade diferente. Isso ocorre porque o desempenho na elaboração de uma determinada tarefa depende de capacidades ou atitudes que não são explicitamente declaradas no construto que está sendo medido, mas são habilidades necessárias para a elaboração da tarefa, essas são denominadas habilidades auxiliares ou complementares. Se algum grupo de examinandos for deficiente em habilidades auxiliares de um teste, eles não vão responder como outros examinandos que possuem o mesmo grau de proficiência em relação ao construto que o teste foi desenvolvido para avaliar. Também devem ser considerados fatores diversos associados à ocasião em que o teste foi aplicado e as interações entre essas variáveis. Todas essas fontes, aparentemente externas ao teste propriamente dito, são responsáveis por erros de medição e causam variabilidade na classificação dos participan-

tes (SUDWEEKS; REEVE; BRADSHAW, 2005; LI, 2003; WEIGLE, 2002; MOSKAL; LEYDENS, 2000; HAERTEL; LINN, 1996).

Os desenvolvedores de avaliações em larga escala devem se preocupar também com uma série de deficiências físicas, muitas vezes consideradas irrelevantes, mas que podem causar dificuldades no desenvolvimento de determinadas tarefas por algumas pessoas, prejudicando a avaliação. Por exemplo, o daltonismo ou alguma outra deficiência física ou motora (HAERTEL; LINN, 1996).

As habilidades auxiliares representam uma grande ameaça à validade das avaliações. A dependência dessas habilidades para a elaboração de tarefas é mais evidente nas avaliações com respostas construídas do que nos testes objetivos, pois tanto as instruções e outros materiais fornecidos na questão, quanto as respostas esperadas, são mais complexas, demandando a utilização de habilidades complementares. Para a avaliação da habilidade de escrita, por exemplo, a leitura e a correta interpretação da tarefa são habilidades auxiliares sempre exigidas.

Haertel e Linn (1996) descrevem três componentes que devem ser considerados para analisar a comparabilidade de tarefas isoladas em avaliações com itens de respostas construídas: 1) a intenção da medição do construto a ser avaliado; 2) o conjunto de demandas complementares que a tarefa exige, como os requisitos de conhecimento, habilidades e disposições necessárias para a elaboração da tarefa; e 3) a variância de erro que é resultado de uma mistura complexa de influências na pontuação. A comparabilidade das tarefas é resultante das semelhanças e diferenças entre essas três variáveis. Pode não ser fácil ou mesmo possível separar essas variáveis para determinar o grau em que esses aspectos da tarefa diferem.

A seguir, são feitas análises sobre a comparabilidade de tarefas distintas quando esses três componentes variam conforme apresentado por Haertel e Linn (1996).

1. **Comparabilidade entre as tarefas com a mesma intenção, os mesmos requisitos auxiliares e as mesmas estruturas de erro.**

   Se tarefas diferentes são propostas para a avaliação de um mesmo construto, requerem as mesmas habilidades complementares e as pontuações resultantes são igualmente precisas para avaliar indivíduos em qualquer nível de desempenho dessas habilidades, as pontuações produzidas devem ser comparáveis. Essa é a forma de lincagem mais forte definida por Mislevy (1992) e também por Linn (1993). Considere duas tarefas quaisquer, por exemplo, X e Y, então, deve ser possível encontrar uma única

função de equalização para transformar as pontuações atribuídas à tarefa X para a tarefa Y e reciprocamente.

Fontanive *et al.* (2010), em um projeto visando avaliar as habilidades de leitura, escrita e matemática alcançadas pelos alunos dos dois primeiros anos do Ensino Fundamental do Rio Grande de Sul, desenvolveram escalas únicas de proficiência nessas disciplinas por meio da TRI para expressar os desempenhos dos alunos. Para comparar os resultados das provas aplicadas ao final de cada um dos dois anos, fizeram as seguintes hipóteses para a "definição" de itens comuns: em leitura e escrita, foram considerados comuns cinco itens de resposta construída que tinham os mesmos critérios de pontuação; em matemática, consideraram-se comuns três itens de múltipla escolha muito semelhantes entre si quanto à habilidade medida e quanto ao conteúdo.

Essa definição para a comparabilidade dos itens ajusta-se para exemplificar essa abordagem de lincagem, pois os itens para a avaliação da leitura e da escrita tiveram a mesma intenção e mediram os mesmos construtos, exigindo as mesmas habilidades auxiliares e também foram pontuados com os mesmos critérios de pontuação. Em matemática, os itens são objetivos, assim não há variabilidade na pontuação e objetivam medir os mesmos construtos.

2. Comparabilidade entre as tarefas com a mesma intenção, os mesmos requisitos auxiliares, mas estruturas de erro diferentes.

Quando duas tarefas medem os mesmos construtos e demandam as mesmas habilidades complementares, mas as pontuações resultantes possuem diferentes graus de precisão, elas satisfazem os requisitos para a comparação. Mislevy (1992) define esse tipo de comparação, como a segunda forma mais forte de lincagem. Para exemplificar, considere uma avaliação da habilidade da escrita com duas tarefas diferentes, tarefa X e tarefa Y, que avaliam os mesmos construtos e demandam as mesmas habilidades auxiliares. Suponha que a correlação entre as pontuações das duas tarefas seja tal que possibilite considerá-las como paralelas.

No entanto, pode haver diferenças substanciais na tarefa ou na atribuição dos escores que podem causar variações sérias na classificação dos candidatos. Por exemplo, se as especificações dos comandos das tarefas não são igualmente claras; se os critérios utilizados para a pontuação de uma das tarefas são mais específicos do que os da outra, ou, ainda, se os pontos de ancoragem e seus exemplos não são igualmente bem escolhidos

para as duas tarefas[2], essas diferenças podem causar erros na classificação dos candidatos, pois um indivíduo com maior capacidade poderia alcançar uma pontuação mais alta se respondesse a uma das tarefas e vice-versa. Um procedimento importante nesse caso consiste em examinar as diferenças na precisão das pontuações das tarefas individuais, escolhendo um método estatístico adequado, e realizar algumas reflexões sobre quais diferenças são realmente importantes.

Exemplos de pesquisas que se enquadram nesse caso são frequentes em estudos que investigam as relações entre pontuações automatizadas e pontuações feitas por avaliadores humanos. Nessas pesquisas, as tarefas são estabelecidas com as mesmas intenções, os requisitos auxiliares são os mesmos, mas as estruturas de erros são diferentes (FAZAL *et al.*, 2013; WEIGLE, 2013; DEANE, 2013; RAMINENI, 2012; WILLIAMSON *et al.*, 2012).

Outras pesquisas que se enquadram como exemplos nesse caso são as que fazem comparações entre os tipos de pontuações utilizadas. Por exemplo, Weigle (2002) compara o tipo de pontuação para a utilização em avaliação da escrita em inglês como segunda língua e Barkauoui (2011) compara os critérios para serem utilizados nas avaliações diagnósticas.

3. Comparabilidade entre as tarefas com a mesma intenção, os requisitos auxiliares diferentes e as mesmas estruturas de erro.

Se as tarefas para a avaliação do desempenho de algum construto diferem na exigência de habilidades auxiliares, então as pontuações a essas tarefas serão comparáveis somente se a população de respondentes tiver essas habilidades auxiliares plenamente desenvolvidas e possuírem domínio completo sobre elas. Do mesmo modo, uma única tarefa que depende de habilidades auxiliares, aplicada a dois grupos, um que possui essas habilidades e outro que não, será tendenciosa e favorecerá o primeiro grupo.

Exemplos de tarefas com exigências de habilidades auxiliares diferentes são obtidos em avaliações aplicadas a grupos de línguas nativas diferentes ou a grupos de alunos com diferentes histórias de instrução.

Como exemplo, Huang (2008, 2012) avaliou as diferenças entre a variabilidade nas classificações e a confiabilidade das pontuações atribuídas à escrita em língua inglesa de estudantes nativos e não nativos em países de língua inglesa.

[2]Esses pontos são selecionados na escala de habilidades e relacionados com descritores que ilustram as variações dos critérios de pontuação.

Outro tipo de pesquisas que se enquadra nesse item são os estudos sobre a comparação entre indivíduos que elaboraram tarefas escritas com a utilização de papel e lápis e indivíduos que utilizaram processadores de texto em computadores (LEE, 2004).

Knoch e Elder (2010) exploraram a variação do tempo disponível para alunos universitários escreverem ensaios. Compararam o desempenho de examinandos em testes com duração de 30 e de 50 minutos, com o objetivo de verificar se as restrições de tempo podem influenciar a avaliação de habilidades de escrita necessárias em contextos acadêmicos.

4. **Comparabilidade entre as tarefas com a mesma intenção, diferentes requisitos auxiliares e diferentes estruturas de erros.**

   Nesse caso, as tarefas construídas para serem intercambiáveis são, na melhor das hipóteses, apenas aproximadamente paralelas. Segundo Haertel e Linn (1996), não existem "regras de ouro" sobre a intercambialidade de tarefas construídas de maneiras diferentes. Essa possibilidade é apenas fictícia.

2.3.5 Justiça

Uma avaliação de qualidade deve permitir aos participantes condições de respostas que assegurem inferências corretas sobre seu desempenho em relação ao construto medido. Quando os testes são administrados para populações diversas, como nas avaliações em larga escala, as especificações que assegurem a validade do teste são mais difíceis de serem alcançadas, assim como é mais difícil a obtenção de medidas precisas sobre os conhecimentos e as competências dos respondentes (JOHNSTONE *et al.*, 2008). As questões sobre justiça estão relacionadas com a equidade do teste, ou a possibilidade de garantir a todos os participantes oportunidades iguais, e, para isso, é necessário que os testes sejam imparciais e apropriados para os vários grupos que serão testados.

A AERA (*American Educational Research Association*) e a APA (*American Psychological Association*) estabelecem que todos os examinandos devem ter oportunidade de demonstrar a sua posição na escala de habilidades em relação ao construto que o teste é concebido para medir. A validade do teste depende dessa oportunidade dada aos participantes da avaliação que, nesse contexto, está relacionada principalmente aos itens.

Downing e Haladyna (1997) estabeleceram algumas evidências

quanto à elaboração e revisão dos itens para que estes sejam considerados válidos. Essas evidências devem ser observadas em todas as etapas para a elaboração da avaliação: especificações do teste, definição do conteúdo, treinamento adequado para os autores dos itens, classificação das habilidades que serão testadas, redação do item dentro de princípios preestabelecidos seguidos por procedimentos de revisão dos itens que consistem em análises sobre o conteúdo utilizado, na forma, e a edição dos itens, no vocabulário, no tema proposto, na adequação do item para a população-alvo e terminando com a eliminação de itens inadequados ou mal formulados, além de revisões de procedimentos para a segurança do teste.

Embora alguns princípios para a elaboração de itens para avaliações sejam conhecidos e divulgados na literatura (ANASTASI, 1977; PASQUALLI, 2010; VIANNA, 1982), os efeitos causados por itens nas avaliações não são muito difundidos nas pesquisas científicas, mas é certo que itens mal formulados prejudicam a validade do teste. Downing (2002) constatou que itens com problemas na formulação foram considerados, pelos alunos, mais difíceis do que os itens bem formulados, isso para avaliar o conhecimento sobre um mesmo conteúdo. O pesquisador considera que a qualidade do item pode ameaçar a validade do teste, especificamente em relação à variância construto-irrelevante, que ocorre quando o teste mede além dos conhecimentos e habilidades que se pretendem medir, medindo também traços subjacentes ao construto e que não são importantes para o objetivo do teste, prejudicando as inferências sobre os resultados da avaliação. Os testes em larga escala devem ser sistematicamente elaborados e revisados para detecção, correção ou remoção dos itens considerados problemáticos.

Há uma linha de pesquisas na área de avaliação em larga escala que defende o "*Design* Universal de Avaliação" (*Universal Design of Assessment*– UDA, em inglês). São avaliações concebidas e desenvolvidas desde o início para permitir a participação da maior variedade possível de participantes e resultar em inferências válidas sobre o desempenho de todos que participam na avaliação (JOHNSTONE *et al.*, 2008).

O termo *design* universal é proveniente da arquitetura e sua base filosófica defende que as edificações devem prever acesso aos portadores de deficiências desde o projeto inicial. Recursos como rampas, elevadores, portas alargadas, banheiros especiais, entre outros, devem ser previstos e construídos durante a obra, para que não sejam necessárias adaptações após a conclusão do edifício, resultando em soluções menos eficientes e tornando os custos maiores. O *design* universal emergiu como um conceito interdisciplinar e a característica principal é promover o acesso aos ambientes, à aprendizagem, à

avaliação, entre outros. Testes desenvolvidos a partir de um quadro de *design* universal, além de medir o que foi proposto medir, devem conter viés mínimo e suas instruções e procedimentos estabelecidos de maneira clara e compreensível a todos os participantes. Por sua base filosófica ser ampla, o *design* universal pode incluir uma variedade de estratégias para auxiliar no entendimento de quais variáveis podem afetar o desempenho dos participantes da avaliação (JOHNSTONE *et al.*, 2008).

O principal objetivo do *design* universal é melhorar a validade do teste e, consequentemente, melhorar a avaliação. Deficiências podem prejudicar a capacidade do aluno de demonstrar o seu conhecimento sobre algum construto avaliado nos testes. Assim, testes que não são projetados com a inclusão em mente não podem discriminar adequadamente as pessoas que possuem determinada habilidade, mas que são afetadas por características do teste, daquelas que simplesmente não possuem tal habilidade. Nesse sentido, o *design* universal é uma forma de melhorar a validade dos testes para todos os participantes da avaliação (JOHNSTONE *et al.*, 2008), além de ser uma ferramenta eficiente para promover a justiça nas avaliações em larga escala.

2.4 PONTUAÇÃO DOS TESTES COM ITENS ABERTOS

A avaliação de competências complexas de modo credível é um tema que gera preocupações e a utilização de critérios de pontuação é um meio, cada vez mais comum, para resolver esse problema. Hoje as avaliações estão mais direcionadas para a avaliação da aprendizagem, no lugar dos testes tradicionais de conhecimentos, o que tem intensificado o interesse pelas avaliações com testes de itens de respostas construídas, pois se acredita que testes desse tipo são necessários para obter o pensamento de ordem superior dos alunos (JONSSON; SVINGBY, 2007; MESSICK, 1996). A avaliação com itens abertos pode, de certa forma, reproduzir atividades relacionadas ao mundo real do estudante, uma vez que a aprendizagem é um produto do contexto em que ela ocorre e, assim, a avaliação pode tentar refletir melhor a complexidade da realidade e fornecer dados mais válidos sobre a competência do estudante (DARLING-HAMMOND; SNYDER, 2000). Nessas avaliações, as respostas são elaboradas pelos alunos e, desse modo, não são possíveis as atribuições de pontuações completamente objetivas, então os critérios de pontuação são considerados uma abordagem eficaz para alcançar julgamentos precisos, consistentes e válidos sobre o desempenho dos estudantes (REDDY, 2011).

As avaliações, de modo geral, têm consequências para as pessoas ava-

liadas, seja no ambiente escolar ou em outras esferas, como nos exames de seleção. Nas avaliações classificatórias em larga escala, como nos vestibulares ou concursos para provimento de vagas, tais consequências são de suma importância, pois muitas vezes determinam aqueles que poderão seguir a carreira escolhida ou alcançar o emprego almejado. Assim, as avaliações devem ser consistentes, focadas principalmente na confiabilidade da medição, com julgamentos honestos e baseados em evidências. A pontuação atribuída ao respondente deve ser independente do avaliador, e os resultados semelhantes, mesmo que a tarefa tenha sido cumprida em outro lugar ou ocasião (JONSSON; SVINGBY, 2007; STEMLER, 2004; MOSKAL; LEYDENS, 2000; APPLEBEE, 2000; HUOT, 1990). Desse modo, 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). Na verdade, Stemler (2004) considera três abordagens principais para determinar a precisão e a consistência da pontuação: (1) estimativas de consenso, medindo o grau em que os avaliadores atribuem a mesma pontuação para o mesmo desempenho, (2) estimativas de consistência, medindo a correlação dos escores atribuídos pelos avaliadores e (3) estimativas de medição, medindo o grau em que os escores atribuídos são livres de erros.

Com o intuito de garantir melhores índices de confiabilidade, geralmente são utilizados critérios de pontuação na forma de rubricas. Os critérios de pontuação ou rubricas são esquemas descritivos de pontuação desenvolvidos com a finalidade de detalhar o modo como a pontuação deve ser atribuída, orientando as análises dos produtos ou processos elaborados pelos participantes da avaliação (MOSKAL; LEYDENS, 2000; REDDY, 2011). Desse modo, os critérios são utilizados com o objetivo de diminuir a subjetividade na atribuição de notas e guiar os avaliadores para que estes alcancem uma pontuação confiável no julgamento de alguma habilidade, em especial da escrita. Nas avaliações em larga escala, os critérios de pontuação devem ser definidos previamente, fazendo parte da avaliação.

As rubricas são cada vez mais utilizadas em avaliações com itens de respostas construídas, e os especialistas concordam que o seu uso adiciona qualidade à avaliação (JONSSON; SVINGBY, 2007; POPHAM, 1997). O desempenho de um indivíduo, na elaboração de uma determinada tarefa, não é julgado como certo ou errado, mas alocado em uma escala de habilidades, que pode ser contínua ou discreta. Os critérios são pensados como ferramentas utilizadas para medir o que os especialistas apontam como importante na

avaliação de um determinado desempenho e definem pontos de ancoragem ao longo da escala. Os critérios de pontuação são responsáveis também por estabelecer previamente as condições aplicadas para o evento, o que é desejado tanto para os avaliadores quanto para os participantes, independentemente se a avaliação ocorre em ambiente escolar ou em larga escala (JONSSON; SVINGBY, 2007; HAMP-LYONS, 2003; POPHAM, 1997)

Para alguns autores, a avaliação com itens de respostas construídas consiste em duas partes, uma tarefa e um conjunto de critérios de pontuação. Messick (1996) considera que o domínio do construto deve orientar tanto a seleção da tarefa quanto o desenvolvimento racional de critérios de pontuação.

Um grande número de especialistas defende a utilização de critérios de pontuação. Desde o início do século XX, quando efetivamente foram desenvolvidos os primeiros estudos sobre escalas de avaliação, os benefícios de seu uso são enumerados nas mais diversas pesquisas. Apesar disso, quando essas escalas são aplicadas nas avaliações da escrita, ainda não há um consenso entre os estudiosos, principalmente se a meta é alcançar a confiabilidade perfeita ou valorizar a variedade de pontuação proveniente de diferentes avaliadores. Também não há acordo sobre a representação da pontuação mais válida para a avaliação de um texto ou quais são as dimensões mais importantes da escrita em contextos específicos (HAMP-LYONS, 2011; SLOMP, 2005).

Outra crítica comum ao uso de critérios é que, apesar de numerosos estudos relatarem que a sua utilização melhora a eficácia da avaliação, há carência de investigações experimentais. A maioria dos estudos é limitada a artigos descritivos ou argumentativos, e, ainda, a maior parte desses estudos trata da eficácia da utilização de critérios de pontuação aplicados à avaliação da escrita nos exames de proficiência em inglês como primeira ou segunda língua em vez de avaliações escritas de outras disciplinas acadêmicas (REDDY, 2011; REZAEI; LOVORN, 2010).

Apesar de não haver estudos claramente contrários ao uso de critérios de avaliação, muitos autores fazem restrições para que sua aplicação resulte em uma marcação eficiente dos escores por parte dos avaliadores, uma vez que a simples utilização dos critérios não pode garantir uma avaliação eficaz. Critérios imprecisos ou mal formulados podem resultar em interpretações subjetivas ou ambíguas (KNOCH, 2011a; WEIGLE, 2002; KANE; COOKS; COHEN, 1999). Os critérios devem ser desenvolvidos localmente para propósitos específicos e, como qualquer ferramenta, o uso inadequado pode não ser vantajoso (REZAEI; LOVORN, 2010; KANE; COOKS; COHEN, 1999). Deve haver um rigoroso treinamento dos avaliadores para a confiabilidade das

pontuações (HAMP-LYONS, 2003; HUOT, 1996). Além disso, por não ser fácil a obtenção de alta confiabilidade em avaliações com itens de respostas construídas, autores destacam a necessidade de atenção para não sacrificar demasiadamente a validade em nome de melhores taxas de confiabilidade. Todos concordam que o conceito de confiabilidade deve ser considerado fundamental, mas o conceito de validade também deve ser explorado em relação a formas mais autênticas de avaliação, e as duas características devem ser consideradas na concepção de avaliações com itens de respostas construídas (JONSSON; SVINGBY, 2007; REZAEI; LOVORN, 2010; KANE; COOKS; COHEN, 1999; WIGGINS, 1994).

Alguns benefícios do uso de critérios de avaliação são comumente citados na bibliografia específica, adicionando muitas vantagens. Um deles reside na possibilidade de fornecer o julgamento válido de avaliação de competências complexas, sem prejudicar a necessidade de confiabilidade, pois o uso de critérios proporciona uma maior consistência de julgamento na avaliação com itens abertos (WEIGLE, 2002; BECKER, 2011; JONSSON; SVINGBY, 2007). Outra vantagem importante, e que não pode deixar de ser notada, é a promoção da aprendizagem, uma vez que possibilita a avaliação por pares e a autoavaliação. A explicitação dos critérios e padrões é fundamental para a qualidade dos comentários sobre o desempenho do aluno (*feedback*) (JONSSON; SVINGBY, 2007; HAMP-LYONS, 2003; SAXTON; BELAGER; BECKER, 2012; KANE; COOKS; COHEN, 1999; POPHAM, 1997). Além disso, a validade de conteúdo pode ser melhorada, alinhando-se instrução, critérios de avaliação, tarefas, currículo e avaliação (JONSSON; SVINGBY, 2007; MOSKAL; LEYDENS, 2000).

Para outras discussões sobre os critérios de pontuação, como as suas variações e utilidades, é necessário o esclarecimento do significado dos termos "descritores" e "níveis âncora", pois são muito utilizados nos ambientes educacionais e especificamente quando se fala em avaliações em larga escala.

Os descritores sintetizam as habilidades e competências que devem ser avaliadas em cada tópico do exame e devem ser expressos detalhadamente de modo que permitam a atribuição de pontuação precisa aos aspectos observados nas respostas do participante da avaliação. Já os níveis âncora são pontos selecionados na escala de habilidades relacionados com descritores que ilustram as variações dos critérios de pontuação para cada desempenho alcançado pelo estudante. Os níveis âncora são utilizados para garantir que as categorias de pontuação sejam bem definidas e evidenciar as fronteiras entre níveis sucessivos (ANDRADE; TAVARES; VALLE, 2000; MOSKAL; LEYDENS, 2000).

2.4.1 Tipos de critérios de avaliação

A avaliação da escrita tem sido considerada uma área problemática desde a década de 1920, quando os primeiros estudos buscavam avaliar aspectos limitados do desenvolvimento e da fluência na escrita. A partir da década de 1950, professores de composição e pesquisadores educacionais intensificaram as buscas de métodos capazes de produzir confiabilidade e validade para a avaliação da qualidade da escrita (HAMP-LYONS, 2003; HUOT, 1990).

O desenvolvimento e a utilização de critérios para a avaliação de aprendizagem de modo geral tornou-se uma tendência popular apenas a partir do início da década de 1990, quando passaram a ser aplicados no âmbito escolar como uma alternativa viável para examinar os trabalhos produzidos pelos alunos e por programas para comprovar a qualidade do ensino/aprendizagem. Atualmente essa utilização se dá em países como EUA, Reino Unido, Austrália, França, Turquia, entre outros (REDDY, 2011).

Três procedimentos principais são utilizados atualmente para atribuir pontuação diretamente às tarefas com itens abertos, são eles: pontuação característica principal, também denominada de traço primário, pontuação analítica e pontuação holística (BECKER, 2011). Esses procedimentos para a pontuação de tarefas com respostas construídas são descritos a seguir.

2.4.1.1 Pontuação característica principal

A pontuação característica principal, também conhecida como pontuação traço primário, desenvolvida por Lloyd-Jones e Klaus Carl (1977), é indicada para a avaliação da qualidade da escrita e não é comum exemplos de aplicações na avaliação com itens de respostas construídas de outras habilidades. Envolve a identificação de uma ou mais características relevantes para a tarefa de escrita determinada, e sua função é avaliar a função da linguagem primária ou traço de retórica provocada pela tarefa de escrita dada (HUOT, 1990). Conforme argumenta Applebee (2000), a pontuação característica principal foi desenvolvida para avaliar o desempenho em uma tarefa específica, e sua avaliação é feita em apenas um traço, como, por exemplo, "persuadir uma audiência". A pontuação poderia ser atribuída em uma escala do tipo *Likert* com um número pré-definido de pontos, por exemplo, 0, 1, 2 e 3, (0 = não consegue convencer o público, 3 = elabora um argumento convincente e bem desenvolvido).

Ao longo dos anos, as rubricas de característica principal foram sendo mudadas tanto na sua concepção original quanto ao uso pretendido. Atualmente elas são utilizadas em abordagens mais genéricas, no entanto a questão básica abordada na pontuação manteve-se. Os avaliadores devem manter o foco em respostas a perguntas como "*Será que o participante cumpriu com sucesso o objetivo desta tarefa?*". Assim, os avaliadores são instruídos a ignorar erros de convenções da linguagem escrita e se concentrar apenas na eficácia da questão que está sendo avaliada (APPLEBEE, 2000; HUOT, 1990).

A pontuação característica principal é baseada no fato de que em alguns contextos certos traços da tarefa são mais importantes do que outros e, assim, cada dimensão específica do texto que é relevante para a situação comunicativa proposta na avaliação é considerada, uma de cada vez. Se houver a necessidade de avaliar outras dimensões da escrita, o processo de pontuação deve ser repetido para cada uma delas.

2.4.1.2 Pontuação holística

A pontuação holística envolve a leitura de uma impressão individual da qualidade do processo ou produto como um todo, sem fazer o julgamento dos componentes separadamente. Quando aplicada à qualidade da escrita, baseia-se 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 (HAMP-LYONS, 2003; WHITE, 1984). Esse tipo de pontuação exige avaliadores especializados experientes e é necessário um considerável esforço para garantir a confiabilidade da pontuação, uma vez que esse tipo de pontuação considera a impressão causada por todo o produto e atribui uma pontuação global para o desempenho (HAMP-LYONS, 2003). A pontuação holística não se preocupa com componentes individuais do produto e geralmente destaca as características positivas e não o que falta ou é deficiente (WHITE, 1984; COHEN, 1994). Um exemplo de critérios de pontuação holística para a avaliação da qualidade da escrita está disponível no Apêndice B.2.

2.4.1.3 Pontuação analítica

A pontuação analítica consiste no julgamento de características individuais. O avaliador atribui uma pontuação para cada uma das dimensões que estão sendo avaliadas, por essa razão Hamp-Lyons (2003) denomina esse tipo de pontuação como pontuação de traço múltiplo. As rubricas analíticas usam a estratégia de pontuar cada critério separadamente e então agregar ou não as pontuações, dependendo da finalidade da avaliação, para formar uma pontuação geral. Portanto, essa pontuação é feita critério a critério, o que torna a avaliação multidimensional possível (REDDY, 2011).

A rubrica de pontuação analítica para avaliação da escrita considera os componentes separadamente, como, por exemplo, o conteúdo, o desenvolvimento, o vocabulário, a organização, a precisão, a coesão, a adequação das convenções da linguagem. A pontuação analítica permite uma maior valorização de alguns traços considerados mais importantes (BECKER, 2011; JONSSON; SVINGBY, 2007; KANE; COOKS; COHEN, 1999; HARSCH; MARTIN, 2012).

Esse tipo de pontuação exige procedimentos específicos ao contexto em todas as etapas do processo, desde o desenvolvimento, passando pela implementação e pontuação, até a elaboração de relatórios. A pontuação analítica, por sua característica multifacetada, permite identificar as qualidades ou características que são importantes na avaliação de tarefas escritas. Quando o objetivo é avaliar uma habilidade específica ou alguma característica relacionada a essa habilidade, a pontuação analítica é mais indicada do que a pontuação holística uma vez que esta última generaliza o desempenho da habilidade para uma pontuação global, o que pode representar uma perda de dados de diagnóstico. Consequentemente, a pontuação analítica pode ser considerada uma ferramenta superior para a medição de habilidades específicas (SAXTON; BELANGER; BECKER, 2012; HARSCH; MARTIN, 2012). Um exemplo de critério de pontuação analítica aplicada ao desempenho em tarefas de escrita pode ser conferido no Apêndice B.1

2.4.1.4 Comparação entre os tipos de pontuação

Todos os três tipos de pontuação apresentam vantagens e desvantagens dependendo da sua utilização. Quanto ao uso da pontuação característica principal, a principal vantagem é que a atenção por parte dos avaliadores é dada a um traço da tarefa de cada vez, o que torna a pontuação de cada

traço mais minuciosa (WHITE, 1984; COHEN, 1994; APPLEBEE, 2000). Outra vantagem é que esse tipo de pontuação considera o contexto no qual a avaliação está sendo aplicada, a sua finalidade e quais dimensões do produto são importantes para a situação proposta. Em contrapartida, quando aplicada à avaliação da escrita, esse tipo de pontuação tende a ser muito demorada, caracterizando uma desvantagem importante (JONES; CARL, 1977). Outra desvantagem é que a escala desenvolvida para a avaliação é muito específica expressando uma única característica de escrita de cada vez, necessitando o desenvolvimento de novas escalas para cada contexto, o que torna o processo muito oneroso (HAMP-LYONS, 2003; COHEN, 1994; HUOT, 1990).

Por essa metodologia de avaliar cada traço separadamente, a pontuação característica principal possui custo elevado, demanda avaliadores experientes e tempo para a aplicação (JONES; CARL, 1977; HAMP-LYONS, 2003). Esse motivo determina a utilização dessa ferramenta em contextos específicos para a avaliação da escrita, embora a sua utilização tenha se mostrado eficaz em avaliações em larga escala nos Estados Unidos (HUOT, 1990).

Uma das principais vantagens associadas com a utilização da pontuação holística é a praticidade. As escalas de pontuação holística são relativamente simples e abrangem poucos critérios para pontuar os resultados individuais dos respondentes (WEIGLE, 2002). Por outro lado, uma desvantagem de sua utilização é que ela não fornece informações suficientes para o diagnóstico do desempenho do estudante (COHEN, 1994). Outra desvantagem importante é a falta de precisão nos julgamentos. Estudos constataram que as avaliações cujos critérios são baseados em terminologia impressionista estão mais abertas a interpretações subjetivas ou ambíguas (KNOCK, 2009; WEIGLE, 2002). Além disso, estudos sugerem que a pontuação holística pode ser de utilização problemática para os avaliadores, que podem ter certa dificuldade na tentativa de equilibrar todos os aspectos dos critérios estabelecidos. Por exemplo, se um participante do teste cometeu erros graves em termos de gramática, os avaliadores podem ignorar outros aspectos importantes do desempenho. Problemas como esses pode levar a uma inconsistência com a interpretação de critérios de pontuação entre os avaliadores, comprometendo a confiabilidade das pontuações. Esse efeito pode ser minimizado com treinamento dos avaliadores para uma padronização dos critérios (ZAINAL, 2012). Além disso, a abordagem holística permite a sobreposição dos critérios estabelecidos, e o avaliador deve considerar e controlar tais sobreposições para evitar que alguns critérios sejam considerados além do previsto inicialmente, penalizando em demasia algum erro específico ou mesmo valorizando o desempenho além do previsto (MOSKAL, 2000).

As vantagens da utilização da pontuação analítica incluem um maior nível de detalhe em informações obtidas a partir dos escores dos respondentes, clareza quanto aos tópicos que estão sendo medidos, melhores condições para a interpretação dos dados, maior facilidade na interpretação da relação entre o que está sendo medido e as pontuações correspondentes (KANE; COOKS; COHEN, 1999), também permite que os avaliadores sejam treinados com facilidade em um conjunto claro de diretrizes (COHEN, 1994) resultando em uma melhoria da confiabilidade (HUOT, 1996; KNOCK, 2011). Além disso, as informações de diagnóstico sobre o desempenho ou *feedback* são bastante detalhadas (COHEN, 1994; KNOCK, 2011; POPHAM, 1997). Outra característica positiva é a possibilidade de generalização para outras tarefas (WEIGLE, 2002).

Entre as desvantagens da utilização da pontuação analítica está uma maior dificuldade na elaboração dos critérios, e seu desenvolvimento pode ser demorado e oneroso (HAMP-LYONS, 2003; WEIGLE, 2002; HARSCH; MARTIN, 2012). Escalas imprecisas, elaboradas em avaliações com itens de respostas construídas, podem resultar em marcações de acordo com a impressão geral obtida pelo avaliador, isto é, de acordo com a pontuação holística, aumentando o grau de subjetividade (KNOCH, 2011a; WEIGLE, 2002; WHITE, 1984).

Na pontuação holística, os avaliadores fazem julgamentos de modo geral sobre um desempenho. Esses julgamentos podem ser combinados com um ajuste entre as descrições sobre a escala, enquanto, na pontuação analítica, o avaliador atribui uma pontuação para cada uma das dimensões que estão sendo avaliadas. A pontuação holística requer uma análise do avaliador sobre a resposta escrita pelo aluno e, por esse motivo, tem um caráter mais subjetivo do que a pontuação analítica, na qual são determinados critérios para as pontuações que serão atribuídas às tarefas escritas. Geralmente a pontuação holística é utilizada com mais frequência nas avaliações em larga escala pela facilidade e rapidez nos julgamentos, além de ser relativamente precisa. Já a pontuação analítica é uma ferramenta usualmente utilizada na sala de aula, uma vez que os resultados auxiliam professores e alunos a identificar os pontos fortes e as lacunas no aprendizado (JONSSON; SVINGBY, 2007; POPHAM, 1997).

As vantagens e desvantagens dos três tipos de pontuação descritos são resumidas no Quadro 5.

Quadro 5 – Comparação entre os tipos de pontuação

PONTUAÇÃO	VANTAGENS	DESVANTAGENS
Característica principal	A avaliação de cada traço da tarefa é mais minuciosa. São avaliadas apenas as dimensões consideradas importantes para a situação específica.	As escalas não são integradas. Aplicação trabalhosa, demorada e dispendiosa.
Holística	A validade é aumentada pois o avaliador capta as propriedades importantes da tarefa. Escala única que integra as qualidades da resposta. Simplicidade, as escalas abrangem poucos critérios para pontuar.	Exige avaliadores experientes. Não considera as características negativas do desempenho. Não fornece informações do desempenho ou *feedback*. Pouca precisão nos julgamentos. Pode haver sobreposição de características avaliadas, penalizando os erros ou valorizando a competência em demasia.
Analítica	Informações mais detalhadas, clareza dos construtos medidos e facilidade na interpretação dos dados. Facilidade de treinamento para os avaliadores. Possibilidade de generalização para outras tarefas. A confiabilidade é melhorada. Fornece informações do desempenho ou *feedback*.	Maior dificuldade na elaboração, aplicação e pontuação. O desenvolvimento pode ser demorado e caro. Os avaliadores podem julgar por impressões holísticas.

Fonte: Autora

2.4.2 Comprimento da escala e o número de pontos

A prática mais comum nas avaliações com itens de respostas construídas, promovidas por instituições de nível superior nos Estados Unidos e em outras instituições ao redor do mundo, é utilizar uma escala com cinco ou seis pontos igualmente espaçados entre si (HAMP-LYONS, 2003; KNOCH; ELDER, 2010; KNOCH, 2011b). Na verdade, esse número de pontos de escala é justificado pelos resultados da pesquisa de Miller (1956), a qual estabelece que a capacidade das pessoas em processar informações limita-se a sete

(com variação de mais ou menos dois) elementos simultaneamente. Quando esse limite é excedido, a estrutura cognitiva pode ficar sobrecarregada, dificultando a compreensão.

No entanto, algumas vezes são utilizadas escalas de dez ou mais pontos, mas esse procedimento não é muito recomendado por especialistas, uma vez que não existem estudos evidenciando estatisticamente que os avaliadores podem distinguir com confiabilidade entre mais do que 10 níveis de qualidade (KNOCH, 2011a; HAMP-LYONS, 2003).

Mesmo assim, há certa preocupação para decidir o número de níveis, uma vez que são necessários números suficientes para discriminar entre diferentes graus de desempenho, mas o número de níveis não deve ser demasiadamente grande de modo que os avaliadores ainda possam fazer distinções entre eles em seus julgamentos (PENNY; JOHNSON; GORDON, 2000; POPHAM, 1997). Segundo North (2000), há uma relação direta entre confiabilidade e poder de decisão. Myford (2002), em um estudo com o objetivo de investigar a relação entre a confiabilidade e o número de pontos da escalas, concluiu que a confiabilidade é maior para as escalas com número de pontos entre cinco e nove. Além da preocupação com a confiabilidade e a capacidade de julgamento dos avaliadores, há a escolha do número de pontos em escalas para categorias específicas. Nem todas as categorias necessitam do mesmo número de pontos, algumas necessitam de uma escala mais refinada para julgamentos mais sutis, enquanto, para outras, fica difícil formular descritores em todos os níveis de modo que os avaliadores não tenham dificuldade de diferenciá-los em seus julgamentos (KNOCH, 2011a). O número de níveis apropriados na escala deve ser estabelecido de acordo com o contexto no qual a avaliação será empregada.

O grau de especificidade que pode ser facilmente conseguido em avaliações em matemática ou ciências, nem sempre é possível em áreas onde a diversidade de respostas é aceitável e até mesmo valorizada. Cronbach *et al.* (1995) propuseram a incorporação de números decimais entre os níveis da escala em avaliações nas quais é esperada uma variedade de respostas corretas, pois algumas vezes os avaliadores sentem que a resposta é um pouquinho superior, mas não suficiente para alcançar o próximo número inteiro da escala. A expectativa é que a confiabilidade de pontuação entre os avaliadores seja melhorada com a diminuição de erros acumulados com o julgamento apenas em números inteiros.

No Brasil, poucas informações são divulgadas em relação aos critérios de avaliação ou número de pontos de escalas que são normalmente utilizados nas avaliações em larga escala. As provas de redação do ENEM são pontua-

das de acordo com cinco competências que são estruturadas a partir da matriz de competências e habilidades definida pelos PCN (Parâmetros Curriculares Nacionais) – Ensino Médio (BRASIL, 2012, 2013).

Cada redação é corrigida por dois avaliadores independentes que atribuem uma nota entre 0 (zero) e 200 (duzentos) pontos para cada uma das cinco competências. A soma desses pontos compõe a nota total de cada avaliador, que pode chegar a 1000 pontos. A nota final do participante é a média aritmética das notas totais atribuídas pelos dois avaliadores. Se houver discrepância entre as duas notas atribuídas pelos avaliadores de mais de 200 (duzentos) pontos na pontuação total, ou se as notas atribuídas para alguma das competências diferirem em mais de 80 (oitenta) pontos, haverá uma nova correção por outro avaliador independente, então a nota final será a média aritmética das duas notas totais que mais se aproximarem. Se a discrepância persistir após a terceira correção, a redação será avaliada por uma banca composta por três professores, que atribuirá a nota final do participante. Cada competência é avaliada em cinco níveis de desempenho espaçados igualmente (BRASIL, 2012, 2013).

Com base nessas informações, verifica-se que a correção da redação do ENEM é feita segundo a pontuação analítica, na qual o julgamento é feito sobre características individuais, cada avaliador atribui uma pontuação para cada uma das dimensões que estão sendo avaliadas. A pontuação analítica permite a avaliação separada de cada competência avaliada e também uma escala descritiva diferente para cada uma das competências (MOSKAL, 2000), no caso do ENEM, a escala utilizada é a mesma em todas elas.

A FUVEST costuma divulgar todos os anos no manual do candidato o que ela denomina de "mecanismo de correção da redação", para informar ao candidato os critérios utilizados na correção da prova de redação. Esses critérios são divulgados, de modo geral, sem muitos detalhes. O mecanismo de correção é o seguinte: cópias do texto elaborado pelo participante são enviadas a dois avaliadores independentes, previamente treinados. As notas são atribuídas conforme três características: tipo de texto e abordagem do tema, estrutura e expressão. Cada uma dessas características recebe notas 0, 1, 2, 3 ou 4. Se houver alguma discrepância entre as notas provenientes dos avaliadores, a redação é encaminhada a uma "banca superior", que atribui a nota definitiva. A fuga ao tema proposto anula a redação que receberá nota zero (FUVEST, 2013).

As informações divulgadas pela FUVEST não contêm referências sobre o tipo de pontuação que é utilizado, mas, com base nessas informações, pode-se intuir que a nota é atribuída conforme a pontuação holística, na qual

cada avaliador faz julgamentos de modo geral sobre o desempenho em cada uma das características avaliadas.

Desde o vestibular de 2011, a prova de redação da UNICAMP consiste em um modelo que solicita ao candidato a elaboração obrigatória de vários textos de gêneros discursivos diversos. Nos concursos de 2011 e 2012 foi exigida a produção de três tarefas, nas edições de 2013 e 2014, de apenas duas. Cada uma das propostas é acompanhada por instruções específicas que objetivam delinear o propósito, o gênero e os interlocutores do texto a ser elaborado, além de textos para leitura que servem como inspiração, fornecendo as condições para a produção textual, situando o candidato em relação ao propósito de sua escrita. A correção dos textos escritos pelos candidatos considera as instruções que são fornecidas no enunciado.

O manual do candidato não fornece informações suficientes para determinar o tipo de pontuação que é utilizada na correção da prova de redação, mas provavelmente a nota seja atribuída conforme a pontuação holística, na qual cada avaliador faz julgamentos de modo geral sobre o desempenho em cada uma das características avaliadas. Outro motivo que justifica essa suposição é a informação contida nas provas comentadas de que os textos não são corrigidos com excesso de rigor quanto às normas da língua culta e que pequenos deslizes são ignorados, sendo essa uma característica da pontuação holística, que considera apenas os aspectos positivos do texto (COMVEST, 2012, 2013).

A UEL também promoveu mudanças na prova de redação a partir do vestibular de 2012, que passou a exigir 2 (dois), 3 (três) ou 4 (quatro) textos a serem produzidos conforme as instruções dadas, inclusive quanto à sua extensão. O candidato deve ler atentamente o enunciado e os textos que servirão de base para a sua resposta, pois a pontuação é atribuída conforme as instruções contidas no enunciado quanto às atividades de analisar, resumir, comentar, comparar, criticar, completar, entre outras. Aspectos discursivos, textuais, estruturais e normativos deverão ser levados em conta.

Inicialmente, as redações são corrigidas por 2 membros da equipe de modo independente, que atribuem notas entre 0 e 6 pontos. Se a diferença entre as notas for menor ou igual a 1 ponto, a nota final será a média aritmética entre as duas notas, caso seja identificada uma discrepância, os textos são lidos por um terceiro avaliador, sem que este saiba quais notas foram atribuídas anteriormente. Se a pontuação atribuída pelo terceiro avaliador for igual à média das pontuações 1 e 2, mantém-se a média, caso contrário será considerada pontuação final a média das duas pontuações que apresentarem menor diferença entre si (COPS/UEL(a), 2012; COPS/UEL(b), 2012).

A UEL também não divulga informações mais detalhadas sobre os critérios de pontuação utilizados ou sobre o número de pontos da escala. Com base nessas informações, a conclusão intuitiva é que a pontuação empregada também seja a holística.

2.5 ELABORAÇÃO DA AVALIAÇÃO

Em avaliações compostas por itens abertos, como as redações ou as provas dissertativas de quaisquer disciplinas, pela dificuldade de correção, o número de itens deve ser resumido, principalmente se o número de pessoas submetidas a essas avaliações for demasiadamente grande, como é o caso dos concursos vestibulares para acesso às universidades, do ENEM e de alguns concursos públicos para provimento de vagas de trabalho. É difícil evitar a generalização de desempenho do candidato em uma pequena amostra de tarefas, deduzindo a forma como o candidato se sairia ao escrever em resposta a outras tarefas semelhantes. A questão não é a generalização, mas como poderiam ser minimizados os erros ou qualificadas as inferências feitas (SUDWEEKS; REEVE; BRADSHAW, 2005).

A pontuação que um indivíduo recebe ao participar de um teste escrito, muitas vezes, é influenciada por fatores externos ao teste, como, por exemplo, o avaliador particular que julgou a tarefa, fatores diversos associados à ocasião em que o teste foi aplicado, interesse e conhecimento do participante sobre o assunto apresentado e, ainda, interações entre algumas dessas fontes. A variabilidade das classificações devido a qualquer fonte externa é considerada erros de medição (WEIGLE, 2002; LI, 2003; MOSKAL; LEYDENS, 2000; SUDWEEKS; REEVE; BRADSHAW, 2005). Espera-se que a variabilidade na classificação de um grupo submetido a uma avaliação com itens abertos seja devido a diferenças confiáveis entre as habilidades avaliadas dos indivíduos.

Segundo Applebee (2000), as medidas mais confiáveis de habilidades em avaliações em larga escala são baseadas em itens de múltipla escolha, mas as habilidades assim avaliadas são limitadas. Medidas com maior validade, com itens de respostas construídas, requerem um instrumento de avaliação bem elaborado e um amplo treinamento dos avaliadores com o intuito de alcançar um padrão comum.

Muitos elementos influenciam a validade e a confiabilidade das medidas em uma avaliação com itens de respostas construídas, mas existem algumas variáveis que afetam diretamente o desempenho do respondente, entre

elas está o formato da questão e o comando, este é responsável pelo estímulo, a inspiração para o participante escrever a sua resposta. Na língua inglesa, em avaliações da habilidade da escrita, esse comando é conhecido como *prompt* (KOBRIN *et al.*, 2011; HAMP-LYONS; MATHIAS, 1994). Algumas destas avaliações em larga escala utilizam apenas um comando, cada respondente escreve apenas um ensaio, outros utilizam comandos múltiplos com variações do formato. O fato é que, até os dias atuais, ainda não há um consenso entre os pesquisadores sobre o que é essencial para garantir a consistência do teste, julgamentos precisos da qualidade da escrita e a determinação do grau de dificuldade das tarefas (HAMP-LYONS, 2011).

Aliás, na elaboração e aplicação de uma avaliação com itens abertos, principalmente as em larga escala, as maiores dificuldades residem no julgamento preciso da habilidade que está sendo avaliada e na determinação do grau de dificuldade dos itens (HAMP-LYONS, 2011), isso porque são muitas as variáveis envolvidas nas etapas demandadas para o estabelecimento do escore e no tipo de tarefa que será determinada ao respondente (HAMP-LYONS, 2011; HUANG, 2008; BRIDGEMAN; MORGAN; WANG, 1997; JENNINGS *et al.* 1999).

Historicamente, as abordagens para orientar a elaboração de avaliações com itens abertos têm sido influenciadas pelas pesquisas sobre os conceitos de validade, tendo como centro os estudos sobre medição. São muitos os tipos e definições de validade, mas a atenção quanto à elaboração da tarefa deve ser voltada para a validade de conteúdo e a validade de construto. A validade de conteúdo é alcançada com o alinhamento do conteúdo exigido para a elaboração da tarefa e a habilidade que se pretende medir. Para a validade de construto, é necessária uma preocupação com o projeto de teste como um todo e com a relação entre o projeto do teste e a habilidade que se pretende medir (DEANE, 2013). Uma preocupação importante com a validade está relacionada com a necessidade de definir claramente o construto que o teste deve medir, além de ter os critérios de avaliação definidos de forma clara e acessível a todos os participantes para que não haja uma desconexão entre a tarefa percebida pelos respondentes do teste e o que é avaliado na correção.

Segundo Deane (2013), o que é importante na elaboração de uma avaliação com itens de respostas construídas, principalmente no estabelecimento da tarefa, é a delimitação do construto que deve ser avaliado. Para projetar a avaliação e determinar o impacto desta no ensino e na sociedade, devem ser tomadas decisões sobre o foco da avaliação, se a medida do desempenho avaliado será sobre habilidades específicas ou sobre aspectos integrados do produto ou processo. Em avaliações da habilidade da escrita, por exemplo,

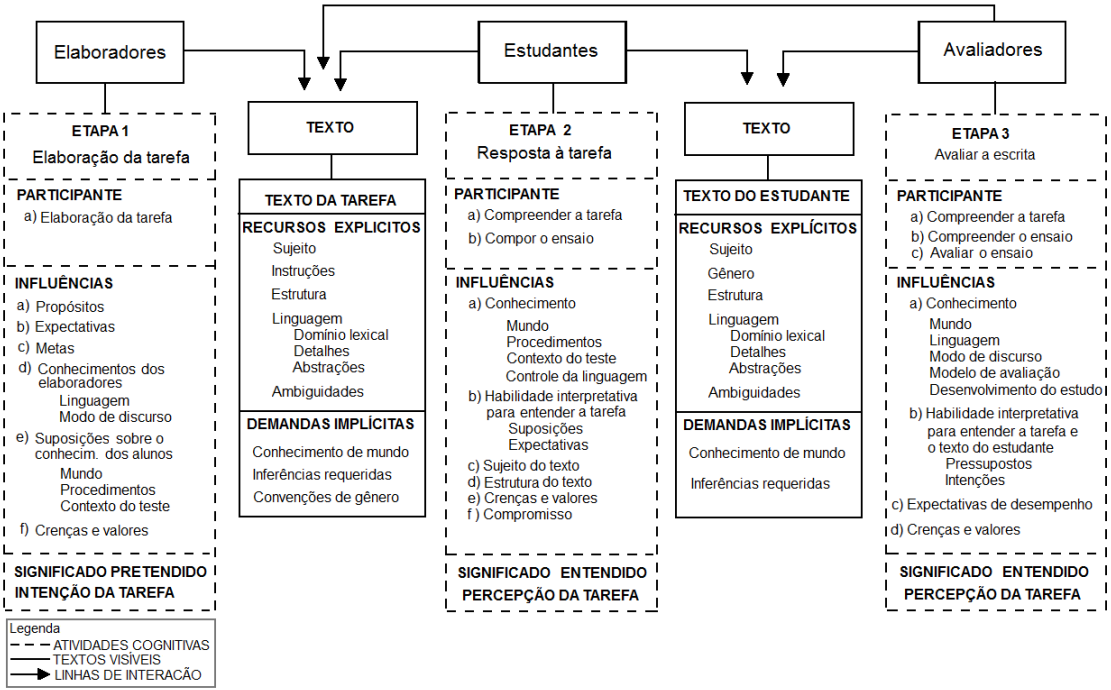
se a decisão for por uma avaliação integrada, quais aspectos são realmente importantes para serem incluídos, tais como o conhecimento das normas da linguagem, o gênero que pode variar entre carta argumentativa, artigo de opinião, resenha crítica, entre outros, e quais as finalidades retóricas que serão consideradas na avaliação. Estas consistem nas variantes argumentativas para a comunicação, a organização do texto que é caracterizada pelas relações que são estabelecidas entre suas partes de modo a garantir a coesão e a coerência do discurso. Além desses aspectos, devem ser consideradas as características dos contextos do mundo real, quais delas devem ser incorporadas na avaliação e quais aspectos explícitos ou implícitos do texto escrito pelo participante serão considerados na determinação da pontuação.

Uma tendência atual, destacada por pesquisadores como Messick (1989) e Moss (1994), entre outros, é a preocupação em considerar a "validade consequente" que se refere às consequências sociais resultantes da utilização de uma avaliação em larga escala específica para um determinado propósito. Segundo eles, a avaliação deve ser considerada em um contexto mais humanista, no qual são destacados conceitos que envolvem fatores como o domínio linguístico, o contexto sociocultural, a variação da tarefa, as diferenças retóricas e as variadas condições e formatos do teste. A percepção da qualidade do desempenho pelos avaliadores pode ser influenciada por esses fatores (DEANE, 2013; BECK; JEFFERY, 2007).

Ruth e Murphy (1984) desenvolveram um modelo idealizado para ser aplicado em avaliações da escrita, contendo todas as características envolvidas, conceituando a avaliação como uma unidade ao invés de descrever e investigar os elementos de maneira isolada. Desse modo, são investigadas as relações dinâmicas e interativas entre os participantes e os processos e textos que compõem o evento inteiro da avaliação da escrita. Apesar de esse modelo ser direcionado especificamente para avaliações da habilidade da escrita, pode ser facilmente generalizado para avaliações com itens abertos de modo geral, uma vez que são mantidas as relações entre os atores participantes da avaliação.

O modelo de Ruth e Murphy (1984) é ilustrado na Figura 1 e descrito na sequência.

Figura 1 – Processos e participantes em uma avaliação escrita



Fonte: Adaptado de Ruth e Murphy (1984)

Os principais atores do evento da avaliação são: os elaboradores, os participantes (alunos ou candidatos) e os avaliadores do teste. O modelo é composto por três etapas distintas. Cada um dos atores executa a tarefa que lhe é destinada em momentos específicos do evento de avaliação, são elas: 1) a criação de uma tarefa para a escrita, 2) a leitura do tópico pelo aluno participante e a sua resposta escrita e 3) a avaliação da resposta do participante.

O modelo enfatiza a leitura e a interpretação, pelos atores, dos dois textos que são elaborados no evento: 1) o texto criado com as instruções da tarefa e 2) o texto escrito pelo aluno. Cada texto tem propriedades específicas e cada leitor interage com eles de acordo com o episódio da avaliação do qual faz parte. As metas comunicativas na elaboração desses textos referem-se ao modo como cada um deles é entendido pelos participantes em cada uma das etapas.

A primeira etapa do evento da avaliação da escrita tem início com uma reunião dos elaboradores do teste composto por temas, perguntas, ou outras formas de estímulos. Na Figura 1, essa etapa é ilustrada por meio da primeira caixa desenhada com linhas tracejadas.

As decisões para a elaboração da tarefa para a avaliação da escrita são diretamente influenciadas pelo conhecimento dos elaboradores quanto à compreensão da finalidade da avaliação, à linguagem e ao ato de escrever, às teorias implícitas de retórica, à estrutura do discurso, ao desenvolvimento dos alunos, às suposições sobre o conhecimento de mundo dos alunos, às suas próprias crenças e valores sobre os contextos nos quais as tarefas são baseadas.

Desse modo, bons resultados nessa primeira fase do evento de avaliação dependem de decisões de pessoas altamente especializadas.

A segunda fase do evento da avaliação escrita ocorre quando o aluno participante tem em suas mãos o texto com a tarefa da avaliação. O ato de escrever, na verdade, tem início com o ato de ler o tópico de avaliação, durante o qual o participante deve compreender a tarefa pretendida pelo elaborador. Os descritores da tarefa da escrita, em forma de texto, trazem algumas características explícitas, como as informações e instruções, mas escondem também outras informações, intencionais ou não, que exigem do participante o emprego de habilidades de inferência, de memória e de resolução de problemas com a leitura e a interpretação. Além das sugestões dadas, há também as suposições dos elaboradores sobre o conhecimento que o participante tem do mundo, da linguagem e o conhecimento dos procedimentos relevantes.

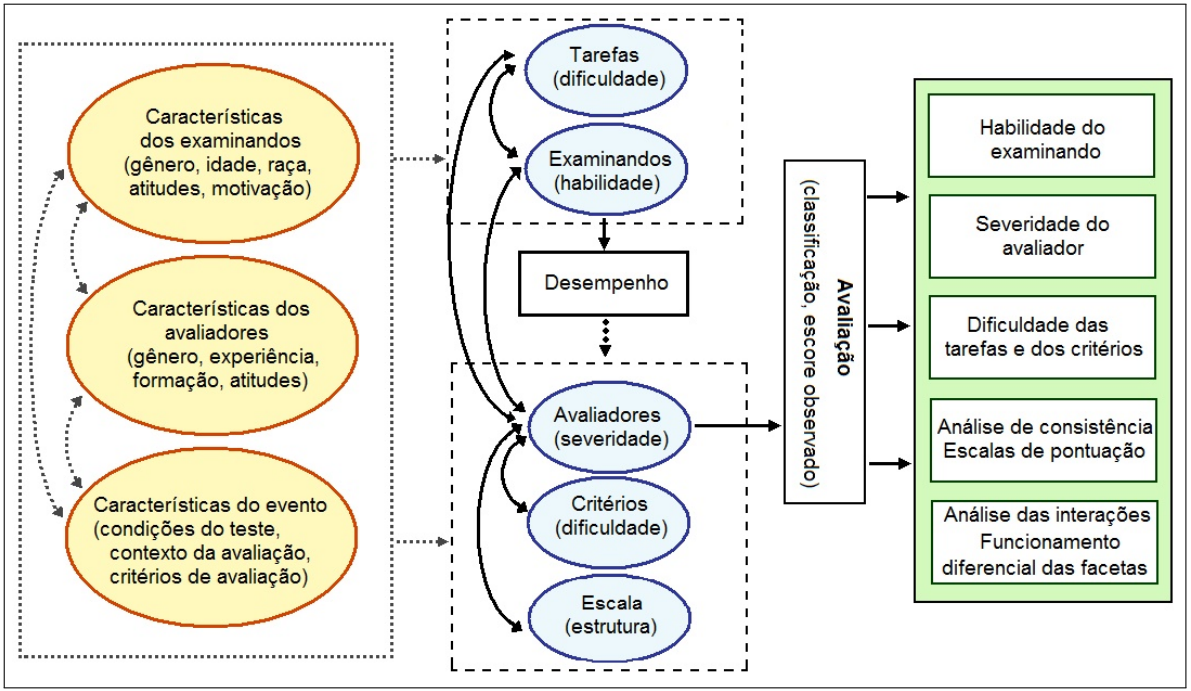
Nessa etapa do processo, o participante lê a tarefa de escrita e as informações disponíveis para compor a sua resposta escrita, que também apresenta

características explícitas e implícitas, bem como pode apresentar efeitos indesejados que serão observados nas leituras subsequentes feitas pelos juízes avaliadores. A tarefa fundamental do participante é produzir um texto que irá se tornar o objeto de análise e julgamento na próxima fase do evento de avaliação.

A terceira fase do evento de avaliação ocorre quando os avaliadores recebem a tarefa elaborada para a pontuação. Na verdade, o avaliador recebe dois textos que exigem leitura e interpretação: o texto com a tarefa da avaliação e a resposta escrita pelo participante. Os atos de ler esses dois textos exigem todos os processos construtivos de interpretação e compreensão envoltos no processo de leitura. A adoção de critérios de pontuação pelos avaliadores do teste é essencial para a realização objetiva da leitura e avaliação dos ensaios. No entanto, uma série de outros componentes colabora com o leitor para moldar o entendimento de ambos os textos: as expectativas, os modelos retóricos preferenciais, o conhecimento do mundo, os preconceitos, as experiências de correção, entre outros.

O modelo conceitual elaborado por Eckes (2009, 2011), exibido na Figura 2, destaca os fatores que potencialmente influenciam a classificação dos examinandos em avaliações que necessitam do julgamento de avaliadores, especificamente as com itens de respostas construídas e alguns de seus relacionamentos mútuos. Além de evidenciar as variáveis que comumente são geradoras de erros importantes nessas avaliações, preocupa-se também com as análises que devem ser feitas para monitorar o modo como os vários avaliadores atribuem as pontuações, a qualidade da pontuação atribuída pelos avaliadores, a veracidade da classificação dos participantes quanto aos seus níveis de desempenhos, o nível de dificuldade das tarefas propostas e algumas interações entre essas variáveis. Nesses trabalhos, o autor direciona a utilização desse modelo conceitual para a aplicação do modelo multifacetas de Rasch, sendo esta a principal ferramenta utilizada para as análises propostas. Esse é o motivo que torna tais referências importantes para o desenvolvimento desta tese, uma vez que o modelo multifacetas de Rasch é essencial para o desenvolvimento deste trabalho e para as análises da aplicação prática.

Figura 2 – Quadro conceitual de fatores relevantes nas avaliações com itens abertos



Fonte: Adaptado de Eckes (2009, 2011)

A parte central do diagrama destaca as variáveis que oferecem impacto direto sobre a pontuação atribuída aos examinandos. Entre eles, o mais importante obviamente é a habilidade (proficiência) do examinando quanto ao construto que está sendo medido.

Os outros fatores exibidos na parte central são basicamente irrelevantes para o construto que está sendo medido, entretanto contribuem potencialmente para erros sistemáticos de medição nas avaliações: (a) efeitos causados pelos avaliadores: severidade, tendência central, halo; (b) variabilidade na dificuldade das tarefas apresentadas no exame; (c) variabilidade na dificuldade dos critérios de classificação. E, ainda, uma fonte de erros de medição menos óbvia refere-se à variação na estrutura da escala de classificação, uma vez que os avaliadores, ao longo das sessões de pontuação, podem mudar o significado das categorias ordenadas da escala de classificação utilizada, não diferenciando realmente níveis de desempenho adjacentes.

No lado esquerdo do quadro (Figura 2), são mostradas três categorias de variáveis que também influenciam a classificação dos examinandos, embora geralmente de uma forma menos direta: (a) características dos examinandos: gênero, etnia, nacionalidade, traços de personalidade, crenças, objetivos, entre outros; (b) características dos avaliadores: gênero, experiência profissional, objetivos, motivação, atitudes, crença, entre outros; (c) características do evento de avaliação e contexto: ambiente técnico e físico, carga de trabalho do avaliador, tempo de duração das sessões de pontuação, política de gestão da qualidade, valores organizacionais.

Alguns desses fatores podem interagir uns com os outros e também podem interagir com alguns dos fatores da parte central. Por exemplo, o avaliador pode ser mais severo quando julga o desempenho de homens, ou, então, quando a experiência profissional do avaliador influencia a interpretação que faz da tarefa escrita pelo examinando.

No lado direito do diagrama, são destacadas algumas entre as principais análises que podem ser feitas nas avaliações com itens abertos, especialmente da habilidade da escrita. Nessas avaliações, podem ser adicionadas variáveis nas configurações do modelo conceitual, dependendo do interesse particular do estudo.

Nas avaliações com itens abertos, são muitos os fatores que podem afetar a medida do desempenho das pessoas ao executar a tarefa determinada no teste. A elaboração desses testes consiste em um conjunto diverso e complexo de procedimentos que visam à medida da proficiência sobre o construto que se deseja medir. Esses testes podem variar em uma gama de diferentes formatos e sofrem interferências de variáveis que podem fazer parte ou não

da situação de avaliação.

No contexto do modelo multifacetas de Rasch, essas variáveis são escolhidas para fazer parte do modelo de avaliação e são denominadas de *facetas*. Por exemplo, um teste para avaliar a habilidade de leitura e compreensão de textos do aluno pode ser composto por um texto no qual ele deve basear-se para elaborar respostas curtas a algumas perguntas. As respostas podem ser pontuadas como corretas ou incorretas de acordo com critérios bem definidos previamente, assim, a nota que a pessoa receberá nesse teste dependerá da habilidade de compreensão de texto (construto que se deseja medir) e da dificuldade de cada item. Nesse caso, o modelo de estudo consistirá em duas facetas: (1) a habilidade do examinando e (2) a dificuldade do item. Cada elemento da primeira faceta (examinandos) interage com cada elemento da segunda faceta (itens) para produzir a resposta observada.

Outro teste para avaliar a habilidade da expressão escrita pode ser constituído por alguns itens que contenham um texto ou imagem situando o examinando no assunto sobre o qual ele deve elaborar um texto. A pontuação que a pessoa receberá não depende apenas da habilidade que se deseja avaliar no teste (habilidade de expressão escrita) e da dificuldade da tarefa, depende também de características do avaliador ou dos avaliadores escolhidos para julgar o texto escrito, por exemplo, a sua severidade. Nesse caso, o modelo de avaliação pode ser composto por 3 facetas: (1) a habilidade do examinando, (2) a dificuldade do item e (3) a severidade do avaliador. As variáveis que definem as medidas nesse contexto são: cada elemento da primeira faceta (examinandos) interage com cada elemento da segunda faceta (itens), que, por sua vez, interage com cada elemento da terceira faceta (avaliadores).

Nessa mesma situação de teste, poderia ser de interesse para o estudo analisar a influência resultante de critérios de pontuação distintos. Para isso, seria necessária a inclusão de critérios de pontuação, por exemplo, pontuação analítica e holística. Nesse caso, seria necessário incluir uma quarta faceta no modelo, o critério de pontuação. Assim, a situação de medição seria: cada elemento da primeira faceta (examinandos) interagindo com cada elemento da segunda faceta (itens), interagindo também com cada elemento da terceira faceta (avaliadores) e, ainda, com cada elemento da quarta faceta (critérios) para produzir a medida. Desse modo, as facetas podem ser incluídas conforme a necessidade do estudo, principalmente se for verificado que elas exercem algum impacto sobre a habilidade do indivíduo que está sendo avaliada.

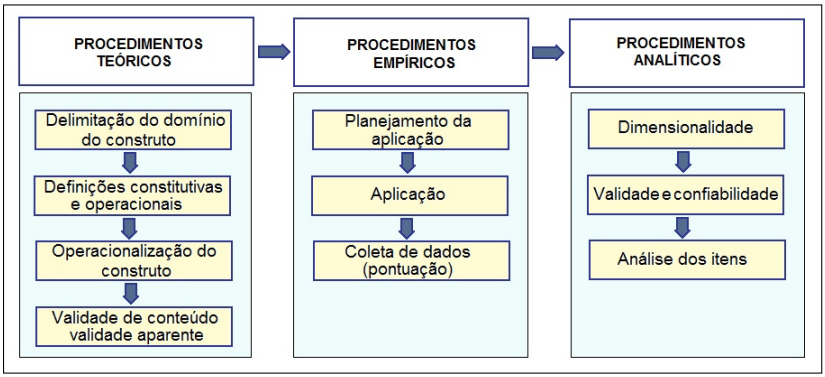
Os eventos de avaliação são caracterizados por conjuntos de fatores distintos que direta ou indiretamente interferem nos resultados observados

(escores). Uma *faceta* pode ser definida como qualquer fator, variável ou componente da situação de avaliação que afeta os resultados da avaliação de modo sistemático (ECKES, 2011; LINACRE, 2002a). Desse modo, as facetas incluem as variáveis de interesse direto, a habilidade que se deseja medir, e também aquelas que indiretamente contribuem sistematicamente para a ocorrência de erros nas medições, tais como os provocados pelas características dos avaliadores, as variadas formas de tarefas, o tempo disponível para as respostas, entre outros.

Outro modelo para a elaboração de instrumento de medidas aplicáveis à construção de testes psicológicos ou avaliações em geral, proposto por Pasquali (2010), contém informações mais específicas para a elaboração das tarefas, preocupando-se principalmente com os construtos avaliados e com a validade do instrumento. O modelo baseia-se em três polos ou procedimentos: (1) procedimentos teóricos, (2) procedimentos empíricos (experimentais) e (3) procedimentos analíticos (estatísticos).

Uma ilustração desse modelo é exibida na Figura 3.

Figura 3 – Modelo para elaboração de instrumento de medida



Fonte: Adaptado de Pasquali (2010).

Fazendo uma analogia entre esses três modelos conceituais quando aplicados a avaliações com itens de respostas construídas, as três etapas do modelo de Ruth e Murphy (Figura 1) estão organizadas nas duas primeiras etapas do modelo de Pasquali (Figura 3), que considera também as análises estatísticas dos resultados da avaliação em sua terceira etapa. Além disso, o modelo de Pasquali possui informações concretas sobre os procedimentos de elaboração dos itens da avaliação, com a delimitação do domínio do cons-

truto, definições constitutivas e operacionais e operacionalização do construto, além dos procedimentos para análise da validade de conteúdo e validade aparente. Já o modelo de Eckes (Figura 2) trata de modo mais conciso todas as variáveis presentes no modelo de Ruth e Murphy (Figura 1), além de se preocupar também com as análises para monitoramento dos resultados da avaliação.

2.5.1 Procedimentos Teóricos

Na elaboração de uma avaliação, principalmente no estabelecimento da tarefa, o mais importante é a delimitação do domínio do construto que deve ser avaliado (DEANE, 2013; PASQUALI, 2010; MISLEVY; HAERTEL, 2006).

Pasquali (2010) sugere procedimentos práticos para a elaboração de instrumentos de medidas. O desenvolvimento das tarefas está inserido no polo "Procedimentos Teóricos" e consiste na fundamentação da teoria envolvida na pesquisa, isto é, a explicitação da teoria sobre o construto para o qual se quer desenvolver o instrumento de medida. Esses procedimentos são compostos por quatro etapas: (1) delimitação do domínio do construto, (2) definições constitutivas e operacionais, (3) operacionalização do construto e (4) análise da validade de conteúdo e validade aparente.

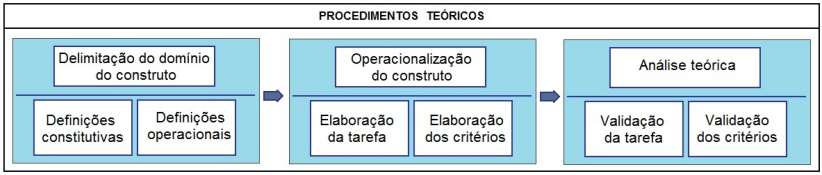
O modelo proposto por Pasquali refere-se à construção de instrumentos de medidas para serem aplicados à construção de testes psicológicos ou testes de desempenho em geral. Para as avaliações com itens abertos, além das etapas propostas por ele, são necessários procedimentos relacionados com a elaboração dos critérios de pontuação e a sua validação.

Os pesquisadores da área de avaliação concordam que a avaliação com itens de respostas construídas consiste em duas partes, uma tarefa e um conjunto de critérios de pontuação. Nas avaliações em larga escala, os critérios de pontuação devem ser definidos previamente, fazendo parte da avaliação (MOSKAL; LEYDENS, 2000; JONSSON; SVINGBY, 2007; HAMP-LYONS, 2003). Messick (1996) considera que o domínio do construto deve orientar tanto a seleção da tarefa quanto o desenvolvimento racional de critérios de pontuação. Mislevy e Haertel (2006) compartilham da mesma opinião ao defenderem a abordagem de avaliação centrada no construto. Segundo eles, o projeto de uma avaliação deve começar com a definição do construto e, então, ir seguindo com a construção do modelo, que consiste basicamente na tarefa e nos critérios de pontuação que explicitam as evidên-

cias a serem coletadas.

Desse modo, para as avaliações com itens abertos, são acrescentados ao modelo de Pasquali os procedimentos responsáveis pela criação e validação dos critérios de pontuação. Na Figura 4, é apresentada uma ilustração do polo *Procedimentos teóricos* do modelo, agora acrescido desses procedimentos.

Figura 4 – Etapas para a elaboração da tarefa



Fonte: Adaptado de Pasquali (2010)

2.5.1.1 Delimitação do domínio do construto

Para a delimitação do domínio do construto, Pasquali (2010) sugere que, primeiramente, deve-se estabelecer a dimensionalidade do atributo que se deseja avaliar, isto é, determinar se o atributo é constituído por vários componentes distintos (multidimensional) ou se é caracterizado como uma única unidade (unidimensional). A dimensionalidade pode ser estabelecida por meio da teoria existente sobre o construto e também por meio dos resultados de pesquisas que utilizaram análise fatorial dos dados, se esses estiverem disponíveis.

Quando se tem a intenção de utilizar algum entre os modelos da TRI na avaliação, como os modelos da família de Rach, existe a imposição de que os dados sejam resultantes da medida de uma única variável unidimensional subjacente, uma vez que a unidimensionalidade é um requisito necessário para a medição por meio dos modelos da TRI, em especial, pelos modelos de Rasch.

Entretanto, na prática, a unidimensionalidade é um conceito mais qualitativo do que quantitativo. Segundo Wright e Linacre (1987), dificilmente um teste real será perfeitamente unidimensional. As situações empíricas não podem atender exatamente os requisitos para a unidimensionalidade. Seja nas ciências em geral, seja nas situações de testes, as correções para a obtenção de unidimensionalidade são inevitáveis e fazem parte das técnicas ex-

perimentais. No entanto, o ideal é que as medidas sejam aproximadamente unidimensionais, e resultados generalizáveis sejam obtidos para o teste.

Esses autores ainda sugerem que a busca pela unidimensionalidade aproximada seja realizada em dois níveis. Em primeiro lugar, os elaboradores do teste devem despender todos os esforços para produzir um conjunto útil de tarefas, juntamente com as categorias de classificação (escalas) para obter informações inequívocas ao longo de uma única dimensão. Todos os aspectos da situação de teste (tarefas, itens, técnicas de observação, etc.) devem ser organizados, o mais perfeitamente possível para que o examinando demonstre a sua habilidade em relação ao atributo que o teste é destinado a medir. Em segundo lugar, devem ser feitas análises sobre uma amostra relevante de observações, cuidadosamente definida, para a avaliação das intenções de unidimensionalidade.

Para Messick (1996), as maiores ameaças à validade do instrumento se referem à representação do construto, na qual se tenta identificar os mecanismos teóricos subjacentes à habilidade que se pretende medir quando o indivíduo executa a tarefa proposta, principalmente pela decomposição da tarefa em processos ou componentes. A representação do construto é fortemente baseada na psicologia cognitiva de processamento de informações e refere-se à dependência relativa aos processos, estratégias e conhecimentos que estão implícitos na execução da tarefa. Uma dessas ameaças é denominada *variância construto-irrelevante*, na qual o instrumento de avaliação é muito amplo, capturando processos irrelevantes para a interpretação do construto. A outra é denominada *sub-representação do construto* que, ao contrário da anterior, ocorre quando o instrumento captura uma porção muito estreita, não incluindo dimensões ou facetas importantes do construto.

A definição do construto também deve ser feita baseando-se na literatura e nos conhecimentos existentes sobre o assunto e deve descrever claramente todos os fatores envolvidos. A demanda por habilidades verbais necessárias na medição de outra habilidade é uma importante fonte de variância construto-irrelevante. Por habilidades verbais, entende-se ler, escrever, falar e ouvir. Desse modo, a medida de habilidades por meio de itens abertos normalmente requer a leitura de um texto seguido da necessidade de interpretação correta, um domínio razoável de conhecimentos e pensamento crítico. Por isso, essa medida pode ser problemática (HALADYNA; DOWNING, 2004).

Pasquali (2010) sugere que a delimitação do construto deve considerar dois aspectos: as definições constitutivas e as definições operacionais do construto.

Definições constitutivas

Entende-se por definições constitutivas definir um construto por meio de outro construto; o construto é concebido em termos de conceitos próprios da teoria em que ele se insere. Essas definições são conceitos abstratos, por exemplo, se a inteligência verbal for definida como a "capacidade de compreender a linguagem", esta é uma definição constitutiva, pois a capacidade de compreender constitui uma realidade abstrata, um construto, um conceito. As definições constitutivas situam o construto dentro da sua própria teoria, fornecendo os limites sobre o que deve ser explorado e considerado ao se medir o construto. Esses limites não se resumem apenas às fronteiras que não podem ser ultrapassadas, mas principalmente estabelecem as fronteiras que devem ser atingidas.

As boas definições constitutivas são de grande importância na elaboração de instrumentos de medidas e ajudam a evitar as ameaças à validade apontadas por Messick (1996) citadas anteriormente. Pasquali (2010, p. 173) afirma que

> [···]as boas definições constitutivas vão permitir em seguida avaliar a qualidade do instrumento, que mede o construto em termos do quanto de sua extensão semântica é coberta pelo instrumento, surgindo daí instrumentos melhores e piores à medida que medem mais ou menos da extensão conceitual do construto, extensão essa delimitada pela definição constitutiva desse mesmo construto.

Definições operacionais

As definições operacionais do construto viabilizam a passagem da teoria para a prática, uma vez que um instrumento de medida propicia uma operação concreta. Segundo Pasquali (2010), caracteriza-se como um dos momentos mais críticos na elaboração de instrumentos de medidas, pois nessas definições é que se baseia a legitimidade da representação empírica e comportamental dos construtos, assegurando a validade do instrumento.

Em primeiro lugar, para a definição ser operacional, o construto deve ser definido em termos de operações concretas e não em termos de outros construtos como para as definições constitutivas. O construto deve ser definido através de comportamentos físicos que o expressam.

A definição é operacional se você puder dizer à pessoa: "vá e faça ···" (MAGER, 1981, *apud* PASQUALI, 2010). Assim, para a definição constitutiva exemplificada anteriormente, "capacidade de compreender a linguagem",

uma definição operacional bem definida poderia ser "Escreva sobre ⋯" indicando claramente o que a pessoa deve fazer. As definições operacionais devem definir comportamentos concretos específicos que devem ocorrer.

Em segundo lugar, as definições operacionais devem ser o mais abrangente possível em relação ao construto. Para garantir uma boa cobertura do construto, as definições operacionais deverão listar e especificar as categorias de comportamentos que representam o construto. Quanto melhores e mais completas forem essas especificações, melhor será a garantia de se ter um instrumento válido para a medida daquele construto.

2.5.1.2 Operacionalização do construto

Nessa etapa são elaborados tanto as tarefas para a avaliação como os critérios para a pontuação das respostas, com o objetivo de avaliar a habilidade das pessoas em relação ao construto que foi proposto. Segundo Pasquali (2010), as tarefas devem ser elaboradas de modo que expressem a representação comportamental do construto, que, por sua vez, já foi delimitado nas definições constitutivas e operacionais do construto. Também para Messick (1996), o domínio do construto deve orientar a seleção da tarefa e o desenvolvimento dos critérios de pontuação. A pontuação atribuída ao ensaio do participante deve refletir o construto que está sendo medido. Os critérios devem corresponder às categorias das habilidades da escrita que estão sendo avaliadas para estabelecer o grau de domínio que o indivíduo possui sobre cada aspecto do construto avaliado (BECKER, 2011).

Segundo Messick (1996), existem duas maneiras para se operacionalizar o construto. Pode-se começar por esclarecer a natureza dos construtos a serem avaliados e, em seguida, selecionar ou construir tarefas que melhor os representem. Essa é denominada *abordagem centrada no construto*; ou, então, pode-se começar com uma tarefa que exige um ótimo desempenho do construto a ser avaliado e questionar sobre quais competências ou construtos essa tarefa demanda, é a *abordagem centrada em tarefas*.

Aliás, a abordagem de avaliação centrada no construto está se tornando cada vez mais proeminente. Mislevy e Haertel (2006) enfatizam a importância de começar o projeto de uma avaliação com a definição do construto e então ir construindo um modelo que explicita as evidências a serem coletadas. No entanto, para Messick (1996), o que é fundamental em uma avaliação não é a operação demandada para o desempenho da tarefa, mas o que é capturado nos resultados do teste e suas interpretações, isto é, a validade

do construto.

Segundo Kane (2013), o construto é implicitamente definido pela sua teoria. Se as previsões derivadas da teoria não concordam com os resultados provenientes da avaliação, ou a teoria está errada ou o instrumento de medidas não é apropriado. Se as previsões forem confirmadas empiricamente, tanto a teoria quanto a interpretação dos escores em termos do construto são suportadas.

Existe na literatura instruções e procedimentos para uma adequada elaboração de itens e dos critérios para a pontuação dos testes. Uma forte preocupação ocorre para que se garanta a "equidade" do teste, ou seja, que se garantam oportunidades iguais e justas a todos os participantes da avaliação. Nesse sentido, há princípios estabelecidos que se destinam a auxiliar as pessoas responsáveis pelo desenvolvimento de avaliações a entenderem melhor o conceito de equidade na avaliação, evitando, desse modo, a inclusão de conteúdos ou imagens que podem provocar injustiças aos participantes do teste. Esses princípios atuam no sentido de evitar as fontes de variância construto-irrelevantes, importantes causadoras de erros nas análises dos resultados da avaliação. Essas fontes podem não ser as mesmas para todas as avaliações, dependem da utilização do teste, do público-alvo a que se destinam ou de outras variáveis. Há, no entanto, alguns princípios que devem ser aplicados em todos os testes, independentemente da utilização pretendida.

Itens mal formulados são importantes fontes de variância construto-irrelevante, que ocorre quando os testes representam também outros construtos além dos conhecimentos e habilidades que se pretendem medir. Essas influências causadoras de variância construto-irrelevante podem resultar em problemas com a dificuldade do item ou podem prejudicar a interpretação de pontuação (Messick, 1989).

É necessária, também, a preocupação com grupos de examinandos considerados minoritários na avaliação, por exemplo, grupos raciais, pessoas com deficiência visual, de audição ou outras, pessoas de idade mais avançada do que a maioria. Esses grupos são significativamente mais afetados por fontes de variância construto-irrelevante do que a população-alvo, diminuindo, desse modo, a justiça, assim como a equidade e a validade do teste. Nesse sentido, são recomendadas revisões para a remoção de fontes de variância construto-irrelevante que são identificadas por afetar grupos diferentes de formas diferentes (ETS, 2009).

Na sequência são apresentados definições e conceitos utilizados na área de avaliação visando à composição de testes de avaliação em larga escala, à elaboração dos critérios para a pontuação e dos procedimentos adotados

para a pontuação dos testes.

Critérios para a construção dos itens

1. O item deve expressar um comportamento, não uma abstração ou construto, e seu enunciado deve propor uma ação clara e precisa, de modo que se possa dizer ao candidato vá e faça. Por exemplo: Escreva sobre···; Reproduza···; Complete···; Calcule··· (PASQUALLI, 2010).
2. Um item deve medir apenas um único traço. Cada traço, no entanto, pode ser medido por um ou vários itens, de acordo com as especificações de teste (COHEN; WOLLACK, 2004).
3. O item deve ser escrito com objetividade, permitindo ao respondente mostrar se conhece a resposta ou se é capaz de executar a tarefa proposta (PASQUALLI, 2010; VIANNA, 1982; ANASTASI, 1977).
4. O item deve medir o que os examinandos sabem, não o que eles não sabem (COHEN; WOLLACK, 2004).
5. Os itens devem permitir ao examinando a possibilidade de concordar, discordar ou opinar sobre algum comportamento ou atitude, isto é, os itens devem expressar desejo ou preferência. Não existem, nesse caso, respostas certas ou erradas; existem, sim, diferentes gostos, preferências, sentimentos e modos de ser (PASQUALI, 2010).
6. Um item deve expressar uma única ideia. Itens que introduzem explicações de termos ou oferecem razões ou justificativas são normalmente confusos porque introduzem ideias variadas e confundem o respondente. Além disso, o item não deve apresentar informações adicionais ou complementares ao texto-base, quando este existir (PASQUALI, 2010; BRASIL, 2010; VIANNA, 1982; ANASTASI, 1977).
7. O item deve ser inteligível para todos da população-alvo, desde o extrato mais baixo até o mais alto. Devem-se utilizar frases curtas, preferencialmente afirmativas, com expressões simples e inequívocas (PASQUALI, 2010; BRASIL, 2010; ETS, 2009; DOWNING; HALADYNA, 1997).
8. O item deve abordar, preferencialmente, temas atuais e que sejam adequados ao público-alvo, evitando-se abordagens de temas que suscitem polêmicas ou que possuam conteúdos considerados sexistas, racistas, ofensivos ou inapropriados (GRAND *et al.*, 2013; BRASIL, 2010; ETS, 2009; DOWNING; HALADYNA, 1997).

9. O item deve ser composto de modo a refletir o fundo cultural de todos os participantes. Isso significa que o item deve possuir formato acessível a todos e não discriminar grupos minoritários ou subgrupos de participantes (GRAND *et al.*, 2013; ETS, 2009).

10. Deve-se ponderar o tempo demandado para a leitura do item durante a realização do exame; a extensão do enunciado, juntamente com a dos textos-base utilizados, deve ser considerada de acordo com a disponibilidade de tempo para a resposta à tarefa (BRASIL, 2010).

11. A sentença deve ser consistente com o traço que se deseja medir e com as outras frases que cobrem o mesmo atributo. Isto é, o item não deve insinuar atributo diferente do definido. O critério diz respeito à saturação que o item tem com o construto, representada pela carga fatorial na análise fatorial, que constitui a correlação entre o item e o traço (PASQUALI, 2010).

12. O item deve possuir uma posição definida no contínuo do atributo e ser distinto dos demais itens que cobrem o mesmo contínuo. Esse critério supõe que o item pode ser localizado em uma escala de habilidades; o item deve ter uma posição escalar modal definida e um desvio padrão reduzido. Em termos da Teoria da Resposta ao Item (TRI), esse critério representa os parâmetros "b" (dificuldade) e "a" (discriminação) e pode realmente ser avaliado de forma definitiva apenas após coleta de dados empíricos sobre os itens (PASQUALI, 2010).

13. Os itens devem possuir linguagem variada, pois o uso dos mesmos termos em todos os itens confunde as frases e dificulta diferenciá-las, além de provocar monotonia, cansaço e aborrecimento (PASQUALI, 2010).

14. O item não deve conter expressões extremadas (excelente, miserável, inteligentíssimo, etc.). A intensidade da reação da pessoa deve ser dada na resposta (PASQUALI, 2010).

15. As frases devem ser formadas com expressões condizentes com o atributo, não devem conter expressões ridículas, despropositadas ou infantis. Também não devem conter expressões humorísticas, pois podem sugerir ao participante não levar a avaliação a sério. Itens formulados erradamente podem fazer com que o respondente se sinta ofendido, irritado ou coisa similar, podendo contribuir para o aumento de erros de resposta (vieses). Esse tema é considerado para se ter a validade aparente (*face validity*) (GRAND *et al.*, 2013; PASQUALI, 2010; ETS, 2009; COHEN; WOLLACK, 2004).

16. É usual introduzir uma situação-problema ao propor a tarefa, que consiste na apresentação de um desafio instigando o examinando a um contexto reflexivo e à tomada de decisões requerendo a mobilização de recursos cognitivos e operações mentais. Uma situação-problema deve estar contextualizada de maneira que permita ao participante aproveitar e incorporar situações vivenciadas e valorizadas no contexto em que se originam (BRASIL, 2010).
17. Optando-se pela introdução de uma situação-problema, esta deve fazer parte de toda a estrutura do item, desde a escolha do texto-base até a construção de todas as partes que compõem um item. Um item contextualizado deve transportar o examinando para uma situação, muitas vezes hipotética, mas comumente vivenciada por ele no dia a dia (BRASIL, 2010).

Os critérios apresentados consideram a elaboração de cada item isoladamente, mas um teste deve também considerar critérios referentes ao conjunto dos itens como um todo. É importante que o instrumento discrimine entre indivíduos de diferentes níveis de habilidades, inclusive diferenciando entre os que estão situados próximos uns dos outros na escala de habilidades, e não somente entre os de maior habilidade em relação aos de menor habilidade.

O teste deve conter itens fáceis, médios e difíceis, distribuindo-se continuamente em toda a extensão da escala de habilidades. Os itens devem distribuir-se sobre a escala numa disposição que se assemelha à da curva normal: maior parte dos itens de dificuldade mediana e diminuindo progressivamente em direção às caudas (itens fáceis e itens difíceis em número menor). A razão desse critério encontra-se no fato de que a grande maioria dos traços latentes se distribui entre a população mais ou menos dentro da curva normal, isto é, a maioria das pessoas possuem magnitudes medianas dos traços latentes, sendo que umas poucas possuem magnitudes grandes e outras magnitudes pequenas (PASQUALI, 2010).

Uma vez que os construtos a serem avaliados foram definidos e embasados na literatura existente sobre o assunto e as tarefas foram construídas de modo a representar adequadamente esses construtos, é necessário estabelecer os critérios que serão utilizados para pontuação dos ensaios.

Elaboração dos critérios de avaliação

Por meio do desenvolvimento de critérios pré-definidos para o processo de julgamentos, é possível diminuir a subjetividade envolvida na avaliação com testes de respostas construídas (MOSKAL, 2000).

A lista exposta no Quadro 6 enumera as etapas, conforme descrito por Weigle (2002), para o desenvolvimento de critérios de pontuação para serem aplicados em avaliações em larga escala, de modo a assegurar a qualidade e a validade. Essa mesma lista também é descrita por Knoch (2011a).

Quadro 6 – Etapas para o desenvolvimento de critérios de avaliação

1. **Escolha do tipo de pontuação**: Deve-se decidir que tipo de abordagem é preferível para o evento: holística, analítica ou característica principal.
2. **Definição do propósito da avaliação**: Deve-se considerar a utilização dos resultados do teste para determinar se a formulação das definições é apropriada para o contexto e propósito da avaliação.
3. **Definição sobre quais aspectos do traço são mais importantes e como eles serão subdivididos**: É necessário definir uma escala e decidir quais critérios serão utilizados para as pontuações.
4. **Definição dos descritores e do número de níveis de pontuação que serão utilizados**: Muitos exames em larga escala utilizam entre seis e nove pontos de escala. Isso é determinado pelo conjunto de desempenhos que podem ser esperados e de quais resultados do teste serão utilizados. Também devem ser considerados a forma como os níveis da escala podem ser distinguidos uns dos outros e os tipos de descritores que serão utilizados.
5. **Definição de como as pontuações serão relatadas**: As pontuações analíticas podem ser apresentadas separadamente ou combinadas em uma pontuação total. A apresentação dos resultados deve ser relacionada à utilização dos escores de teste.

Fonte: Adaptado de Weigle (2002)

2.5.1.3 Análise teórica

A análise teórica deve contemplar as revisões dos itens e também dos critérios de pontuação. Para que as interpretações dos resultados da avaliação sejam válidas, deve-se ter cuidado com a elaboração dos itens do teste e dos critérios para a pontuação desses itens para que o instrumento, como um todo capture verdadeiramente a habilidade que se deseja medir de acordo com os objetivos da avaliação. Essas duas partes que compõem o instrumento são igualmente importantes, pois, se os critérios de pontuação estão mal elaborados, os avaliadores não conseguem atribuir pontuações confiáveis, mesmo se as tarefas estiverem de acordo com todos os requisitos estabelecidos para a

excelência. O mesmo ocorre com tarefas com problemas na formulação. Critérios de pontuação, por melhor projetados que estejam, não podem corrigir um teste mal concebido.

Mesmo que os elaboradores dos itens e dos critérios de pontuação tenham sido instruídos e treinados para a execução dessas tarefas, de acordo com os requisitos estabelecidos para a avaliação, é comum ainda restarem erros, como itens contendo conteúdos que não sejam totalmente indicados, problemas com a formulação das sentenças, tarefas que capturem habilidades que não se pretendem medir ou que exista alguma desconexão entre a tarefa e os critérios de pontuação, entre outros. São necessárias, então, revisões sistemáticas para detectar problemas que não foram evitados durante a elaboração do instrumento.

A validade do teste é muito dependente dos cuidados na fase de construção, etapa que tem recebido pouca atenção em comparação com a ênfase dada às análises dos resultados do teste. Borsboom, Mellenberg e Van Heerden (2004) sugerem que o problema principal na elaboração de instrumentos de avaliação é primeiramente saber o que deve ser medido, pois quando se sabe exatamente o que se pretende medir, então provavelmente sabe-se como medir, e assim será necessária pouca investigação para validar o instrumento. Desse modo, o problema para a validade não é descobrir o que é medido, mas sim determinar o que se pretende medir.

Os itens devem representar adequadamente o construto a ser avaliado. Então, nessa fase, o instrumento é submetido a especialistas para que eles expressem suas opiniões quanto à adequação dos itens ao construto a ser avaliado. Essas análises teóricas compreendem dois tipos de julgamentos denominados *validade de conteúdo* e *validade aparente*. A validade de conteúdo é determinada por peritos da área do construto e consiste em julgamentos desses especialistas sobre a pertinência do item para avaliar o construto em questão. A validade aparente, denominada por Pasquali (2010) *análise semântica*, também é feita por juízes, não necessariamente da área de definição do construto, e tem a finalidade de determinar se os itens são compreensíveis para todos os indivíduos da população. Para a validade aparente, pode-se também submeter os itens a amostras da população-alvo, devendo-se, nesse caso, ter cuidado para que essa amostra seja representativa de toda a população, com indivíduos pertencentes aos diferentes níveis de habilidade.

O procedimento de revisão de itens, utilizando peritos para análise e consequente correção ou exclusão de itens problemáticos, consiste em um método comumente abordado na literatura (GRAND, 2013; ETS, 2010; PASQUALLI, 2010; JOHNSTONE *et al.*, 2008; DOWNING; HALADYNA,

1997). Tais revisões permitem aos especialistas examinar a qualidade dos itens antes da sua utilização e são úteis principalmente para detectar conteúdos abordados nos itens que possam, de algum modo, desviar os examinandos de respostas que permitem inferências corretas sobre a sua posição na escala de habilidades para a medida do construto pretendido no teste. Desse modo, as revisões, no mínimo, devem garantir que o teste (1) reflita o fundo cultural tanto da maioria dos examinandos como de grupos considerados minoritários, (2) seja desprovido de conteúdo considerado sexista, racista, ofensivo ou inapropriado e (3) possua itens de formato acessível e não discriminatório, inclusive para grupos considerados minoritários (GRAND, 2013; ETS, 2010).

Essas revisões, na verdade, devem identificar fontes de variância construto-irrelevante, principal geradora de erros na interpretação dos resultados da avaliação e, segundo Messick (1989), uma das maiores ameaças à validade.

Para revisões eficientes, deve ser elaborado um conjunto de diretrizes que orientem os revisores em seus julgamentos. Esse conjunto de diretrizes é baseado em princípios existentes, normalmente citados na literatura, mas com conteúdos e exemplos localmente apropriados, resultando em orientações claras e específicas para a elaboração e também para a revisão de testes que sejam justos a todos os participantes.

Segundo a ETS (2009), um conjunto de três princípios cobre as possíveis fontes de variância construto-irrelevante: princípio cognitivo, princípio afetivo e princípio físico. Essas fontes devem ser evitadas e são descritas a seguir.

1. **Princípio cognitivo:** É responsável pelas fontes de variância construto--irrelevante decorrentes das diferenças entre as bases de conhecimento dos examinandos. Nesse caso, a variância na pontuação é causada quando, para se responder corretamente a um item, são necessários conhecimentos ou habilidades que não estão relacionados diretamente com o construto que o item foi desenvolvido para medir. Por exemplo, se o objetivo do item é avaliar a habilidade do indivíduo para efetuar a divisão de números, mas o enunciado é demasiadamente complexo, o correto entendimento desse texto é uma causa de variância construto-irrelevante. Se o objetivo do item, porém, é avaliar a habilidade de interpretação de texto, esse item pode ser apropriado e justo. É necessário determinar se os conhecimentos, as habilidades ou outros requisitos que o item exige para uma resposta correta são realmente importantes para a medição do construto pretendido ou são fontes de variância construto-irrelevante.

2. **Princípio afetivo:** É gerador de variância construto-irrelevante proveniente das diferenças nas reações emocionais dos examinandos. As fontes afetivas são indutoras de variância construto-irrelevante quando as imagens ou textos causam fortes emoções, podendo interferir na capacidade de responder ao item corretamente. Por exemplo, um texto com conteúdo ofensivo pode prejudicar a concentração do examinando na passagem que realmente importa para a resposta ao item do teste, sendo, assim, uma fonte de variância construto-irrelevante. Itens que defendem crenças ou posições políticas podem também ser fonte de variância construto-irrelevante, principalmente porque o examinando pode possuir posição contrária e responder ao item emocionalmente em vez de se concentrar logicamente na resposta. Desse modo, deve-se evitar a inclusão de conteúdo que parece ser ofensivo, perturbador, controverso, ou outros.
3. **Princípio físico:** É responsável por variância construto-irrelevante proveniente das diferenças de habilidades físicas dos examinandos. Essas fontes ocorrem principalmente para examinandos com alguma deficiência quando algum aspecto do item demandar habilidades como ver, ouvir, distinguir, ou outras. Por exemplo, os examinandos que podem enxergar mas possuem alguma deficiência visual podem ter dificuldade para entender um gráfico que possui informações escritas com fontes pequenas.

As diretrizes devem ser desenvolvidas de modo a contemplar todos os examinandos, no entanto alguns grupos requerem atenção especial no desenvolvimento, revisão e aplicação da avaliação. Indivíduos desses grupos são mais propensos do que outros a causarem variância construto-irrelevante, pois eles são mais suscetíveis a preconceitos, a diferenças culturais, a diferenças de formação, entre outras características importantes. Entre os grupos que devem ser considerados estão os caracterizados por idade, deficiência, etnia, sexo, região, língua materna, raça, religião, orientação sexual, nível socioeconômico. Dependendo da especificidade de cada avaliação, alguns desses grupos podem necessitar ou não de atenção especial, assim como outros grupos diferentes desses podem ser incluídos (GRAND *et al.*, 2013; ETS, 2009; DOWNING; HALADYNA, 1997).

As ameaças à validade de construto também são listadas como fontes de variância construto-irrelevantes, citadas anteriormente, e que devem ser removidas durante as revisões dos itens. Moskal e Leydens (2000) sugerem que as evidências para a validade de conteúdo e de construto podem ser constatadas por meio de respostas a algumas perguntas, as quais são descritas no Quadro 7.

Quadro 7 – Perguntas para examinar as evidências para a validade de conteúdo e de construto

Validade de conteúdo
1. *Os critérios abordam algum conteúdo estranho ao teste (que não se pretende medir)?*
2. *Os critérios de pontuação abordam todos os aspectos do conteúdo pretendido?*
3. *Há algum conteúdo abordado na tarefa que deveria ser avaliado, mas não é?*
Validade de construto
1. *Todas as características importantes do construto são medidas por meio dos critérios?*
2. *Algum dos critérios de avaliação é irrelevante para a medida do construto de interesse?*

Fonte: Adaptado de Moskal e Leydens (2000)

Critérios de pontuação bem definidos também são importantes para que a avaliação obtenha bons índices de confiabilidade, pois a normatização da pontuação assegura a consistência da pontuação independentemente do avaliador ou da ocasião na qual a pontuação foi atribuída (JOHNSTON, 2004; MOSKAL; LEYDENS, 2000; NYSTRAND; COHEN; DOWLING, 1993).

Moskal e Leydens (2000) também sugerem perguntas para avaliar se os critérios de pontuação são claros o suficiente para assegurar a qualidade da avaliação, no que tange à confiabilidade da correção. Essas perguntas estão expostas no Quadro 8.

Quadro 8 – Perguntas para examinar se os critérios de pontuação são adequados

Adequação dos critérios de pontuação
1. *As categorias de pontuação são bem definidas?*
2. *As diferenças entre as categorias de pontuação são claras?*
3. *Dois avaliadores independentes podem chegar à mesma pontuação para uma resposta dada com base na rubrica de pontuação?*

Fonte: Adaptado de Moskal e Leydens (2000)

Se a resposta a qualquer uma dessas perguntas for negativa, então as categorias de pontuação devem ser revistas. Esse processo também coincide

com a quarta etapa da lista formulada por Weigle (2002) (Quadro 6). É usual, para garantir que as categorias de pontuação sejam bem definidas, a utilização de níveis âncora, relacionando os pontos selecionados na escala de habilidades com descritores que ilustram as variações da rubrica de pontuação. Os níveis âncora são utilizados pelos avaliadores para esclarecer as diferenças entre os níveis de pontuação para cada rubrica (MOSKAL; LEYDENS, 2000).

2.5.2 Procedimentos Empíricos

Essa fase engloba o planejamento da aplicação do teste, a aplicação e a coleta dos dados. O planejamento e a aplicação do teste dependem da especificidade de cada ocasião e deverão considerar todas as variáveis envolvidas visando, principalmente, oportunidades iguais para todos os participantes. Essas preocupações envolvem o número de candidatos, o local de aplicação da avaliação e as condições físicas desse local, a forma de acesso dos participantes, entre outras inúmeras variáveis. Também é necessário proporcionar acesso, acomodações e condições adequadas para pessoas com deficiência.

Atualmente é inconcebível a organização de avaliações em larga escala sem equipes especializadas para os procedimentos computacionais, logísticos e pedagógicos. Os procedimentos computacionais estão relacionados com a construção de *site* para as informações da avaliação, a efetivação das inscrições, a geração de numeração ao inscrito, a geração do boleto bancário para o pagamento, a confirmação da inscrição, a divulgação do gabarito após os testes, o escore alcançado pelo participante, entre outros. O *site* da avaliação deve fornecer ao participante qualquer informação relacionada com o evento até que todas as etapas estejam concluídas. A equipe computacional é responsável também pelo processamento dos dados para as análises estatísticas da avaliação.

Entre as responsabilidades da equipe de logística estão a alocação de cada participante do teste em um local, sala e carteira, a seleção e o treinamento do pessoal de apoio: coordenadores, fiscais de sala, segurança, zeladores, leitores para os deficientes visuais (quando for o caso), as condições dos locais de prova, como limpeza, arrumação e iluminação das salas e banheiros, além do transporte seguro, de ida e de volta, dos cadernos de provas e outros materiais necessários para o evento.

Os procedimentos pedagógicos tratam de todos os processos na elaboração do instrumento de avaliação. Em algumas etapas, é necessário um

trabalho conjunto entre esses setores, por exemplo, a equipe computacional deve elaborar um banco de dados com nome, número de identificação e tipo de prova no qual cada candidato está inscrito. Esse banco de dados será utilizado pela equipe logística para distribuir cada candidato em uma sala, carteira e local, e a equipe pedagógica usará esse banco para gerar uma prova para cada candidato de acordo com a sua inscrição.

Neste trabalho, os assuntos relacionados com os procedimentos computacionais e os procedimentos logísticos não serão tratados com profundidade por não fazerem parte dos objetivos desta pesquisa e por necessitarem de profissionais de outras áreas. Desse modo, nessa etapa de planejamento da aplicação da avaliação, são tratados apenas os procedimentos sob a responsabilidade da equipe pedagógica das empresas, que correspondem à diagramação do caderno de provas e à sua impressão.

2.5.2.1 Diagramação dos cadernos de provas

Os responsáveis pela montagem do instrumento de avaliação devem ter atenção com uma série de detalhes, aparentemente sem muita importância, mas que podem influenciar o desempenho do examinando (VIANNA, 1982).

1. *Dimensões do caderno de provas e determinação do número de páginas*: Essas especificações afetam diretamente os custos da avaliação. Em avaliações em larga escala, uma página a mais a ser impressa pode elevar consideravelmente os custos e significar horas e às vezes dias a mais de trabalho, dependendo das especificações da impressora na qual serão feitas as cópias. Os cadernos de provas em avaliações em larga escala brasileira comumente utilizam papel de tamanho A4 (210×297 mm), como os vestibulares das principais universidades e o ENEM, mas eles podem ser de outros formatos. Deve-se ter a preocupação com a qualidade visual do instrumento, incluindo tamanho e tipo da fonte, espaçamento entre os itens, espaço adequado para as respostas, etc. O instrumento deve ser igualmente acessível a todos os participantes, inclusive aos portadores de deficiências. Nas provas de redação ou provas com itens abertos, é usual a disponibilidade de espaços em branco para rascunho, normalmente do mesmo tamanho dos disponíveis para as respostas definitivas. Muitas vezes, são distribuídos dois cadernos aos participantes: um caderno com as instruções, os itens e os espaços para os rascunhos das respostas e outro para as respostas definitivas.

2. *Capa do caderno de provas*: As informações que devem estar disponíveis na capa dependem do tipo de avaliação. Se a finalidade da avaliação é a seleção para vaga de trabalho, por exemplo, devem constar na capa a empresa que está disponibilizando as vagas, o título do certame, o cargo a que se destina aquele caderno; mas se a avaliação é para vaga em curso superior, devem constar na capa o nome da universidade e o nome do curso no qual o participante está inscrito. Em algumas avaliações, há cadernos de provas com montagens diferentes para dificultar a cópia entre os participantes. Nesse caso, deve ser indicado na capa o tipo da prova que o examinando está fazendo. O ENEM utiliza cadernos de provas de cores diferentes e algumas universidades utilizam números para diferenciá-las. A capa deve conter também instruções gerais para a elaboração das respostas e outras informações relacionadas com o caderno, como o número de itens e de páginas. Além dessas informações, algumas empresas provedoras possuem tecnologia para impressões individualizadas e apresentam na capa as informações de cada participante, como o nome, número de identificação, local e carteira disponibilizada a ele.
3. *Disposição dos itens*: A prova pode ser escrita em duas colunas ou em uma, utilizando-se toda a linha. Em avaliações com itens abertos, é mais comum a utilização de toda a linha para a apresentação dos elementos textuais, no entanto, se não prejudicar a legibilidade, os itens podem ser apresentados em duas colunas, se esse formato representar alguma economia de espaço.

   Preferencialmente, o item deve ser escrito inteiramente na mesma página, pois seu fracionamento em páginas ou colunas diferentes pode representar problemas para o examinando e deve sempre ser evitado. Os materiais informativos, como gráficos, textos, figuras, tabelas, entre outros, devem, se possível, ser apresentados na mesma página do item. Quando isso não for possível, deve-se garantir que eles estejam em paginação dupla, isto é, o item deve ter seu início no verso de uma página e continuar na frente da página seguinte para que todas as informações estejam visíveis sem a necessidade de virar a página.

2.5.2.2 Impressão dos cadernos de provas

Em avaliações em larga escala, é muito importante a preocupação com a qualidade da impressão e também com a segurança da avaliação. Normalmente o setor das empresas promotoras de avaliações, no qual os testes são

elaborados, são fechados e o acesso é restrito às pessoas credenciadas.

A impressão e o armazenamento dos cadernos de provas são muito suscetíveis às fraudes. Não é raro notícias veiculadas na imprensa relatando o comércio de cópias de provas de concursos públicos. O desvio de cópias deve ser evitado a todo custo. Essa possibilidade não pode ser menosprezada em avaliações em larga escala, pois os prejuízos, financeiros ou não, podem ser incalculáveis. Entre eles estão a possibilidade de anulação da avaliação e a necessidade de elaboração de novo exame, prejuízo na imagem da instituição promotora, prejuízo à empresa ou instituição contratante da avaliação, processos judiciais e eventuais custos com indenizações. Um exemplo brasileiro importante de desvio de cópias de provas na fase da impressão é o caso do ENEM no ano de 2009. Esse crime causou um prejuízo financeiro imenso ao governo brasileiro. As provas tiveram que ser refeitas e a aplicação da avaliação adiada (MOREIRA, 2011). Essa mudança de datas do exame afetou inclusive o calendário das instituições que utilizam os resultados do ENEM em seus exames vestibulares, sem falar do prejuízo e dos transtornos causados aos estudantes.

Algumas empresas responsáveis pela elaboração de avaliações possuem sua própria gráfica para que não seja necessário que os instrumentos de avaliação, completamente prontos nessa ocasião, saiam do setor responsável correndo riscos desnecessários de segurança. Nesse caso, todo o processo é feito "dentro de casa", e a equipe pedagógica da avaliação é responsável também pela impressão, organização e armazenamento dos cadernos de provas.

Outras empresas terceirizam essa tarefa contratando gráfica especializada. Nesse caso, alguns procedimentos auxiliam a melhorar os padrões de segurança nas etapas de impressão e armazenamento.

- O contrato de terceirização da impressão deve exigir exclusividade, isto é, apenas o material da avaliação é impresso no período contratado.
- Presença obrigatória de pessoas da equipe pedagógica acompanhando todo o processo e garantindo que nenhuma cópia seja desviada ou extraviada.
- O número de pessoas, funcionários da gráfica, deve ser reduzido o máximo possível, e essas pessoas devem assinar um documento se comprometendo com o sigilo e a segurança das provas.
- O trabalho de verificação da qualidade da impressão deve ser feito por membros da equipe pedagógica simultaneamente com a impressão. Os funcionários da gráfica não devem ler partes dos testes.

- Quando problemas de impressão são detectados, o lote deve ser impresso novamente e os cadernos com defeitos devem ser imediatamente destruídos no fragmentador de papéis por um membro da equipe pedagógica. O mesmo deve ser feito com qualquer papel excedente ou rejeitado pela impressora.

- Os cadernos de provas devem ser armazenados em malotes lacrados e devidamente identificados.

- O transporte das provas prontas até um local seguro também exige cuidados, muitas vezes é indicado utilizar-se de escolta.

Muitas impressoras de grande porte imprimem provas individuais para cada participante da avaliação por meio de banco de dados fornecido pelo setor computacional da empresa provedora. Nesse caso, o caderno de provas é impresso contendo dados relacionados com a identificação do candidato, as informações sobre o tipo de prova no qual ele foi inscrito, o local, a sala e a carteira em que fará o teste. Esse tipo de impressão auxilia demasiadamente a segurança da prova, pois nenhum caderno pode faltar ou mesmo sobrar. Além disso, para melhorar ainda mais a segurança, todos os cadernos de provas de uma mesma sala deverão ser agrupados e guardados em envelopes com lacres de segurança. Os envelopes contendo as provas de cada sala são armazenados em caixas identificadas com o local no qual aquelas provas serão aplicadas, e essas caixas também são guardadas em malotes lacrados, que são colocados em local seguro até o dia do evento.

A qualidade da impressão deve ser monitorada ao mesmo tempo que elas são feitas. Esse trabalho consiste em conferir a qualidade da impressão de alguns cadernos aleatoriamente. Além disso, quando as provas são personalizadas, as equipes de logística e computacional fornecem listas contendo o número das salas, o número de participantes alocados em cada sala, o nome de cada participante, o número da sua carteira e o tipo de prova na qual ele foi inscrito. Todas essas informações devem ser conferidas, ao menos de alguns cadernos de cada lote antes do armazenamento nos envelopes com lacres de segurança, que também são identificados com informações referentes a local, número da sala e número de participantes. Apesar de todos os cuidados, recomenda-se que seja disponibilizado para cada local de prova um excedente de 10% de cada caderno, para eventuais substituições quando algum problema com a qualidade da impressão é constatado na hora da aplicação do teste.

2.5.2.3 Pontuação dos testes e treinamento dos avaliadores

A coleta de dados, quando a avaliação é de itens abertos, é resultante da pontuação atribuída por avaliadores especialistas e treinados. Desse modo, para a obtenção de dados confiáveis, a preocupação principal dentro desse procedimento é direcionada ao treinamento dos avaliadores.

Para estabelecer classificações confiáveis em avaliações, principalmente as em larga escala, é necessário o treinamento dos avaliadores. Segundo Weigle (1999), as características pessoais dos avaliadores, tais como cultura, experiências, expectativas, estilo de correção entre outras variáveis, podem influenciar substancialmente a pontuação das tarefas, e as expectativas dos avaliadores podem ser tão importantes para a pontuação quanto a qualidade da resposta escrita pelo participante da avaliação. O treinamento é frequentemente citado como um meio para compensar diferenças diversas entre os avaliadores e ajustar suas expectativas, diminuindo a variabilidade na pontuação (HUOT, 1990). Além disso, o treinamento de avaliadores possibilita um entendimento comum sobre os critérios de avaliação e a interpretação dos descritores, tais como esses foram originalmente pensados pelos desenvolvedores do teste. A prática de estratégias de pontuação também facilita aos avaliadores inexperientes um comportamento aproximado aos dos avaliadores mais experientes (HARSCH; MARTIN, 2012; WEIGLE, 2002; COHEN, 1994).

A maior parte dos estudos divulgados sobre a pontuação de testes com itens abertos está relacionada com a pontuação de tarefas de escrita, talvez porque o alcance de índices altos de confiabilidade geralmente é mais difícil do que em outras disciplinas. As tarefas de escrita dependem fortemente de outras habilidades além das que estão sendo medidas como leitura e interpretação de textos, conhecimento de mundo e pensamento crítico. Além disso, mesmo quando os avaliadores são rigorosamente treinados em um conjunto claro de critérios de pontuação, alguma taxa de subjetividade sempre é constatada na pontuação (WEIGLE, 1999), pois as avaliações da escrita são mais dependentes das interpretações e decisões pessoais dos avaliadores (WISEMAN, 2012; MYFORD; WOLFE, 2000). Mesmo assim, os procedimentos que são adotados nas avaliações da escrita devem ser estendidos a avaliações com itens abertos de outras disciplinas para a obtenção de pontuações confiáveis.

A experiência dos avaliadores é tida como um fator importante na obtenção de altos índices de confiabilidade, mas também é necessário levar em conta as tendências dos avaliadores em julgamentos sistemáticos dos desem-

penhos avaliados. Essas tendências são comportamentos frequentemente citados nas pesquisas e são consideradas componentes geradores de erros importantes na pontuação de tarefas escritas. Os efeitos mais discutidos causados por essas tendências dos avaliadores são: efeito da severidade/complacência, que é a tendência dos avaliadores em avaliar de maneira muito exigente ou muito branda as tarefas elaboradas pelos examinandos; efeito de tendência central, que é a tendência dos avaliadores de classificações iguais ou perto do ponto médio da escala, evitando, desse modo, classificações nos extremos da escala; efeito de aleatoriedade, que é a tendência do avaliador em aplicar uma ou mais categorias da escala de maneira inconsistente com o modo com que os outros avaliadores aplicam a mesma escala; efeito halo, que é a tendência dos avaliadores em atribuir pontuações semelhantes para todos os examinandos para o mesmo item, desse modo, desempenhos muito diferentes podem receber pontuações semelhantes; efeito de viés, também denominado efeito de severidade/complacência diferencial, é a tendência dos avaliadores em julgar de forma discriminatória, atribuindo pontuações a um determinado grupo, em média, menores ou maiores do que as pontuações atribuídas pelos outros avaliadores a esse grupo (KNOCK; READ; RANDOW, 2007, MYFORD; WOLFE, 2000, 2004; ENGELHARD; MYFORD, 2003).

Apesar de esses comportamentos serem resistentes, estudos demonstram que os efeitos causados podem ser minimizados após o treinamento dos avaliadores. Aliás, são esperadas pequenas discrepâncias ao invés de pontuações idênticas. Com treinamentos eficientes e sessões de correção cuidadosamente monitoradas, podem ser alcançados níveis relativamente elevados de consistência entre os avaliadores (EAST, 2009; HUANG, 2008; WEIGLE, 1999).

Basicamente, para realizar treinamento de avaliadores, são destacadas duas abordagens.

a) A primeira é o grupo hierárquico de coordenação, em que o avaliador coordenador decide como os critérios de pontuação devem ser interpretados. São utilizados exemplos de tarefas pré-avaliadas, cujas normas não devem ser discutidas ou negociadas, mas, simplesmente, aceitas e internalizadas.

b) A segunda possibilidade é uma abordagem, baseada em reuniões de consensuais de coordenação: primeiramente, alguns textos são analisados, e as principais características são destacadas e discutidas a fim de desenvolver um entendimento comum para a aplicação dos descritores para as classificações (HARSCH; MARTIN, 2012; GREATOREX; BAIRD; BELL, 2004).

O método de treinamento de avaliadores e procedimentos de classificação, comumente utilizado nas avaliações em larga escala, consiste basicamente no seguinte:

1. Alguns textos escritos pelos participantes da avaliação são escolhidos pela equipe de coordenação da correção para exemplificar o desempenho esperado em cada nível âncora e evidenciar as fronteiras entre os níveis sucessivos. Esses níveis foram previamente determinados na elaboração da rubrica (ESFANDIARI; MYFORD, 2013; HARSCH; MARTIN, 2012; HARSCH; HUPP, 2011; HUANG, 2008; KNOCH; READ; RANDOW, 2007; MYFORD; WOLFE, 2000; GEARHART 1995).

2. Os critérios de pontuação são explicados aos avaliadores, os quais são treinados para usá-las com a correção de um conjunto de textos utilizados como referência. A abordagem escolhida pode ser a hierárquica (MYFORD; WOLFE, 2000) ou a consensual (ESFANDIARI; MYFORD, 2013; GEARHART 1995). Durante esse treinamento, frequentemente, os avaliadores são organizados em pequenos grupos, cada um com um avaliador mais experiente (líder de grupo), o qual tem a função de esclarecer procedimentos e sanar eventuais dúvidas que possam surgir. Essa pontuação experimental deve atingir um nível de precisão predeterminado antes do início da sessão de pontuação efetiva.

3. As pontuações finais podem ser feitas de três maneiras distintas:

    i. As tarefas são pontuadas por dois ou mais avaliadores independentes (sem comunicação ou discussão) e a pontuação final é a média aritmética dessas pontuações. A acuracidade da pontuação entre os avaliadores é monitorada continuamente pela diferença entre as pontuações atribuídas a cada texto; quando essa diferença for maior do que um valor predeterminado, é detectada uma discrepância. Esse procedimento auxilia a determinar se os avaliadores estão mantendo o padrão de correção ao longo de cada sessão de pontuação (KOBRIN; DENG; SHAW, 2011; BARKAOUI, 2007; MYFORD; WOLFE, 2000; GEARHART 1995).

    ii. As tarefas são pontuadas por uma equipe formada por dois ou mais avaliadores. Um dos membros da equipe lê o texto em voz alta, então cada avaliador registra a sua pontuação de acordo com a rubrica estabelecida. Se necessário, os avaliadores podem reler o texto individualmente (MYFORD; WOLFE, 2000; JENNINGS *et al.*, 1999).

iii. As tarefas são pontuadas apenas por um avaliador. Nesse caso, alguns textos previamente pontuados são distribuídos aleatoriamente entre aqueles a serem corrigidos pela primeira vez, desse modo a confiabilidade da pontuação pode ser monitorada. Esse procedimento auxilia a determinar se algum avaliador está pontuando fora do padrão de correção estabelecido e também a identificar discrepâncias.

4. Quando é detectado algum avaliador pontuando fora do padrão estabelecido, o líder do grupo pode reler os textos corrigidos juntamente com o avaliador, monitorando seu desempenho e detectando os pontos de divergência ao padrão comum. Esse acompanhamento é feito até que o problema seja corrigido (ESFANDIARI; MYFORD, 2013).
5. As discrepâncias podem ser tratadas em um dos modos a seguir:

   i) A tarefa é corrigida novamente pelos mesmos avaliadores que atribuíram as pontuações inicialmente, mas agora em conjunto. Eles explicam e ponderam sobre as notas atribuídas em busca de consenso (JENNINGS *et al.*, 1999; GEARHART 1995).

   ii) A tarefa é corrigida novamente pelo avaliador líder da sessão, que ouve os argumentos dos avaliadores e pontua a tarefa (HUANG, 2008).

   iii) A tarefa é corrigida novamente por um novo avaliador independente (BARKAOUI, 2007).

   iv) A tarefa é corrigida novamente por uma banca formada por corretores experientes.

6. Regras para resolver problemas, como, por exemplo, respostas anormais, podem ser formuladas simultaneamente e comunicadas verbalmente aos membros do grupo ou mesmo escritas no quadro de avisos.

Se todo o processo da avaliação, inclusive o da pontuação, é replicado em outra ocasião, é possível que o resultado não seja exatamente o mesmo, o que prejudica a comparabilidade das avaliações de uma ocasião para outra. É possível distribuir alguns textos pontuados na avaliação anterior juntamente com os novos trabalhos para detectar qualquer divergência na pontuação. Mas se algum aspecto da tarefa muda de uma ocasião de avaliação para a outra, pode não ser possível, para os avaliadores, fazer julgamentos equivalentes, pois, para isso, é necessário estabelecer parâmetros predefinidos (KNOCH, 2011a; HAERTEL; LINN, 1996).

2.5.3 Procedimentos Analíticos

Os procedimentos analíticos consistem nas análises estatísticas aplicadas sobre os dados coletados. No caso das avaliações educacionais com itens abertos, esses dados são o resultado das pontuações atribuídas às respostas dos participantes.

Uma pontuação precisa de tarefas escritas é uma necessidade em toda avaliação, mas a dificuldade maior reside nas avaliações em larga escala, que ainda enfrentam graves desafios em suas concepções. Vários mecanismos têm sido utilizados eficazmente para melhorar a precisão da pontuação, entre eles estão a escolha dos critérios de pontuação adequados (BARKAUOUI, 2011; HAMP-LYONS, 2011; REZAEI; LOVORN, 2010; SLOMP, 2005; WEIGLE, 2002; HUOT, 1990), a utilização de um número razoável de pontos na escala (MYFORD, 2002; PENNY; JOHNSON; GORDON, 2000; NORTH, 2000), a inclusão de números decimais na escala (CROMBACH *et al.*, 1995) e o treinamento de avaliadores (JOHNSTON, 2004; NYSTRAND; COHEN; DOWLING, 1993).

Nas seções seguintes, são estabelecidos, de modo geral, os procedimentos relacionados com as análises da validade do instrumento por meio de análises empíricas da qualidade dos itens, da confiabilidade nas pontuações de tarefas e a determinação do grau de dificuldade dos itens de respostas construídas.

2.5.3.1 Validade da avaliação

McNamara (2000) caracteriza a validade como uma avaliação do próprio teste e a define como o processo para investigar os procedimentos pelos quais decisões são tomadas a partir das inferências feitas sobre os resultados do teste. Segundo o autor,

> A validação de um teste envolve o pensar na lógica do teste, especialmente em seu *design* e em suas intenções, e também envolve olhar para as evidências empíricas – os fatos – que emergem dos dados advindos de um julgamento do teste ou de administrações operacionais. Se não houver procedimentos de validação disponíveis, há potencial para parcialidades e injustiças. Esse potencial é significativo em proporção ao que está em jogo [3]

[3] Test validation similarly involves thinking about the logic of the test, particularly its design

As inferências sobre os resultados do teste frequentemente vão muito além dos desempenhos observados. Os resultados dos testes não são utilizados simplesmente para relatar como um indivíduo se saiu ao responder alguns itens em determinado momento e sob certas condições. Ao contrário, as pontuações do teste são usadas para apoiar afirmações diversas, como, por exemplo, afirmar que um indivíduo possui certo nível de habilidade em algum construto ou possui alguma probabilidade de sucesso em um programa educacional ou outra atividade. Essas afirmações geralmente não são evidentes nas avaliações. É necessário avaliar a plausibilidade das afirmações com base nos resultados dos testes para validar as interpretações e utilizações desses resultados (KANE, 2013).

A confiabilidade da avaliação é considerada como uma condição necessária para a validade, mas não suficiente. Essa afirmação é derivada do fato de que, se a pontuação dos testes varia substancialmente quando se repetem os procedimentos, é difícil fazer inferências consistentes sobre os resultados do teste. Desse modo, as exigências sobre a qualidade das avaliações devem estar sempre presentes, independentemente de objetivos, finalidade ou abrangência da avaliação (KANE, 2013).

A precisão na classificação de proficiência dos examinandos está relacionada com o fato de as decisões baseadas nos resultados dos testes corresponderem às decisões que teriam sido tomadas se as pontuações fossem livres de erros de medição. Como é muito difícil a obtenção de testes livres de erros, principalmente em áreas educacionais, é necessário estimar a precisão com que ocorrem as classificações dos examinandos em relação às suas habilidades.

Para isso, pode ser utilizada a comparação entre os resultados de testes paralelos. Se os indivíduos são classificados de forma aproximada em duas formas de testes equivalentes, a precisão da classificação é alta. A desvantagem maior desse método reside na dificuldade de aplicar dois testes, que medem as mesmas habilidades, aos mesmos examinandos em uma mesma ocasião. Assim, a precisão da classificação tem de ser avaliada com base na aplicação de um teste único.

Certo número de procedimentos para avaliar a confiabilidade da pontuação, e consequentemente a precisão da classificação, foi desenvolvido com base na Teoria Clássica de Testes (TCT), entretanto, procedimentos que uti-

---

and its intentions, and also involves looking at empirical evidence – the hard facts – emerging from data from test trials or operational administrations. If no validation procedures are available there is potential for unfairness and injustice. This potential is significant in proportion to what is at stake (McNAMARA, 2000, p. 48, tradução nossa).

lizam modelos derivados da Teoria de Resposta ao Item (TRI) estão sendo cada vez mais utilizados. Um dos métodos que está recebendo muita atenção nas pesquisas recentes para a determinação da precisão em que são feitos os julgamentos nas avaliações com itens abertos utiliza o modelo multifacetas de Rasch. Esse modelo é uma extensão do modelo de Rasch, que é o modelo da TRI e um parâmetro.

Nas avaliações com itens de respostas construídas, são muitos os fatores que podem afetar a medida do desempenho das pessoas ao executar a tarefa determinada no teste. Em primeiro lugar, está a habilidade do examinando, mas a pontuação que ele receberá no exame não depende apenas da sua capacidade ou do conhecimento sobre o construto que está sendo medido, depende também da severidade do avaliador, da dificuldade das tarefas, do formato da questão, do tema abordado, dos critérios e da escala de pontuação e de outras variáveis que podem interferir em cada evento de avaliação em particular.

Esses e outros fatores são frequentemente constatados em estudos relacionados com avaliações com itens abertos, principalmente nas avaliações da linguagem de primeira e segunda língua. Alguns exemplos podem ser obtidos nos trabalhos de Huang (2012), Rezai e Lovorn (2010), Gyagenda e Engelhard (2009), Jonsson e Svigby (2007), Sudweeks, Reeve e Bradshaw (2005) e Weigle (1999).

Alguns procedimentos estatísticos para avaliar a confiabilidade da pontuação baseados na TCT são expostos na Seção 2.5.3.2. O modelo multifacetas de Rasch, entretanto, está se mostrando uma ferramenta superior às fornecidas pela TCT para as análises de dados provenientes das avaliações com itens abertos, por permitir análises tanto no nível de grupo quanto no nível individual.

As análises para os efeitos individuais causados por cada elemento que faz parte da avaliação, ou seja, cada examinando, cada avaliador, cada uma das tarefas, cada critério de pontuação utilizado, entre outros, fornecem a possibilidade de obter informações que possam servir de diagnóstico, no nível individual, sobre o funcionamento de cada elemento em particular. Essa é uma vantagem valiosa sobre outros métodos e torna especial a utilização do modelo multifacetas de Rasch nas avaliações com itens abertos. Nas avaliações da linguagem, a utilização do modelo multifacetas de Rasch tem possibilitado o levantamento sobre o modo como cada avaliador pontua cada uma das tarefas elaboradas pelos examinandos, possibilitando inclusive a detecção de efeitos nas pontuações de difícil diagnóstico, por se apresentarem camuflados.

Por esse motivo, e também por apresentar outras vantagens, o modelo multifacetas de Rasch tem se tornado popular em avaliações da linguagem (MACNAMARA; KNOCH, 2012; SUDWEKS; REEVE; BRADSHAW, 2005; MYFORD, 2002), nas avaliações de inglês para estrangeiros (LIM, 2011; JOHNSON; LIM, 2009; MYFORD; WOLF, 2000; WEIGLE, 1999) e também em análises de avaliações que necessitam do julgamento de avaliadores em diversas áreas, como, por exemplo, para estudo das habilidades essenciais para a escrita criativa (BARDOT *et al.*, 2012), estudos sobre a criatividade (HUNG; CHEN; CHEN, 2012), avaliações orais (VAN MOERE, 2006), análise comportamental em relação a alimentos doces e salgados (VIANELLO; ROBUSTO, 2010), avaliação do desempenho médico (McMANUS; ELDER; DACRE, 2013; LUNZ; WRIGHT, 1997), estudos turísticos (PARRA-LÓPES; OREJA-RODRÍGUES, 2014), desempenho na patinação artística (LINACRE, 2002b).

Algumas estatísticas são utilizadas com o objetivo de avaliar a adequação dos dados aos modelos de Rasch e também a qualidade das pontuações provenientes dos avaliadores, a qualidade e dificuldade dos itens, a qualidade dos critérios e das escalas de classificação utilizadas, entre outros. Esses índices podem auxiliar na determinação da qualidade da avaliação e consequentemente apoiar a sua validação. Essas estatísticas estão organizadas em três grupos: (1) *Estatísticas de ajuste*, que indicam o grau com que as pontuações observadas se aproximam das pontuações esperadas que são geradas pelo modelo multifacetas de Rasch; (2) *Estatísticas de separação*, que indicam o quanto os elementos da avaliação estão separados entre si (examinandos, avaliadores, itens, etc.); (3) *Médias justas e observadas*, que auxiliam na obtenção de uma interpretação entre as diferenças nas medidas dos elementos participantes da avaliação e suas implicações. Essas medidas podem ser obtidas para todas as variáveis incluídas no modelo e que fazem parte do sistema de avaliação como um todo. Essas estatísticas são descritas no Capítulo 3, Seções 3.5.1, 3.5.2 e 3.5.3, respectivamente.

O modelo multifacetas de Rasch, por ser uma extensão do modelo de Rasch, deve ser utilizado em testes que medem a proficiência dos indivíduos em uma única dimensão do construto. Ou seja, os modelos de Rasch são modelos unidimensionais. Quando os resultados da avaliação estão ajustados com os resultados esperados pelo modelo, o pressuposto da unidimensionalidade é suportado (ECKES, 2011; SMITH, 1998; TENNANT; PALLANT, 2006). No entanto, diferenças significativas entre os valores esperados pelo modelo de Rasch e os valores observados podem ocorrer por diversas razões, não significando de imediato que a causa seja a multidimensionalidade, para

tanto são necessárias outras análises.

No contexto das medidas de Rasch, existem algumas abordagens para testar a unidimensionalidade (TENNANT; PALLANT, 2006; LINACRE, 1998; SMITH, 1998). A maioria dessas abordagens se baseia em análises do ajuste dos dados ao modelo de Rasch. Quando os dados estão em conformidade com o modelo de Rasch, toda variação sistemática detectada nos dados é explicada por uma única dimensão. Os resíduos calculados para as pessoas e itens, a partir das observações em uma única dimensão, possuem uma estrutura aleatória normal e variância previsível. Consequentemente, os residuais calculados para os pares de itens, por meio das pessoas, não estão correlacionados. Essa característica é o que define a *independência local*. No contexto de análises de traços latentes, ou, ainda, das medidas de Rasch, independência local é modelada para manter cada pessoa em pontos correspondentes sobre a variável latente (LINACRE, 1998).

Uma vez que a muldimensionalidade é manifestada pelo comportamento dos dados, esses dados devem ser examinados. Após a construção das medidas de Rasch, um valor esperado pode ser calculado para cada observação. O residual da observação é a diferença entre a observação e a expectativa dessa observação. Analisando os padrões entre esses resíduos, podem-se identificar valores que indicam a ocorrência de multidimensionalidade relevante (LINACRE, 1998). Segundo Wright (1995 *apud* LINACRE, 1998), "A análise do ajuste dos dados para a (independência local) é o dispositivo estatístico pelo qual os dados são avaliados quanto ao seu potencial de medição – para sua validade medição".

Um dos métodos para testar a unidimensionalidade consiste em examinar os índices médias quadráticas *infit* e *outfit*. Neste trabalho, essas estatísticas encontram-se definidas no Capítulo 3, Seção 3.5.1. Valores desses índices relativamente diferentes de seus valores esperados podem representar sintomas de multidimensionalidade no teste.

Os valores *infit* e *outfit* podem ser estimados para cada examinando, cada avaliador, cada critério e são sensíveis para detectar desvios em relação aos valores esperados de acordo com o modelo de Rasch. Por exemplo, as análises sobre a dificuldade relativa de cada critério podem indicar multidimensionalidade uma vez que os critérios devem trabalhar juntos para definir uma única dimensão do traço latente. Embora os desvios possam ser causados por uma série de fatores, um desses fatores poderá ser a multidimensionalidade do construto (ECKES, 2011).

Existem muitos motivos nos quais os valores observados podem diferir dos valores esperados calculados pelo modelo de Rasch, por isso tem sido

sugerido que as diferenças mais grosseiras sejam investigadas em primeiro lugar. Linacre (1998) sugere um processo em três fases para as análises dos dados com desvios grosseiros: (1) corrigir contradições sistemáticas às medidas de Rasch, que normalmente são sinalizadas por correlações bisseriais negativas; (2) diagnosticar pessoas e itens idiossincráticos por meio das estatísticas de ajuste como as médias quadráticas *infit* e *outfit*; (3) procurar por multidimensionalidade.

No Capítulo 3 (Seção 3.6) são apontadas análises que devem ser feitas para aferir a qualidade de uma avaliação com itens de respostas construídas no contexto do modelo multifacetas de Rasch.

2.5.3.2 Confiabilidade da pontuação

A estimação do grau de concordância entre avaliadores quanto à pontuação atribuída é importante em todas as avaliações que envolvem avaliadores, pois a confiabilidade da pontuação resulta na validade dos resultados da avaliação. Se dois avaliadores não podem concordar em suas pontuações para avaliar indivíduos com base em comportamentos observados, então as inferências obtidas das notas dadas pelos avaliadores não terão validade. Além disso, a confiabilidade interavaliador deve ser verificada em cada evento da avaliação, mesmo que o instrumento de avaliação e a rubrica de pontuação que estão sendo utilizados encontrem-se testados e comprovadamente eficazes para o evento em questão. Isso ocorre porque a confiabilidade interavaliador refere-se ao grau em que um determinado conjunto de avaliadores pode concordar em cada circunstância de teste em particular. A confiabilidade interavaliador é uma propriedade de cada situação de teste, e não do instrumento de avaliação (STEMLER, 2004).

Existem vários métodos estatísticos frequentemente citados na literatura para determinar a precisão e a consistência da pontuação atribuída por vários avaliadores. Stemler (2004) classifica esses métodos em três abordagens principais: (1) estimativas de consenso, que indicam o grau em que os avaliadores atribuem as mesmas pontuações, essa é a indicação de confiabilidade interavaliador; (2) estimativas de consistência, que indicam o grau em que o padrão de pontos (notas altas e notas baixas) atribuídos por cada avaliador é semelhante entre si, essa é a confiabilidade intra-avaliador e (3) estimativas de medição, que indicam o grau em que os resultados, e não as componentes de erro, podem ser atribuídos à pontuação final. O autor argumenta que a determinação da consistência da pontuação exige que todas

as três abordagens sejam satisfeitas uma vez que cada uma delas possui particularidades que implicam na forma como os dados provenientes de vários avaliadores são resumidos.

A primeira abordagem, as estimativas de consenso, é utilizada quando os avaliadores são treinados para julgamentos sistemáticos baseados em critérios de pontuação em escalas contínuas que representam o desempenho do indivíduo quanto ao construto avaliado.

Um dos métodos mais populares para calcular essas estimativas é por meio de porcentagens exatas, no qual se soma o número de casos que receberam a mesma pontuação por dois avaliadores distintos e divide-se esse número pelo número total de casos classificados pelos dois avaliadores. Quando os valores encontrados forem superiores a 70%, então pode-se considerar a pontuação confiável (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004). As principais vantagens desse método são: facilidade de cálculos, facilidade de compreensão e forte apelo intuitivo. A principal desvantagem é que, muitas vezes, o treinamento dos avaliadores para se ter uma concordância perfeita é trabalhoso e demorado.

Uma modificação desse método, utilizada como meio de driblar essa desvantagem, é o método denominado porcentagens adjacentes. Envolve ampliar a definição de acordo, incluindo as categorias de pontuação adjacentes na escala de classificação, isto é, os avaliadores não precisam chegar a um acordo exato sobre a classificação, que pode diferir em alguns pontos. Dependendo do comprimento da escala e da precisão que se deseja na avaliação, essa diferença poderá ser maior ou menor. Uma desvantagem desse método é que ele pode levar a estimativas de confiabilidade entre avaliadores exageradas, por exemplo, se a escala tiver um número limitado de pontos (três ou quatro), quase todos os pontos poderão ser adjacentes, resultando em uma porcentagem inflacionada de confiabilidade. Algumas vezes, é indicada a utilização de categorias adjacentes, nas quais o acordo combinado para a concordância entre os avaliadores nos extremos da escala é menor do que no meio da escala.

Outro método para calcular as estimativas de consenso de confiabilidade interavaliador é o método estatístico kappa de Cohen (COHEN, 1968; STEMLER, 2001, 2004). Esse método foi desenvolvido para estimar o grau de concordância entre dois avaliadores depois de corrigir a porcentagem, isto é, a concordância entre os avaliadores que seria esperada ao acaso com base nos valores das distribuições marginais. A interpretação desse método é a seguinte: um valor zero em kappa não indica que os dois avaliadores discordaram completamente, mas indica que os dois avaliadores concordam entre

si com a mesma frequência que seria esperada ao acaso. Do mesmo modo, valores de kappa positivos indicam que os avaliadores concordam entre si com maior frequência do que o esperado ao acaso, e os negativos, com menor frequência.

Valores de kappa estimados entre 0,41 e 0,60 são razoáveis, enquanto maiores de 0,60 são muito bons. O método estatístico kappa de Cohen é muito utilizado quando a maioria das observações cai em uma única categoria inflacionando as estimativas. Uma desvantagem é que consiste em um método de difícil interpretação. Os valores de kappa podem ser diferentes, dependendo da proporção de respondentes que pertencem a cada categoria da escala de avaliação. Desse modo, os valores de kappa provenientes de itens ou de estudos diferentes podem não ser comparáveis. Embora esse método possa fornecer uma indicação sobre se a concordância entre os avaliadores é melhor do que o previsto ao acaso, é difícil interpretar os valores kappa em circunstâncias diferentes (STEMLER, 2004).

A segunda abordagem, as estimativas de consistência, baseia-se no pressuposto de que não é realmente necessário que dois avaliadores tenham o mesmo entendimento da escala e atribuam a mesma pontuação para uma tarefa específica, desde que cada avaliador seja consistente na classificação do desempenho avaliado de acordo com sua própria definição da escala. Por exemplo, um determinado avaliador pode apresentar consistentemente resultados dois pontos mais altos na escala de classificação do que um segundo avaliador ao julgar o mesmo grupo de indivíduos. Nesse caso, os dois avaliadores não concordam com a maneira de aplicar os critérios de pontuação, mas a diferença na forma como eles aplicam os critérios de pontuação é constante e previsível (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

Uma das estatísticas mais populares para o cálculo do grau das estimativas de consistência é o coeficiente de correlação de Pearson. Esse coeficiente pode ser facilmente calculado pela maioria dos programas estatísticos existentes e, além disso, as pontuações na escala de avaliação podem ser de natureza contínua, podendo assumir valores decimais. Os coeficientes de correlação de Pearson podem ser calculados para cada par de avaliadores e para cada item de cada vez. Uma limitação do coeficiente de correlação de Pearson é que esse método assume que os dados são distribuídos normalmente (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

Outra estatística, também popular para o cálculo das estimativas de consistência de confiabilidade, é o coeficiente de Spearman. Esse coeficiente proporciona estimativas aproximadas do coeficiente de correlação de Pearson,

mas pode ser utilizado quando os dados estudados não estão normalmente distribuídos. A principal desvantagem para o coeficiente de Spearman é que ele requer ambos os juízes para avaliar todos os casos (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

Quando são utilizados vários avaliadores, uma opção para calcular as estimativas de consistência de confiabilidade é o coeficiente alfa de Cronbach. Essa é uma medida de consistência interna, útil para compreender a forma como os julgamentos de um grupo de avaliadores variam. A principal vantagem da utilização do alfa de Cronbach é que esse coeficiente produz uma única estimativa de consistência de confiabilidade entre vários avaliadores. A principal desvantagem desse método é que cada avaliador deve julgar cada caso, ou, então, o alfa só será calculado sobre a base de um subconjunto dos dados. Em outras palavras, se apenas um avaliador não pontuou um indivíduo em particular, esse indivíduo ficará de fora da análise (STEMLER, 2004).

A terceira abordagem, as 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 das notas dos avaliadores ao tentar criar uma pontuação resumida para cada indivíduo avaliado. Cada avaliador fornece informações que são úteis para a geração de uma pontuação para o indivíduo. Desse modo, não é necessário que dois avaliadores concordem perfeitamente ao aplicar os critérios de pontuação, pois as diferenças entre os avaliadores podem ser estimadas e compensadas na nota final do participante (STEMLER, 2004; BROWN; GLASSWELL; HARLAND, 2004).

As estimativas de medição são úteis quando diferentes níveis da escala representam os diferentes níveis de desempenho de um construto unidimensional ou quando na avaliação estão envolvidos vários avaliadores, mas nem todos julgam todos os itens (STEMLER, 2004).

Um método bastante utilizado para calcular as estimativas de medição é a teoria da generalização, também denominada *"Teoria G"*. Esse método oferece um extenso quadro conceitual e um poderoso conjunto de procedimentos estatísticos para tratar vários problemas de medição (BRENNAN, 2011). A Teoria G fornece uma maneira de distribuir a variância total de um conjunto de avaliações em partes separadas, não correlacionadas, que estão associadas com cada uma das diferentes fontes de variabilidade, por exemplo, a variabilidade sistemática entre os textos escritos pelos examinandos, a variabilidade entre os avaliadores e a variabilidade entre os itens. Além dos componentes de variância para cada um desses efeitos, essa teoria também permite ao pesquisador a obtenção de estimativas de componentes de variância

causados por interações entre esses efeitos. Ao comparar o tamanho relativo dos componentes de variância estimados, pode-se determinar quais fontes de variância são mais problemáticas (SUDWEEKS; REEVE; BRADSHAW, 2005; BOCK; BRENNAN; MURAKI, 2002).

A Teoria G faz distinção entre dois tipos de situações: (1) quando se deseja determinar a ordem de classificação dos participantes dentro de um grupo, nesse caso, a posição de cada indivíduo é determinada não apenas por seu próprio desempenho, mas também pelo desempenho dos outros membros do grupo de comparação; (2) quando se deseja decidir se o desempenho de um indivíduo é igual ou superior a um nível preestabelecido como padrão, nesse caso, a decisão sobre cada examinando é tomada independentemente do desempenho dos outros indivíduos do grupo. A definição de erro de medição é baseada na distinção entre essas duas situações (SUDWEEKS; REEVE; BRADSHAW, 2005).

A Teoria G fornece uma série de estatísticas de resumo. Uma relevante para esse contexto é o coeficiente *phi*. Esse coeficiente pode estabelecer o grau em que é atribuída a pontuação "verdadeira", levando-se em conta o efeito de todas as fontes de variabilidade (SUDWEEKS; REEVE; BRADSHAW, 2005; BROWN; GLASSWELL; HARLAND, 2004).

Outra abordagem para calcular a confiabilidade interavaliador é por meio do modelo multifacetas de Rasch, que permite determinar empiricamente a equivalência entre as pontuações atribuídas pelos avaliadores. Por exemplo, uma pontuação 3 proveniente do avaliador A é equivalente a uma pontuação 5 proveniente do avaliador B. Nesse caso, o avaliador A é mais severo do que o avaliador B. Desse modo, além de fornecer informações sobre a severidade de cada avaliador, quando comparado com o grupo de avaliadores, o modelo multifacetas de Rasch permite também avaliar o grau em que cada avaliador utiliza os critérios de pontuação de maneira consistente. Em outras palavras, mesmo que os avaliadores utilizem a escala com suas próprias interpretações, as estatísticas do modelo podem indicar o grau em que um determinado avaliador é fiel à sua própria definição das categorias de escala para todos os itens e todos os indivíduos (confiabilidade intra-avaliador) (STEMLER, 2004).

Stemler (2004) destaca algumas vantagens em estimar a confiabilidade interavaliador por meio da abordagem de medição. Uma delas é que as estimativas de medição podem considerar os erros ocasionados por cada avaliador individualmente e também pelo grupo de avaliadores. Desse modo, as pontuações finais tendem a representar com maior precisão o desempenho sobre o construto de interesse. Outra vantagem é que as estimativas de medi-

ção podem considerar simultaneamente as pontuações provenientes de todos os avaliadores para todos os itens que foram pontuados e não apenas o cálculo de estimativas de cada item e cada dupla de avaliadores.

Um resumo desses métodos descritos nesta seção para as estimativas de confiabilidade da pontuação, destacando algumas de suas vantagens e desvantagens, é feito nos Quadros 9, 10 e 11.

Quadro 9 – Métodos para as estimativas de consenso

Porcentagens exatas	
Vantagens	Facilidade de cálculos e de interpretação. Forte apelo intuitivo.
Desvantagens	Treinamento dos avaliadores trabalhoso e demorado. Não valoriza a diversidade das respostas.
Porcentagens adjacentes	
Vantagens	Facilidade de cálculos e de interpretação. Forte apelo intuitivo. O acordo entre os avaliadores não precisa ser exato.
Desvantagens	As estimativas de confiabilidade podem ser exageradas. Não valoriza a diversidade das respostas.
Kappa de Cohen	
Vantagens	Fornece uma boa indicação sobre a concordância entre os avaliadores. É bom quando a maioria das observações cai em uma única alternativa inflacionando as estimativas.
Desvantagens	Difícil interpretação. Não é possível comparar resultados de testes em circunstâncias diferentes.

Fonte: Autora

Quadro 10 – Métodos para as estimativas de consistência

Coeficiente de correlação de Pearson	
Vantagens	Facilidade de cálculos. A pontuação pode ser com números decimais. Aplicados a cada par de avaliadores e a cada item. Fácil interpretação.
Desvantagens	Os dados devem ser distribuídos normalmente.
Coeficiente de Spearman	
Vantagens	Estimativas aproximadas às da correlação de Pearson. Não é necessário distribuição normal dos dados. Fácil interpretação
Desvantagens	Requer que ambos os juízes avaliem todos os casos.
Coeficiente alfa de Cronbach	
Vantagens	Produz uma única estimativa de consistência de confiabilidade entre vários avaliadores. Fácil interpretação.
Desvantagens	Cada avaliador deve julgar todos os casos. Se algum avaliador deixar de pontuar um indivíduo, este ficará de fora da análise.

Fonte: Autora

Quadro 11 – Métodos para as estimativas de medição

Teoria da generalização – Teoria G	
Vantagens	As estimativas consideram as pontuações de todos os avaliadores para todos os itens simultaneamente. Extenso quadro conceitual e procedimentos estatísticos. Possibilita analisar fontes de variabilidade causadas por: tarefas, itens, interavaliadores, intra-avaliadores e interações entre fontes.
Desvantagens	Exige conhecimento especializado para as interpretações. Poucas opções de *software* para os cálculos.
Multifacetas de Rasch	
Vantagens	As estimativas consideram as pontuações de todos os avaliadores para todos os itens simultaneamente. Mesmas vantagens e propriedades matemáticas do modelo de Rasch. Permite a inclusão de fontes causadoras de erros nas avaliações. Determinação da equivalência entre as pontuações atribuídas pelos avaliadores. Fornece informações sobre a severidade de cada avaliador em comparação com o grupo de avaliadores. Permite avaliar o grau com que cada avaliador utiliza os critérios de pontuação de maneira consistente (confiabilidade intra-avaliador). Permite a comparação entre as dificuldades de cada item.
Desvantagens	Exige conhecimento especializado para as interpretações. Poucas opções de *software* para os cálculos.

Fonte: Autora

2.5.3.3 Conclusão sobre o padrão de qualidade da avaliação

Para completar as análises da validade da avaliação, Jaeger *et al.* (1996) propõem algumas questões úteis para que a avaliação satisfaça padrões de medição profissionais:

1. Os resultados da avaliação são suficientemente confiáveis para apoiar a seleção e classificação dos indivíduos, seja ela local, estadual ou nacional?
2. Os resultados da avaliação são válidos para apoiar as inferências sobre o desempenho quanto às aptidões e capacidades dos examinandos?
3. Os resultados da avaliação refletem as habilidades dos examinandos de forma justa e imparcial sem distinção quanto ao sexo, à raça, ao grupo étnico ou grupo socioeconômico?
4. Os resultados da avaliação refletem verdadeiramente as normas institucionais quanto à classificação dos examinandos por categorias, por exemplo, licenciado, graduado, certificado ou então por nível de proficiência como "básico", "intermediário", "avançado", entre outros?

O modelo multifacetas de Rasch e os índices estatísticos citados anteriormente, que fazem parte do contexto desse modelo, possibilitam a resposta a essas e a outras questões que auxiliam na validação da avaliação.

A primeira questão é respondida com análises dos índices de ajuste dos dados aos modelos de Rasch. A segunda questão diz respeito ao desempenho do examinando, que, por sua vez, está relacionado com a qualidade das tarefas propostas, com a precisão da pontuação atribuída pelos avaliadores e com a adequação dos critérios e escalas utilizadas para a pontuação. Todas essas variáveis podem ser incluídas no modelo multifacetas de Rasch que proporciona análises de cada uma delas, tanto no nível de grupo como no nível individual.

Pode-se também, por meio da utilização do modelo multifacetas de Rasch, examinar o desempenho diferencial entre os diferentes grupos, como raça, classe social, idade, entre outros, considerando separadamente as variáveis dentro dos grupos e analisando a existência de alguma influência entre as dificuldades relativas dos itens, as pontuações provenientes dos avaliadores e os elementos observados nos grupos. Os dados provenientes desse estudo fornecem a resposta à terceira questão.

Quanto à última questão, os modelos de Rasch são conhecidos por possibilitarem comparações entre diferentes edições de um sistema de avaliação, desde que sejam respeitadas certas condições. Quando isso é possível, podem-se comparar os padrões dos vários exames aplicados e estabelecer as normas de certificação para cada categoria em particular. Por pertencer à família de modelos de Rasch, o modelo multifacetas também pode ser utilizado com essa finalidade.

O modelo multifacetas de Rasch mostra-se adequado para assegurar a validade das avaliações com itens de respostas construídas. Na aplicação prática deste trabalho (Capítulo 5), esse modelo é utilizado para as análises de dados provenientes de uma avaliação da linguagem, sendo possível, desse modo, constatar na prática a sua eficiência.

Neste capítulo, foram estabelecidos uma variedade de métodos e procedimentos que devem ser adotados para assegurar a qualidade de avaliações com itens de respostas construídas em larga escala, desse modo, orientações especializadas de diversas áreas são essenciais em todas as etapas da elaboração dessas avaliações.

As técnicas abordadas neste capítulo, referem-se, na maior parte, a cada etapa da avaliação isoladamente, não integrando os diversos procedimentos e processos demandados. Há, portanto, a necessidade da elaboração de modelos práticos que possam ser aplicados como um todo e que englobem todo o processo. Neste sentido, uma sistemática que sirva como guia para as empresas provedoras de avaliações em larga escala, é bem-vista para auxiliar as pessoas em todas as etapas demandadas para a construção das avaliações

com itens de respostas construídas, de modo que essas avaliações possam alcançar padrões profissionais de qualidade.

Desse modo, para finalizar este capítulo, a seção seguinte apresenta uma sistemática contendo todas as etapas necessárias para a concepção, elaboração e implantação de avaliações em larga escala com itens de respostas construídas.

2.6 SISTEMÁTICA PARA ELABORAÇÃO DE AVALIAÇÕES COM ITENS ABERTOS

No Brasil, há uma carência de trabalhos que auxiliem as empresas provedoras de avaliações em larga escala na construção e análises das avaliações, principalmente quando se trata de avaliações com itens de respostas construídas. Nesse sentido, a elaboração de uma sistemática para a concepção e a construção de avaliações com itens de respostas construídas, explicitando claramente os conceitos envolvidos na elaboração dos instrumentos de avaliação, dos critérios de correção e de pontuação para as tarefas estabelecidas, assim como as análises estatísticas para a determinação da validade da avaliação e classificação dos candidatos, pode contribuir para a melhoria da qualidade das avaliações de modo geral.

Desse modo, para encerrar este capítulo, apresenta-se a sistematização das etapas sugeridas para a elaboração de avaliações com itens de respostas construídas. Estas etapas devem ser desenvolvidas de acordo com as teorias e procedimentos discutidos neste capítulo.

A sistematização das etapas sugeridas para a elaboração de avaliações com itens abertos é apresentada no Quadro 12.

Quadro 12 – Sistemática para a elaboração de avaliações com itens abertos

	ETAPA	Descrição	Procedimento/finalidade
PROCEDIMENTOS TEÓRICOS	ETAPA 1	Definição do teste	Abrangência, objetivos, recursos, etc.
	ETAPA 2	Delimitação do domínio do construto	Dimensionalidade
			Definições constitutivas
			Definições operacionais
	ETAPA 3	Operacionalização do construto	Elaboração dos itens
			Elaboração dos critérios
	ETAPA 4	Análise Teórica	Validação dos itens
			Validação dos critérios
PROCEDIMENTOS EMPÍRICOS	ETAPA 5	Planejamento e aplicação do teste	Diagramação
			Impressão
			Armazenamento
	ETAPA 6	Treinamento dos avaliadores	Estudo dos critérios
			Confiabilidade
	ETAPA 7	Pontuação do teste	Sessões de pontuação
			Monitoramento da qualidade
PROCEDIMENTOS ANALÍTICOS	ETAPA 8	Validade	Avaliação da qualidade: tarefas, itens, escalas e classificação
	ETAPA 9	Confiabilidade entre avaliadores	Avaliação da precisão da pontuação
	ETAPA 10	Resultados da avaliação	Divulgação: aos examinandos, à sociedade e às estâncias superiores

2.6.1 Etapa 1: Definição do teste

A primeira etapa do evento da avaliação deve ter início com as especificações do teste que devem ser estabelecidas por uma equipe multidisciplinar formada por administradores, especialistas em avaliação, especialistas nas áreas dos construtos que serão avaliados e estatísticos.

Entre essas especificações estão a identificação dos objetivos e finalidades da avaliação, que podem ser formativas, de diagnóstico, classificatórias, etc., dos construtos que devem ser testados e quais aspectos são mais importantes e devem possuir maior destaque. Também é importante a definição de alguns aspectos práticos da avaliação, a maior parte de caráter administrativo, mas algumas decisões também devem ser tomadas por equipes multidisciplinares. Entre eles estão o domínio da avaliação, que consiste na definição da abrangência da avaliação, se esta será aplicada institucionalmente, localmente, regionalmente, nacionalmente ou em outras configurações; o número estimado de participantes; a previsão orçamentária para o evento ou outras variáveis que podem ser importantes para cada avaliação específica. Esses aspectos práticos influenciarão a tomada de decisão de alguns fatores em outras etapas da avaliação, como, por exemplo, o estilo das tarefas, o número de itens, o tempo destinado à elaboração das respostas e o número de avaliadores que serão necessários para a pontuação da avaliação, além da determinação das análises estatísticas que serão necessárias e a maneira como os resultados serão apresentados. Essas especificações fornecem um "modelo" para a elaboração do teste, pois sem ele o desenvolvimento do teste corre o risco de prosseguir sem uma direção clara.

A elaboração desse conjunto de especificações deve ser o primeiro passo no processo de desenvolvimento do teste. Elas devem ser elaboradas para cada avaliação em particular e necessitam ser continuamente revistas para acompanhar as modificações necessárias e as tendências atuais para a medição do conhecimento.

2.6.2 Etapa 2: Delimitação do domínio do construto

Essa etapa consiste na definição do construto, sua delimitação e sua dimensionalidade e depende de profissionais especialistas da área do construto em questão. A teoria sobre a delimitação do construto encontra-se descrita na Seção 2.5.1.1.

2.6.2.1 Dimensionalidade

Primeiramente, deve-se determinar a dimensionalidade do atributo conforme foi apresentado na Seção 2.5.1. Às vezes é possível elaborar o teste de modo que o atributo apresente apenas uma dimensão, outras vezes,

o atributo é de natureza multidimensional, não sendo possível, ou algumas vezes vantajoso, separar as componentes de modo a torná-lo unidimensional.

Para um teste ser unidimensional, ele deve medir a habilidade em um único atributo, como, por exemplo, em matemática, química ou biologia. Se o teste avalia o conhecimento em mais do que um atributo, por exemplo, química e matemática ou álgebra e trigonometria, então ele é multidimensional.

2.6.2.2 Definições constitutivas e operacionais

Para a definição constitutiva, os termos são definidos com outras palavras, isto é, os conceitos são definidos em termos de outros conceitos. Essas definições consistem em conceitos abstratos. A definição operacional viabiliza um significado concreto para os conceitos, que devem especificar as atividades ou operações necessárias para a obtenção de uma medida. A delimitação do domínio do construto deverá guiar tanto a seleção da tarefa e a elaboração do item quanto o desenvolvimento racional de critérios de pontuação, assuntos que serão abordados na próxima etapa do desenvolvimento da avaliação. Este assunto encontra-se descrito na Seção 2.5.1.1.

2.6.3 Etapa 3: Operacionalização do construto

A operacionalização do construto engloba a construção dos itens e dos critérios para a pontuação das tarefas, sendo necessária uma equipe multidisciplinar composta por profissionais especialistas na área de avaliação educacional, na área do construto avaliado e estatísticos para subsidiar a construção adequada dos itens e dos critérios para a pontuação.

A etapa para a elaboração de avaliações com itens abertos é constituída de duas partes: (1) a elaboração da tarefa que normalmente é exposta em forma de itens e (2) a elaboração do conjunto de critérios de pontuação. Essa etapa é de suma importância e uma das mais trabalhosas e delicadas na elaboração de um teste, especialmente os testes em larga escala, pois, se algum erro ou equívoco for detectado tardiamente, dificilmente poderá ser corrigido sem consequências prejudiciais ao evento e a seus participantes. Os procedimentos para a elaboração dos itens e dos critérios para a pontuação estão descritos na Seção 2.5.1.2.

2.6.4 Etapa 4: Análise teórica

A análise teórica é feita por meio de julgamentos de especialistas para verificar a adequação dos itens e dos critérios de pontuação aos objetivos do teste. Basicamente consistem em dois tipos de julgamentos: (1) quanto à validade de conteúdo e (2) quanto à validade aparente. Conforme foram descritas na Seção 2.5.1.3, a validade de conteúdo é determinada por especialistas da área do construto que devem analisar se os itens são adequados para avaliar o construto em questão e se os critérios de pontuação de fato capturam os traços previstos para serem avaliados. A validade aparente, ou análise semântica, tem a finalidade de determinar se os itens são compreensíveis para todos os indivíduos da população e não precisa, necessariamente, ser feita por especialistas da área do construto. Podem ser de outras áreas, ou mesmo pertencentes à população para a qual o teste foi desenvolvido. Nessas análises, devem constar, também, as referentes à justiça, que devem ter o foco nas fontes de variância construto-irrelevante.

2.6.5 Etapa 5: Planejamento e aplicação do teste

O planejamento da aplicação e a aplicação do teste dependem da especificidade de cada avaliação e deverão considerar algumas variáveis, como a segurança do teste e a garantia de oportunidades iguais para todos os participantes.

Como exposto na Seção 2.5.2, as empresas elaboradoras de avaliações em larga escala necessitam do apoio de três setores especializados: o computacional, o logístico e o pedagógico. Nessa fase de planejamento e aplicação do teste, o trabalho é desenvolvido pelos setores logístico e pedagógico. O logístico se preocupa com todos os procedimentos de ordem prática para a realização do evento, como os locais para a realização das provas, a alocação dos examinandos nesses locais, a seleção e o treinamento do pessoal de apoio, o transporte seguro dos materiais necessários para os locais de prova, as condições de acesso dos examinandos aos locais de prova e outras inúmeras preocupações. Esses procedimentos não serão detalhados neste trabalho.

Os procedimentos pedagógicos envolvem a diagramação dos testes, a impressão, a organização e o armazenamento dos cadernos de provas. Esses procedimentos encontram-se detalhados nas Seções 2.5.2.1 e 2.5.2.2.

2.6.6 Etapa 6: Treinamento dos avaliadores

Essa etapa visa, principalmente, à confiabilidade da pontuação ou à consistência dos escores dos participantes. É esperado que as pontuações atribuídas por dois avaliadores a uma mesma resposta de teste variem muito pouco. Além disso, é esperado também que o examinando receba a mesma pontuação, quando responder ao teste em ocasiões diferentes. A variabilidade da pontuação não pode ocorrer por razões externas ao teste. A validade também depende da confiabilidade da pontuação, isto é, um processo avaliativo não pode ter seus resultados considerados válidos sem um nível suficiente de acordo entre avaliadores.

A confiabilidade entre avaliadores independentes (interavaliador), sem discussão ou colaboração entre si, é considerada a característica mais importante da pontuação das avaliações com itens abertos, embora para algumas abordagens, como a do modelo multifacetas de Rasch, seja exigida apenas a confiabilidade intra-avaliador, isto é, que cada avaliador seja consistente com o seu próprio modo de atribuir as pontuações em relação aos desempenhos dos traços associados à escala utilizada. Independentemente do tipo de confiabilidade exigida, interavaliador, intra-avaliador ou ambas, o treinamento dos avaliadores é sempre um procedimento primordial para a avaliação. Os métodos mais utilizados atualmente para o treinamento dos avaliadores são descritos na Seção 2.5.2.3.

2.6.7 Etapa 7: Pontuação dos testes

As pontuações geralmente são atribuídas por dois ou três avaliadores diferentes. Durante as sessões de pontuação, deve-se monitorar constantemente a diferença entre as pontuações atribuídas pelos avaliadores para a constatação de discrepâncias por meio de métodos, como o de porcentagens exatas ou porcentagens adjacentes. Esses métodos são comumente utilizados nessa etapa do processo por sua simplicidade, facilidade de análises e rapidez de cálculos (ver Seção 2.5.3.2). Entretanto, para estudos mais minuciosos objetivando a busca por efeitos causados pelas tendências dos avaliadores a pontuações sistemáticas, como o efeito de tendência central, severidade/complacência, halo e viés, pode-se empregar o modelo multifacetas de Rasch para escala de crédito parcial e monitorar os avaliadores individualmente quase que simultaneamente às sessões de pontuação. O método multifacetas de Rasch encontra-se descrito no Capítulo 3, sendo a ferramenta

utilizada nas análises estatísticas da aplicação prática nesta tese.

2.6.8 Etapa 8: Validade

Essa sistemática sugere que a preocupação com a validade seja constante e presente em todas as etapas para a elaboração da avaliação, destacando-se, principalmente, os conceitos expostos nas Seções 2.3.1, 2.3.5, 2.5.1 e 2.5.1.2, e considera que a validade da avaliação é garantida, em grande parte, por procedimentos adotados na fase da construção do teste.

Quando o instrumento é desenvolvido respeitando-se os procedimentos que visam à validade, então não serão necessárias investigações substanciais para assegurar a validade da avaliação (BORSBOOM; MELLENBERG; VAN HEERDEN, 2004).

Entretanto, alguns problemas são detectados apenas após a conclusão da avaliação. A determinação desses problemas é importante para que possam ser válidas, ou não, as inferências feitas sobre os resultados da avaliação e para permitir que os erros sejam corrigidos para a elaboração da próxima edição da avaliação, se for o caso.

2.6.9 Etapa 9: Confiabilidade

Essa etapa é importante para garantir que as inferências a serem feitas sobre os resultados da avaliação serão válidas uma vez que não é possível ter validade sem confiabilidade da pontuação.

Na literatura são sugeridas abordagens variadas para a verificação da confiabilidade da pontuação atribuída por avaliadores. As principais abordagens são descritas na Seção 2.5.3.2.

Um quadro geral para conceituar as avaliações que necessitam da mediação de avaliadores, sob a lente do modelo multifacetas de Rasch encontra-se descrito no Capítulo 3, juntamente com uma família de índices que devem ser utilizados para examinar a qualidade das pontuações atribuídas às respostas dos examinandos sob o ponto de vista de erros sistemáticos gerados por tendências dos avaliadores. Esse modelo é utilizado na aplicação prática desse trabalho para analisar a qualidade da pontuação atribuída aos participantes da avaliação.

2.6.10 Etapa 10: Divulgação dos resultados da avaliação

Nessa etapa são divulgados os resultados da avaliação. A primeira preocupação deve ser no sentido de fornecer os resultados aos participantes da avaliação de acordo com a finalidade do teste. Por exemplo, se a finalidade da avaliação for educacional, o desempenho que o participante obteve na avaliação deve ser o mais detalhado possível, pois o retorno fornecido ao participante (*feedback*) é considerado pelos educadores uma fonte instrutiva importante que pode colaborar com a aprendizagem. Se a finalidade da avaliação for classificatória, deve-se ponderar sobre o fornecimento ou não dos testes corrigidos ou detalhamento sobre a pontuação recebida pelos participantes. Isso depende também da estrutura da empresa provedora da avaliação, pois, em avaliações em larga escala, o número de participantes pode ser muito grande e o fornecimento de cópias das provas corrigidas a todos eles, por exemplo, pode não ser possível. Mesmo assim, a empresa responsável pela avaliação deve fornecer informações aos participantes, as mais detalhadas possíveis, sobre os critérios utilizados na pontuação, o que foi considerado ou não na correção, os erros mais comuns, entre outros. São necessárias também respostas aos questionamentos dos participantes quanto às suas pontuações e a pronta correção de erros, quando eles forem constatados.

Os resultados das avaliações em larga escala também devem ser divulgados à sociedade, de acordo com as suas finalidades. Além disso, estudos sobre os resultados da avaliação estão se tornando cada vez mais primordiais e devem ser elaborados a cada edição da avaliação uma vez que a melhoria dos sistemas avaliativos em larga escala é conseguida também pela correção dos erros detectados de uma edição da avaliação para a outra. Elaboração e divulgação de estudos sobre as avaliações é uma tendência atual que tem sido notada em sistemas de avaliações importantes dos Estados Unidos e países da Europa. Apesar de sempre serem detectados erros nesses processos, a divulgação dos resultados colabora grandemente com a credibilidade e assegura a eficiência da avaliação.

3 MODELO MULTIFACETAS DE RASCH

Os modelos de medição denominados *multifacetas de Rasch* são derivados da família de modelos que tiveram suas raízes no modelo de Rasch para itens dicotômicos (RASCH, 1960). Esses modelos da família de modelos de Rasch se diferenciam pelo tipo de observações modeladas e pela forma específica com que eles transformam essas observações em medidas lineares (ECKES, 2011).

Com a finalidade de proporcionar uma visão ampla das propriedades e características dos modelos de medição multifacetas de Rasch, assim como destacar suas semelhanças e diferenças com o modelo básico de Rasch, primeiramente apresenta-se neste capítulo o modelo de Rasch para itens dicotômicos, na sequência, são apresentadas algumas entre as principais extensões desse modelo que são apropriadas para itens politômicos e das quais os modelos multifacetas são derivados. Finalmente, são apresentados os modelos de medição multifacetas de Rasch.

3.1 MODELO DE RASCH PARA ITENS DICOTÔMICOS

O modelo de Rasch, desenvolvido pelo matemático dinamarquês Georg Rasch por volta de 1960, é adequado para itens dicotômicos, isto é, considera apenas duas categorias de respostas, correta ou incorreta. O modelo de Rasch é também conhecido por modelo logístico da TRI de 1 parâmetro (ML1).

O modelo de Rasch estabelece que a probabilidade de um indivíduo j optar pela alternativa correta no item i é dada por

$$P(x_{ij} = 1|\theta_j) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \tag{1}$$

$i = 1,2,\cdots,I$ e $j = 1,2,\cdots,n$, com I o número de itens e n o número de indivíduos da população. x_{ij} é uma variável dicotômica que assume os valores 1, quando o indivíduo j responde corretamente ao item i, ou 0 quando o indivíduo j não responde corretamente ao item i; θ_j representa a habilidade (traço latente) do j-ésimo indivíduo; b_i é o parâmetro de dificuldade do item i, medido na mesma escala da habilidade (ANDRADE; TAVARES; VALLE, 2000).

A equação (1) estabelece o modelo de Rasch em sua forma exponen-

cial. Nesse modelo, a probabilidade do examinando j responder corretamente ao item i, $P(x_{ij} = 1)$ depende da diferença entre a habilidade do examinando (θ_j) e a dificuldade do item (b_i). Se a habilidade do examinando for igual à dificuldade do item, $\theta_j - b_i = 0$, a probabilidade desse indivíduo de responder ao item corretamente é de $P(x_{ij}) = 1/2$. Quanto maior for a habilidade do examinando em relação à dificuldade do item, maior será a probabilidade dele de responder ao item corretamente (ECKES, 2011; ANDRADE; TAVARES; VALLE, 2000). O parâmetro de dificuldade do item (b_i) é também denominado de parâmetro de locação do item.

A relação existente entre a probabilidade de um indivíduo responder corretamente a um item e os parâmetros desse item é uma função monótona e crescente denominada de Curva Característica do Item (CCI). Para exemplificar, destaca-se, na Figura 5, a representação gráfica da curva característica de um item hipotético.

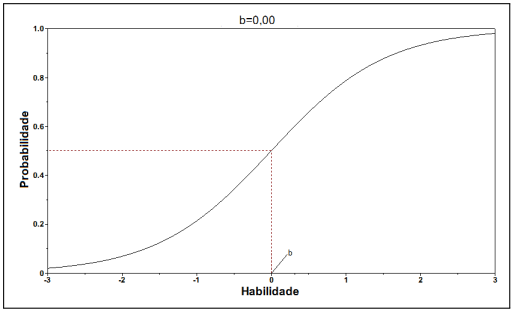


Figura 5 – Curva Característica do Item

Fonte: Autora

A escala de habilidade (eixo horizontal), teoricamente, pode assumir todos os valores reais, entretanto o que é importante nessa escala não é a sua extensão, mas as relações de ordem existentes (ANDRADE; TAVARES; VALLE, 2000). Por este motivo, não existe um único local correto para a origem do sistema de medição e também não existe uma unidade de medida predefinida. Estes podem ser escolhidos de acordo com a especificidade de cada aplicação.

O parâmetro b é uma medida da dificuldade do item e é dado na mesma unidade da habilidade. Observa-se na Figura 5 que quanto maior a habilidade do indivíduo, maior é a probabilidade dele de responder corretamente ao item.

Traçando-se uma linha vertical em uma habilidade, na intersecção dessa linha com a CCI, obtém-se a probabilidade de um indivíduo com aquela habilidade responder corretamente ao item. Observe que quanto mais para a direita está a CCI, mais difícil é o item. A dificuldade do item é definida como o nível de habilidade na qual a probabilidade de uma resposta correta é de 0,5. Na Figura 5, a dificuldade do item correspondente é b=0,00.

A dificuldade do item e a habilidade dos examinandos consistem nos parâmetros do modelo. Normalmente a estimação dos parâmetros é denominada de calibração e, para isso, são utilizados os dados provenientes das respostas aos itens. O modelo de Rasch básico especifica um único parâmetro para o item, a dificuldade do item. É por essa razão que este modelo é também denominado de modelo da TRI de 1 parâmetro.

Quanto a estimação dos parâmetros, um problema comum é a falta de identificabilidade do modelo. Este problema ocorre quando mais de um conjunto de parâmetros produz o mesmo valor da probabilidade. Os parâmetros da habilidade dos examinandos devem ser estimados na mesma métrica dos parâmetros da dificuldade dos itens e vice-versa, entretanto, quando é feita a estimação das habilidades e da dificuldade dos itens em conjunto, não há uma métrica definida, o que gera a não-identificabilidade. Alguns métodos são comumente utilizados para eliminar a não-identificabilidade. Uma forma consiste em definir uma métrica (unidade de medida) especificando uma medida de posição, como a média, e outra de dispersão, como o desvio padrão, para as habilidades, e consequentemente, para a dificuldade dos itens. Neste caso, uma métrica popular é a (0,1), ou seja, faz-se a média da habilidade dos examinandos igual a zero e o desvio padrão igual a 1.

Outro método comum para o modelo de Rasch, principalmente quando se faz a estimação dos parâmetros em conjunto, consiste em impor alguma restrição para as habilidades ou então, para a dificuldade dos itens. Na aplicação prática deste trabalho, faz-se a média da didiculdade dos itens igual a zero.

Deste modo, neste trabalho, o modelo de Rasch é definido pela equação (1) juntamente com a restrição

$$\sum_{i=1}^{I} b_i = 0. \tag{2}$$

Mais informações sobre a estimação dos parâmetros e a identificabilidade do modelo podem ser obtidas na Seção 3.4.

Uma expressão alternativa para o modelo dicotômico de Rasch

é em termos de *log-odds* ou *logitos*. A probabilidade do examinando j de responder corretamente ao item i é dada por $P(x_{ij} = 1)$ e a probabilidade do examinando j de responder incorretamente ao item i é $P(x_{ij} = 0)$. A razão entre essas duas probabilidades resulta em:

$$\frac{P(x_{ij} = 1)}{P(x_{ij} = 0)} = \exp(\theta_j - b_i). \tag{3}$$

Aplicando o logaritmo natural em ambos os membros dessa equação, obtem-se a equação em *logitos*, que é uma abreviação do logaritmo (*log*) da razão entre probabilidades (*odds*) ou *log-odds*. Desse modo, o modelo de Rasch na forma *logitos* é dado pela equação:

$$\ln\left[\frac{P(x_{ij} = 1)}{P(x_{ij} = 0)}\right] = \theta_j - b_i. \tag{4}$$

Essa é uma função linear dos parâmetros da habilidade do indivíduo (θ_j) e da dificuldade do item (b_i). A habilidade do examinando e a dificuldade do item são dadas na mesma escala *logito* que pode assumir valores no intervalo $(-\infty, \infty)$, entretanto, na prática, é normal encontrar valores no intervalo $(-5, 5)$. Um *logito* é a distância ao longo da escala da variável latente que aumenta a probabilidade de observar o evento especificado no modelo por um fator de aproximadamente 2,7178, o valor de e. Quando a habilidade do examinando é igual à dificuldde do item, a medida para o sucesso da resposta é zero *logito* (LINACRE; WRIGHT, 1989).

O modelo de Rasch para itens dicotômicos e também suas extensões para itens politômicos possuem algumas vantagens sobre a Teoria Clássica dos Testes (TCT), ainda muito utilizada nas análises dos resultados das avaliações. A vantagem considerada a mais importante é a *invariância das medidas* ou, como foi denominada por Rasch, *objetividade das medidas* (RASCH, 1968 *apud* LINACRE; WRIGHT, 2002). A objetividade de medidas ocorre quando são obtidas as mesmas medidas para a habilidade dos examinandos, independentemente de qual amostra de itens foi utilizada no teste, isto é, as medidas da habilidade dos examinandos são independentes dos itens utilizados. As medidas dos itens também são invariantes quando aplicadas a grupos distintos de examinandos, isto é, as medidas dos itens são independentes dos examinandos que responderam aos itens ou da ocasião na qual o teste foi aplicado (ECKES, 2011; ANDRADE; TAVARES; VALLE, 2000).

Para tanto, o modelo de Rasch exige a unidimensionalidade do teste, isto é, o teste deve medir um único traço latente. Entretanto, a unidimensionalidade perfeita não existe uma vez que sempre são necessários conhecimentos

subjacentes para a resposta a uma tarefa. Desse modo, unidimensionalidade de um teste refere-se a saber qual é o grau de multidimensionalidade aceitável para que as interpretações dos resultados da medição não sejam ameaçadas (WRIGHT; LINACRE, 1989).

Outra exigência do modelo de Rasch é a independência local. Essa propriedade estabelece que respostas a itens distintos devem ser independentes, isto é, a resposta a um determinado item não deve depender de respostas a outros itens do teste, ou mesmo ser influenciada por elas.

Segundo Andrade, Tavares e Valle (2000), a unidimensionalidade implica em independência local, desse modo, se os itens forem elaborados para satisfazer a unidimensionalidade, as duas exigências para a utilização do modelo de Rasch serão atendidas e, desse modo, pode-se obter a adequação dos dados ao modelo, permitindo-se análises válidas para os resultados do teste.

Duas consequências importantes são derivadas da objetividade das medidas: (1) os escores do teste são estatísticas suficientes para a estimação da habilidade dos examinandos. Isto significa que o número de respostas corretas assinaladas pelo indivíduo contém toda a informação necessária para a estimativa da sua habilidade, e (2) o teste é unidimensional, isso significa que todos os itens do teste devem medir um único traço latente ou construto, ou melhor, o indivíduo deve necessitar de uma única habilidade dominante para responder a todos os itens do teste, essa habilidade é a que supostamente está sendo medida no teste (ECKES, 2011).

O acréscimo ao modelo de Rasch do parâmetro da discriminação resulta no modelo da TRI de dois parâmetros, (*modelo logístico de 2 parâmetros* – ML2). Nesse caso, os parâmetros consistem em dificuldade e discriminação do item. Para o modelo da TRI de três parâmetros (*modelo logístico de 3 parâmetros* – ML3), é necessário acrescentar também o parâmetro da resposta ao acaso, resultando no modelo com parâmetros de dificuldade, de discriminação e de resposta ao acaso. Esses modelos da TRI não pertencem à família de modelos de Rasch.

3.2 MODELOS DE RASCH PARA ITENS POLITÔMICOS

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, além de certo ou errado. Essas extensões fazem parte da família de modelos de Rasch.

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 proposto por Andrich em 1978 e o segundo é o modelo de crédito parcial desenvolvido por Masters em 1982. Esses modelos são definidos na sequência.

3.2.1 Modelo de Escala Gradual – MEG

O modelo de escala gradual foi proposto por Andrich em 1978 e é adequado para itens com categorias de respostas ordenadas igualmente espaçadas. Esse modelo, escrito na forma *logito* conforme apresentado por Eckes (2011), é dado por:

$$\ln\left[\frac{P_{ik}(\theta_j)}{P_{i(k-1)}(\theta_j)}\right] = \theta_j - b_i - d_k \tag{5}$$

e pelas restrições

$$\sum_{i=1}^{I} b_i = 0, \quad d_0 = 0 \quad \text{e} \quad \sum_{k=1}^{m} d_k = 0 \tag{6}$$

$P_{ik}(\theta_j)$ é a probabilidade do examinando j de responder com a categoria k ao item i e $P_{i(k-1)}(\theta_j)$ é a probabilidade do examinando j de responder com a categoria $k-1$ ao item i. O termo b_i é o parâmetro da dificuldade do item e d_k é o parâmetro da categoria.

No caso dos modelos de Rasch para itens politômicos, a não-identificabilidade do modelo é resolvida similarmente ao modelo dicotômico mas, neste caso, são necessárias também, restrições para a escala de classificação. Desse modo, neste trabalho, as restrições dadas em (6) asseguram a identificabilidade do modelo. Este assunto é tratado com mais detalhes na Seção 3.4.

Para um indivíduo de habilidade θ_j e um item i com $m+1$ categorias $k = 0, 1, \cdots, m$, o modelo especifica a probabilidade que é observada na categoria k de um item i de dificuldade b_i em relação à probabilidade observada na categoria $(k-1)$.

O parâmetro d_k, denominado parâmetro de categoria, é o ponto da

variável latente, em relação à dificuldade do item (b_i), no qual a probabilidade de ser observada na categoria k é igual à de ser observada na categoria $k-1$.

Para um melhor entendimento sobre o significado desses parâmetros de categoria no modelo de escala gradual, pode-se agrupar os parâmetros de locação do item (b_i) e de categoria (d_k) para definir o parâmetro de locação (dificuldade) de cada categoria, relativo ao item i, $b_{ik} = b_i + d_k$, assim, a equação (5) pode ser reescrita como:

$$\ln\left[\frac{P_{ik}(\theta_j)}{P_{i(k-1)}(\theta_j)}\right] = \theta_j - (b_i + d_k) = \theta_j - b_{ik}. \tag{7}$$

A locação (dificuldade) do item b_i é determinada pelo ponto no qual a probabilidade de ocorrer a categoria mais alta é igual à probabilidade de ocorrer a categoria mais baixa. As curvas "deslizam" sobre a escala para estabelecer a dificuldade do item (LINACRE, 2014b).

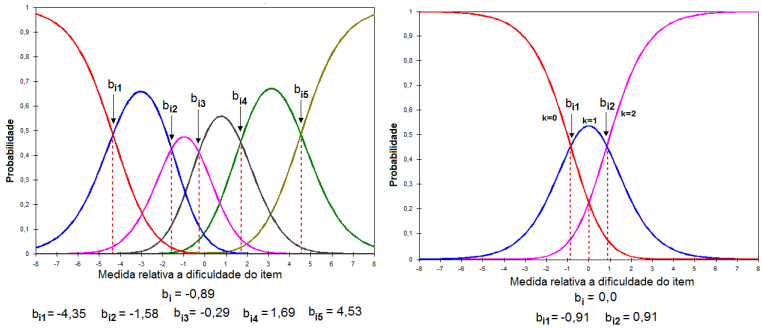
O parâmetro de categoria d_k representa o incremento ao parâmetro de locação do item b_i, no qual a probabilidade de um examinando de escolher uma categoria é igual à probabilidade dele de escolher uma outra categoria adjacente. Desse modo, os valores b_{ik} são pontos de transição, nos quais, o examinando de habilidade θ_j deixa de escolher uma categoria para escolher outra adjacente. Os parâmetros de locação (d_k) são também denominados *limiares* ou *locação das categorias* e em inglês de *Rasch-Andrich thresholds*. Aliás, este termo é utilizado também para os outros modelos de Rasch para itens politômicos, independentemente de ser ou não o modelo de Andrich.

Além disso, o modelo de escala gradual utiliza o mesmo conjunto de parâmetros das categorias (d_k) para todos os itens do teste. Desse modo, esse modelo é aplicado apenas quando a estrutura da escala de avaliação é a mesma para todos os itens. Isso significa que os itens devem ter o mesmo número de categorias e que a dificuldade relativa entre cada par de categorias é constante ao longo de todos os itens. Quando o teste é composto por itens que possuem suas próprias categorias de respostas, este modelo não é adequado (ECKES, 2011; ANDRADE; TAVARES; VALLE, 2000).

Para exemplificar, a parte (a) da Figura 6 exibe o gráfico da curva característica de um item i de dificuldade $b_i = -0,89$, no qual, cada curva corresponde a uma categoria $k = 0, \cdots, 5$. Também são indicados os parâmetros de locação que se situam no ponto de interseção entre duas categorias adjacentes, denotadas por $b_{i1}, \cdots, b_{i5}$.

A parte (b) da Figura 6 traz o gráfico da curva característica de um item com três categorias. Neste exemplo $b_i = 0$ o que resulta em $b_{ik} = d_k$

para $k = 1$ e $k = 2$.



(a) Curva Característica do Item (b) Curva Característica do Item

Figura 6 – Locação das categorias
Fonte: Autora

Como os parâmetros de categoria d_k são comuns a todos os itens do teste, se cada item tiver $(m+1)$ categorias de respostas, devem ser estimados m parâmetros de categoria e um parâmetro de locação para cada item ($d_0 = 0$). Assim, para um teste com I itens, cada um com $m+1$ categorias de resposta, o número de parâmetros a ser calculado é $I+m$ (ANDRADE; TAVARES; VALLE, 2000).

3.2.2 Modelo de Crédito Parcial – MCP

Para o modelo de crédito parcial (MCP) desenvolvido por Masters em 1982, a estrutura da escala de pontuação deve variar dependendo do item, isto é, deve ser diferente para itens diferentes. Em outras palavras, os itens possuem diferentes números de categorias de respostas ou a dificuldade relativa entre as categorias variam de item para item. O MCP estima os parâmetros de categoria para cada item separadamente, permitindo que a escala de avaliação seja específica para cada item.

O modelo de crédito parcial na forma *logito*, conforme apresentado por Eckes (2011), é dado pela equação:

$$\ln\left[\frac{P_{ik}(\theta_j)}{P_{i(k-1)}(\theta_j)}\right] = \theta_j - b_i - d_{ik} \tag{8}$$

e pelas restrições a seguir, sendo que as duas últimas são estabelecidas para cada item i

$$\sum_{i=1}^{I} b_i = 0, \quad d_{i0} = 0 \quad \text{e} \quad \sum_{k=1}^{m_i} d_{ik} = 0 \tag{9}$$

$P_{ik}(\theta_j)$ é a probabilidade do examinando j de responder com a categoria k ao item i e $P_{i(k-1)}(\theta_j)$ é a probabilidade do examinando j de responder com a categoria $k-1$ ao item i.

As restrições dadas em (9) resolvem o problema da não-identificabilidade do modelo. Sobre este assunto e o motivo para a necessidade de imposição destas restrições veja a Seção 3.4.

Para um indivíduo de habilidade θ_j e um item i com $m_i + 1$ categorias de respostas ($k = 0, 1, \cdots, m_i$), o modelo especifica a probabilidade que é observada na categoria k de um item i de dificuldade b_i em relação à probabilidade observada na categoria $(k-1)$.

Neste modelo, d_{ik}, com $k = 1, \cdots, m_i$, significa que cada item i possui a sua própria estrutura de categorias de respostas, numeradas de 0 até m_i. Isto significa que é permitido que cada item do teste possua diferentes números de categorias e, além disso, que a dificuldade de cada item b_i, seja acrescida de uma dificuldade adicional associada com cada categoria de respostas d_{ik}.

Desse modo, agrupando-se os parâmetros de locação do item (b_i) e de categoria (d_{ik}) tem-se o parâmetro de dificuldade $b_{ik} = b_i + d_{ik}$. Este parâmetro estabelece o ponto de transição no qual a probabilidade de ser atribuída a categoria k é igual a probabilidade de ser atribuída a categoria adjacente $(k-1)$ para cada item i. O parâmetro b_{ik} é também denominado de parâmetro do item, locação ou limiar do item, mas neste caso, relacionado a um quadro global de referência do item ao longo da variável latente no lugar de referir-se apenas à uma categoria particular.

Assim, a equação (8) pode ser reescrita como:

$$\ln\left[\frac{P_{ik}(\theta_j)}{P_{i,k-1}(\theta_j)}\right] = \theta_j - (b_i + d_{ik}) = \theta_j - b_{ik}. \tag{10}$$

Um dos objetivos de estabelecer esses modelos da TRI para itens politômicos neste trabalho, é a construção dos conceitos necessários para a definição e correto entendimento do *Modelo multifacetas de Rasch* que será utilizado em aplicações práticas posteriormente. Para o modelo multifacetas de Rasch, o entendimento e análises geralmente são mais fáceis, quando o parâmetro b_{ik} é utilizado separadamente, isto é, $b_{ik} = b_i + d_{ik}$.

Cada parâmetro b_{ik} corresponde ao valor de habilidade (ou da dificul-

dade do item) na qual o indivíduo tem a mesma probabilidade de responder à categoria k e à categoria $(k-1)$, isto é, o valor de θ_j no qual $P_{ik}(\theta_j) = P_{i(k-1)}(\theta_j)$. Desse modo, para itens com (m_i+1) categorias de resposta, o número de parâmetros de item a serem estimados será m_i (ANDRADE; TAVARES; VALLE, 2000).

3.3 MODELO MULTIFACETAS DE RASCH – MFR

O modelo Multifacetas de Rasch (MFR) é adequado quando são necessárias análises simultâneas de múltiplas variáveis que são fontes responsáveis pela ocorrência de erros nas avaliações. Esse modelo incorpora mais parâmetros (facetas), além das duas variáveis tradicionalmente presentes nas situações de avaliação. Por exemplo, nas avaliações com itens de respostas construídas em geral, podem ser incorporadas, além da habilidade do examinando e da dificuldade das tarefas, outras variáveis como a severidade dos avaliadores, a estrutura da escala de avaliação, entre outras.

O modelo multifacetas de Rasch de três facetas conforme apresentado por Linacre (1994), para ser aplicado a testes que possuem uma única escala de classificação para todos os avaliadores em todos os itens (*Modelo de escala gradual*), é dado por:

$$\ln\left[\frac{P_{jihk}}{P_{jih(k-1)}}\right] = \theta_j - b_i - c_h - d_k \tag{11}$$

e pelas restrições a seguir, sendo que as duas últimas são estabelecidas para cada item i e para cada avaliador h

$$\sum_{i=1}^{I} b_i = 0, \ \sum_{h=1}^{H} c_h = 0, \ d_0 = 0, \ \text{e} \ \sum_{k=1}^{m} d_k = 0 \tag{12}$$

onde

$P_{ihk}(\theta_j)$ é a probabilidade do indivíduo j de ser classificado na categoria k do item i, pelo avaliador h.

$P_{ih(k-1)}(\theta_j)$ é a probabilidade de um indivíduo j ser classificado na categoria $k-1$ do item i, pelo avaliador h.

θ_j é a **habilidade** do indivíduo j.

b_i é a **dificuldade** do item i.

d_k é o **tamanho do passo** k, é o parâmetro da dificuldade, regula a probabilidade de ser atribuída ao indivíduo a categoria k em relação à categoria $k-1$.

c_h é a **severidade** do avaliador h.

Neste caso, como foi acrescido mais um conjunto de parâmetros ao modelo de Rasch de escala gradual, é necessária a imposição de mais uma restrição para garantir a identificabilidade do modelo, esta restrição é dada por $\sum_{h=1}^{H} c_h = 0$.

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 (11) coloca todos esses parâmetros em uma escala comum na unidade *log-odds* ou *logito*.

Essa trasformação logística da razão entre as probabilidades de se observarem categorias sucessivas (*log-odds*) pode ser entendida como a variável dependente, e as várias facetas, como a habilidade dos examinandos, a dificuldade das tarefas, a severidade dos avaliadores, são conceitualizados como variáveis independentes que influenciam esses *log-odds* (ECKES, 2011).

A equação (11) refere-se ao modelo multifacetas de Rasch de escala gradual de três facetas, a habilidade do examinando, a dificuldade do item e a severidade do avaliador. Outras facetas podem ser incorporadas, por exemplo, tarefas diferentes ou grupos específicos de pessoas. Nesse caso, pode-se ter como objetivo analisar o comportamento diferencial dos itens, das tarefas ou dos avaliadores em relação aos grupos distintos de pessoas. Se apenas duas facetas forem consideradas, a habilidade do examinando e a dificuldade do item, o modelo descrito na equação (11) se resume ao modelo de escala gradual original de Andrich, dado pela equação (5).

A aplicação prática deste trabalho visa análises de uma prova composta de duas tarefas de escrita com cinco itens cada uma delas pontuadas por, no mínimo, dois avaliadores distintos. Desse modo, o modelo multifacetas de Rasch utilizado neste trabalho é de quatro facetas, a habilidade do examinando, a dificuldade da tarefa, a dificuldade do item e a severidade do avaliador. Este modelo assim configurado é dado por:

$$\ln\left[\frac{P_{jiphk}}{P_{jiph(k-1)}}\right] = \theta_j - b_i - t_p - c_h - d_k \tag{13}$$

neste caso, t_p denota a dificuldade da tarefa p, as outras variáveis são como as estabelecidas na equação (11).

Como neste modelo foi acrescido mais um conjunto de parâmetros

(t_p), é necessário impor mais uma restrição para garantir a identificabilidade. Desse modo, tem-se a exigência $\sum_{p=1}^{P} t_p = 0$, onde P denota o número de tarefas.

Assim, as restrições exigidas para este modelo, são

$$\sum_{i=1}^{I} b_i = 0, \ \sum_{h=1}^{H} c_h = 0, \ \sum_{p=1}^{P} t_p = 0, \ d_0 = 0 \text{ e } \sum_{k=1}^{m} d_k = 0. \tag{14}$$

As duas últimas restrições referem-se a cada item, a cada tarefa e a cada avaliador.

Cada categoria sucessiva representa "um passo" de melhoria de desempenho em relação à categoria anterior no traço que está sendo avaliado. O termo d_k define a escala de classificação como tendo a mesma estrutura para todos os itens, todos os avaliadores e todas as tarefas.

O parâmetro de categoria d_k especifica que a estrutura de escala de avaliação utilizada é a referente ao modelo de escala gradual (ANDRICH, 1978 *apud* LINACRE, 1994), sendo esta comum a todos os itens do teste e a todos os avaliadores. Em uma avaliação com itens abertos, por exemplo, todos os avaliadores devem entender e utilizar a estrutura da escala de classificação do mesmo modo, e o nível de desempenho relativo a cada uma das categorias deve ser o mesmo para todos os itens. Os limiares das categorias são estimados em conjunto para todos os avaliadores, todas as tarefas e todos os examinandos participantes do teste.

Esse modelo de "passo comum" permite a estimativa das diferenças entre a severidade dos avaliadores, possibilitando a identificação dos avaliadores causadores de "viés" da calibração dos itens e medição de examinandos (LINACRE; WRIGHT, 2002). A comparação entre o significado dos parâmetros de categoria no modelo original de escala gradual (eq. (5)) e do modelo de três facetas para escala gradual é que este último modela a pontuação que um avaliador atribuiu a uma determinada categoria e a um examinando, com base no que esse avaliador acredita caracterizar o desempenho desse examinando, enquanto o modelo original, modela a resposta de um examinando a uma categoria em particular da escala de avaliação. Isto significa que o parâmetro de categoria d_k não se refere à dificuldade de resposta à categoria k em relação à categoria $k-1$, e sim à dificuldade de se receber a resposta na categoria k em relação à categoria $k-1$ da escala de avaliação (ECKES, 2011).

O modelo estabelecido na equação (11) refere-se ao modelo de três

facetas de escala gradual. Entretanto, o modelo multifacetas de Rasch pode ser expressado de muitas outras formas, dependendo das exigências de cada situação de teste em particular. O modelo que permite que a escala de classificação possa ser "corrigida" através dos itens por meio da utilização de elementos de cada uma das facetas em particular é o modelo para *escala de crédito parcial*. Para a definição desse modelo, é necessário alterar a especificação do parâmetro de categoria d_k dado na equação (11), que passa a ter índice duplo indicando o modo com que cada um dos elementos das facetas interage com as categorias da escala para cada item.

O modelo multifacetas de Rasch de três facetas para a escala de crédito parcial é dado por:

$$\ln\left[\frac{P_{jihk}}{P_{jih(k-1)}}\right] = \theta_j - b_i - c_h - d_{ik} \tag{15}$$

juntamente com as restrições

$$\sum_{i=1}^{I} b_i = 0, \ \sum_{h=1}^{H} c_h = 0, \ d_{i0} = 0, \ \text{e} \ \sum_{k=1}^{m_i} d_{ik} = 0 \tag{16}$$

para cada item i e cada avaliador h.

Todos os parâmetros são como o especificado na equação (11), exceto o parâmetro de categoria d_{ik}, que passa a representar a dificuldade da categoria k relativa à categoria $k-1$ do item de respostas construídas i pelo avaliador h (LINACRE, 1994).

Do mesmo modo que nos modelos apresentados anteriomente, pode-se também agrupar os parâmetros de dificuldade com os de categorias. Desse modo, $b_{ik} = b_i - d_{ik}$, neste caso, os parâmetros b_{ik} representam os limiares ou locações dos itens em relação aos itens e as categorias.

Nesse modelo, os tamanhos dos passos entre as categorias de classificação adjacentes variam entre os itens. Os dois termos subscritos de d_{ik} indicam a *altura do degrau* entre a categoria $k-1$ e a próxima categoria superior k como esta foi utilizada pelo avaliador h para cada item i.

Mais especificamente, o termo d_{ik} indica que a escala de classificação para cada item é modelada para ter sua própria estrutura de categorias, permitindo que a escala de classificação varie de um item para outro. Esse modelo multifacetas, assim estabelecido, fornece a possibilidade de análises sobre a escala utilizada para cada item individualmente e também informações sobre o modo como o grupo de avaliadores utilizou cada categoria em cada item (ECKES, 2011).

Os elementos subscritos do termo d, e que interferem na escala de classificação, podem variar dependendo das análises de interesse do estudo em particular. Por exemplo, se o foco do estudo for as relações existentes entre as categorias de classificação (k) e os avaliadores (h), a equação do modelo MFR é dada por:

$$\ln\left[\frac{P_{hihk}}{P_{jih(k-1)}}\right] = \theta_j - b_i - c_h - d_{hk} \tag{17}$$

juntamente com as restrições

$$\sum_{i=1}^{I} b_i = 0, \quad \sum_{h=1}^{H} c_h = 0, \quad d_{h0} = 0, \quad \text{e} \quad \sum_{k=1}^{m_i} d_{hk} = 0 \tag{18}$$

para cada item i e cada avaliador h.

Nesse modelo, o parâmetro de categoria d_{hk} passa a representar o parâmetro da dificuldade, para todos os itens, que regula a probabilidade de ser atribuído ao indivíduo a categoria k em relação à categoria $k-1$ por cada avaliador h.

Neste caso, pode-se também agrupar os parâmetros da severidade do avaliador (c_h) com os parâmetros de categoria (d_{hk}) para definir o parâmetro $c_{hk} = c_h + d_{hk}$. Este parâmetro pode ser interpretado como a severidade com que o avaliador h julga a categoria k do item i. Desse modo, a equação (17) pode ser reescrita como

$$\ln\left[\frac{P_{hihk}}{P_{jih(k-1)}}\right] = \theta_j - b_i - (c_h + d_{hk}) = \theta_j - b_i - c_{hk}. \tag{19}$$

A equação (17) refere-se ao modelo de três facetas de crédito parcial relacionado aos avaliadores. Esse modelo combina o componente do modelo de crédito parcial com as características dos avaliadores e a escala de avaliação aplicada aos itens. Isso significa que é permitido variar a estrutura da escala de avaliação entre os avaliadores (MYFORD; WOLFE, 2004). Com análises baseadas nesse modelo, é possível obter o comportamento dos avaliadores no nível individual ao utilizarem a escala de avaliação para cada uma das tarefas estabelecidas no teste.

A principal diferença entre os modelos de escala gradual e de crédito parcial está relacionada com a localização dos parâmetros de categoria. Esses elementos não constituem uma faceta, mas representam a diferença da dificuldade (ou local sobre a variável latente) entre categorias adjacentes em uma

escala de classificação (ENGELHARD; WIND, 2013).

Podem-se obter outras configurações para o modelo multifacetas para a escala de crédito parcial, por exemplo, o modelo contendo o parâmetro de categoria d_{ihk} significa a dificuldade relativa da categoria k para o item i e o avaliador h. É comum a denominação dessa e de outras configuração do modelo multifacetas de crédito parcial por *modelo híbrido*, pois combina o componente da escala de crédito parcial com os elementos das facetas e com a estrutura da escala de categorias com o modo como esta foi aplicada pelos avaliadores. Em geral, as variantes do modelo de crédito parcial exigem amostras maiores do que os modelos básicos para se obter as estimativas dos parâmetros (MYFORD; WOLFE, 2004; ECKES, 2011).

Como ocorre com o modelo de escala gradual (eq. (11)), se forem consideradas apenas duas facetas no modelo de crétito parcial descrito pela equação (15), a habilidade dos examinandos e a dificuldade dos itens, o modelo resultante é o de crédito parcial original conforme proposto por Masters (eq. (8)).

Os modelos da família de modelos de Rach compartilham propriedades desejáveis para as medidas, os escores do teste são estatísticas suficientes para a estimação de cada parâmetro, a ordenação dos itens e dos indivíduos são consistentes e a objetividade das medidas proporciona comparações invariantes: a) as comparações entre as habilidades das pessoas são invariantes em relação ao conjunto de itens usados para determinar essas habilidades e b) as comparações entre as medidas da dificuldade dos itens são invariantes em relação ao grupo de pessoas específicas usadas para determinar essas medidas (EMBRETSON, 2000).

Essas propriedades, presentes nos modelos da família de modelos de Rasch, estão entre as exigidas sobre os modelos matemáticos para que as medidas resultantes sejam ideais, tais propriedades são comumente denominadas *medidas invariantes* (ENGELHARD, 2013; EMBRETSON, 2000; RASCH, 1960)

Engelhard (2013) propõe cinco condições básicas para a obtenção de medidas invariantes e estabelece essas condições como a chave para obter medições ideais.

Medida das pessoas

1. A medida da habilidade dos examinandos em relação à variável latente deve ser independente do conjunto particular de itens usados para a obtenção da medida.
2. Examinandos com maior habilidade devem ter maior probabilidade de

sucesso em um item do que examinandos com menor habilidade.

Calibração dos itens

3. A calibração dos itens deve ser independente do conjunto particular de examinandos usados para a calibração.
4. Todo examinando deve ter maior probabilidade de sucesso em um item fácil do que em um item mais difícil.

Mapa das variáveis

5. Os examinandos e os itens devem ser simultaneamente alocados na escala da variável latente.

Segundo Engelhard (2013), essas condições são estendidas para as avaliações mediadas por avaliadores e são úteis para cobrir problemas encontrados nessas avaliações bem como para estabelecer diretrizes para análises da qualidade psicométrica de escalas de avaliação que são comuns nas avaliações mediadas por avaliadores. As condições para as medidas invariantes nas avaliações mediadas por avaliadores são:

Medida dos examinandos

A1. A medida da habilidade dos examinandos em relação à variável latente deve ser independente do conjunto particular de avaliadores que julgaram os desempenhos para a obtenção da medida.
A2. Examinandos com maior habilidade devem ter maior probabilidade de receber maior pontuação dos avaliadores do que pessoas com menor habilidade.

Calibração dos itens

A3. A calibração dos avaliadores deve ser independente do conjunto particular de examinandos usados para a calibração.
A4. Todo examinando deve ter maior probabilidade de receber maior pontuação de um avaliador complacente do que de um mais severo.

Mapa das variáveis

A5. Os examinandos e os avaliadores devem ser simultaneamente alocados na escala da variável latente.

Essas condições também podem ser estendidas para os critérios de avaliação (itens) usados para definir os aspectos a serem avaliados e guiar os

avaliadores nos seus julgamentos, como na pontuação analítica, por exemplo, que subdivide o domínio de conteúdos a ser avaliado em alguns critérios, que podem ser tratados como itens.

O modelo multifacetas satisfaz as exigências de objetividade da mesma maneira que os outros modelos Rasch (Linacre, 1994) e, consequentemente, também satisfaz os requisitos matemáticos para medidas invariantes (ENGELHARD, 2013; LINACRE; WRIGHT, 2002).

3.4 ESTIMAÇÃO DOS PARÂMETROS

A estimação é um processo estatístico básico utilizado para se obter estimativas para os parâmetros de um modelo. Os métodos de estimação para os modelos de Rasch podem ser categorizados como não iterativos e iterativos. Os métodos não iterativos envolvem a estimação de equações que podem ser resolvidas de forma fechada. Métodos iterativos requerem métodos numéricos com múltiplos passos para a obtenção das estimativas, por exemplo, método de Newton-Raphson. Na verdade, processos iterativos são definidos como métodos numéricos que envolvem um valor inicial estimado ou "chutado" para a solução da equação e, a partir daí, são feitas correções desse valor repetidas vezes para obter aproximações melhores desse valor desconhecido. As correções repetidas são denominadas iterações, que são feitas até se alcançar um critério aceitável e pré-estabelecido de parada. Para os modelos da TRI, exemplos de métodos iterativos são o método JMLE (*Joint Maximum Likelihood Estimation*), o método MML (Marginal Maximum Likelihood) e o método CML (*Conditional Maximum Likelihood*). Já, exemplos de métodos não iterativos são o método LOG, o método PAIR e o método PROX (ENGELHARD, 2013).

De acordo com Linacre (1994), os parâmetros dos modelos multifacetas de Rasch não podem ser observados ou estimados diretamente, eles devem ser obtidos por meio de iterações uns com os outros para produzir as medidas que representem os dados. Desse modo, a severidade dos avaliadores, a dificuldade dos itens, a habilidade dos examinandos, entre outras, somente podem ser estimadas por meio das pontuações atribuídas pelos avaliadores às respostas que os examinandos deram aos itens.

O método mais comum utilizado pelos programas de computador existentes atualmente para estimar os parâmetros dos modelos de Rasch é o JMLE, inclusive pelo programa *Facets*, que é utilizado na aplicação prática deste trabalho. Esse método de estimação, no contexto das medidas de Rasch,

é também denominado por UCON (*Unconditional estimation algorithm*).

3.4.1 Considerações sobre a Estimação dos parâmetros

As estimativas dos parâmetros não são todas igualmente boas. Consistência e viés são dois aspectos que podem assegurar ou não a qualidade estatística das estimativas. A consistência diz respeito à precisão com que as medidas são calculadas, isso significa que, para uma amostra significativamente grande, as estimativas devem tender à medida do parâmetro, o que estabelece que a consistência é uma propriedade assintótica. Por outro lado, para um conjunto finito de observações, o viés estatístico diz respeito ao grau com que o valor médio das estimativas difere da medida do parâmetro correspondente (LINACRE, 1994).

Uma vez que a consistência é uma propriedade assintótica, ela só poderá ser obtida nas estimativas dos parâmetros quando o número de observações para esses parâmetros for conceitualmente ilimitado. Para os modelos da TRI, especialmente para os modelos de Rasch, existem dois tipos de parâmetros, incidentais e estruturais (NEYMAN; SCOTT, 1948 *apud* ANDRADE; TAVARES; VALLE, 2000). Para exemplificar, seja um teste tradicional com um número limitado de itens desenvolvidos para ser aplicado a um número ilimitado de pessoas. Os parâmetros relacionados aos itens são denominados *estruturais* uma vez que cada um deles pode aparecer em um número ilimitado de observações (uma para cada examinando). Já os parâmetros correspondentes aos indivíduos aparecem em apenas algumas observações, mais precisamente, o número correspondente ao número de itens do teste. Esses parâmetros são denominados *incidentais*. As estimativas desses parâmetros incidentais não podem ter propriedades assintóticas, pois possuem um número finito de observações. Desse modo, podem não ser consistentes. Além disso, as estimativas de máxima verossimilhança para os parâmetros estruturais também podem ser inconsistentes se os parâmetros incidentais estiverem presentes na formulação do modelo para a determinação da probabilidade (LINACRE, 1994). Essas inconsistências das estimativas podem gerar o problema denominado *falta de identificabilidade* do modelo, que ocorre quando mais de um conjunto de parâmetros produz o mesmo valor da verossimilhança. Uma maneira de eliminar a não identificabilidade é definindo uma métrica para os parâmetros dos indivíduos (habilidade) e, consequentemente, para os parâmetros dos itens (dificuldade) (ANDRADE; TAVARES; VALLE, 2000).

Quando os parâmetros dos itens são conhecidos e deseja-se estimar os parâmetros dos indivíduos, ou vice-versa, o problema de estimação torna-se relativamente simples, pois a escala na qual os parâmetros, tanto dos itens como os dos indivíduos, são estimados é a mesma na qual os parâmetros já conhecidos foram estimados. Assim é eliminada a não identificabilidade do modelo, embora, nos processos de estimação por máxima verossimilhança conjunta, esse problema ainda persista (AZEVEDO, 2003). No entanto, se o modelo puder ser reformulado de modo que os parâmetros incidentais não estejam presentes, podem-se obter algumas condições de regularidade para que as estimativas dos parâmetros estruturais de máxima verossimilhança sejam consistentes (LINACRE, 1994).

Na aplicação prática deste trabalho é utilizado o programa *Facets* que, por sua vez, utiliza, para a estimação dos parâmetros, o método JMLE (WRIGHT; PANCHAPKESAN, 1969). Essa abordagem maximiza simultaneamente a verossimilhança do escore marginal correspondente a cada parâmetro sem fazer suposições sobre a distribuição dos parâmetros. Como os cálculos não são condicionados pelos escores marginais dos parâmetros, essa técnica é conhecida também como UCON (*unconditional maximum likelihood*).

Segundo Linacre (1994), esse método é computacionalmente eficiente para o modelo de duas facetas, uma vez que o número de operações, na pior das hipóteses, aumenta linearmente com o número de observações empíricas, isso para uma dada escala de classificação. Além disso, problemas computacionais relacionados com a perda de precisão raramente são encontrados, desde que o teste seja considerado bem construído.

No entanto, os métodos de estimação conjunta têm algumas deficiências. Considerando-se ainda o modelo de duas facetas, a estimativa conjunta de parâmetros incidentais (habilidade dos indivíduos) e estruturais (dificuldade dos itens) pode levar a estimativas que não são consistentes com o aumento do tamanho da amostra. Isto é, para os modelos de Rasch, o número de parâmetros a ser estimado aumenta à medida que o número de examinados ou de itens aumenta. Quando o número dos parâmetros aumenta com as observações, é possível que as estimativas de máxima verossimilhança conjunta sofram falta de consistência, eficiência e normalidade assintótica (NEYMAN; SCOTT, 1948 *apud* LINACRE, 1994).

Por outro lado, quando o número de itens e o número de indivíduos crescem simultaneamente e numa mesma proporção, os estimadores de máxima verossimilhança de ambos os tipos de parâmetros são consistentes (HABERMAN, 1977). Resultados semelhantes também foram constatados

por Lord (1968; 1975) e Swaminathan e Gifford (1983) conforme citado por Azevedo (2003).

Outro fato destacado por Haberman (1977) é que a presença de valores extremos (dados faltantes ou pontuação perfeita), tornam os parâmetros correspondentes inestimáveis, o que causa viés nas estimativas dos demais parâmetros. Em geral, a probabilidade de valores extremos é reduzida quando o número de indivíduos e de itens aumenta.

3.4.2 Método de estimação JMLE para o modelo de Rasch dicotômico

O método de estimação iterativo JMLE é estabelecido nesta seção para o modelo de Rasch dicotômico. Esse método de estimação para o modelo multifacetas de Rasch será abordado na Seção 3.4.5.

Sejam θ_j, $j = 1, \cdots N$ a habilidade do indivíduo j e x_{ji} a variável aleatória que representa a resposta do indivíduo j ao item i, dada por

$$x_{ji} = \begin{cases} 1, & \text{resposta correta} \\ 0, & \text{resposta incorreta.} \end{cases}$$

Cada uma das observações é denotada por x_{ji}, o vetor de habilidades dos N indivíduos é denotado por $\theta = (\theta_1, \cdots, \theta_N)$ e o conjunto dos parâmetros dos I itens é denotado por $\beta = (b_1, \cdots, b_I)$.

O modelo de Rasch estabelece que a probabilidade do indivíduo j de responder corretamente ao item i é dada por

$$P(x_{ji} = 1|\theta_j, b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}} \tag{20}$$

e a probabilidade do indivíduo j de responder ao item i incorretamente é dada por

$$P(x_{ji} = 0|\theta_j, b_i) = 1 - P(x_{ji} = 1|\theta_j, b_i) = \frac{e^{-(\theta_j - b_i)}}{1 + e^{-(\theta_j - b_i)}}. \tag{21}$$

Para simplificar a notação, no restante deste trabalho será utilizado simplesmente P_{ji} para denotar a probabilidade da pessoa de responder corretamente ao item i dada a sua habilidade, isto é, $P_{ji} = P(x_{ji} = 1|\theta_j, b_i)$.

Se as respostas dadas por indivíduos diferentes são independentes e

os itens são respondidos de forma independente por cada indivíduo quando fixada a sua habilidade (independência local), a verossimilhança de observar um vetor particular de respostas para a pessoa é

$$L(\theta,\beta) = \prod_{j=1}^{N}\prod_{i=1}^{I}[P_{ji}]^{x_{ji}}\,[1-P_{ji}]^{1-x_{ji}}\,. \tag{22}$$

A log-verossimilhança pode ser escrita como

$$\ln\left(L(\theta,\beta)\right) = \sum_{j=1}^{N}\sum_{i=1}^{I}\left[x_{ji}\ln\left(P_{ji}\right) + \left(1-x_{ji}\right)\ln\left(1-P_{ji}\right)\right]. \tag{23}$$

Se os parâmetros das habilidades das pessoas são conhecidos (θ), a equação (22) é utilizada para obter as estimativas de máxima verossimilhança para os parâmetros dos itens (β). Para tanto, os estimadores de máxima verossimilhança (EMV) de b_i ($i = 1,\cdots,I$) são os valores que maximizam a verossimilhança ou equivalentemente a log-verossimilhança. Isto é, são as soluções de

$$\frac{\partial \ln\left(L(\theta,\beta)\right)}{\partial b_i} = 0. \tag{24}$$

Do mesmo modo, se os parâmetros dos itens são conhecidos (β), a equação de verossimilhança é utilizada para obter os EMV para os parâmetros das habilidades (θ_j, $j = 1\cdots N$). Nesse caso, devem-se obter as soluções de

$$\frac{\partial \ln\left(L(\theta,\beta)\right)}{\partial \theta_j} = 0. \tag{25}$$

Quando tanto os valores de β quanto os de θ são desconhecidos e devem ser estimados, o problema de estimação é mais difícil. Isso ocorre porque diferentes valores dos parâmetros podem levar a um mesmo valor da verossimilhança, o que compromete o processo de obtenção das estimativas. Nesse caso, é necessário impor uma restrição para se obter uma única solução.

A denominação do método, *Joint Maximum Likelihood Estimation*, é derivada do processo de estimação passo a passo que envolve tanto a estimação da locação dos itens quanto a habilidade das pessoas, desse modo, *joint* significa em conjunto.

As derivadas de primeira ordem da log-verossimilhança (eq. (23)) em relação a cada parâmetro são dadas na sequência.

A derivada em relação a b_i é dada por

$$\frac{\partial \ln(L(\theta,\beta))}{\partial b_i} = \sum_{j=1}^{N}\left[x_{ji}\left(\frac{1}{P_{ji}}\right)\left(\frac{\partial P_{ji}}{\partial b_i}\right) - (1-x_{ji})\left(\frac{1}{1-P_{ji}}\right)\left(\frac{\partial P_{ji}}{\partial b_i}\right)\right]. \quad (26)$$

Como

$$\frac{\partial P_{ji}}{\partial b_i} = -P_{ji}(1-P_{ji}). \quad (27)$$

O que resulta em

$$\begin{aligned}\frac{\partial \ln(L(\theta,\beta))}{\partial b_i} &= \sum_{j=1}^{N}\left[-x_{ji}(1-P_{ji}) + (1-x_{ji})P_{ji}\right] \\ &= -\sum_{j=1}^{N}\left[x_{ji}-P_{ji}\right]. \quad (28)\end{aligned}$$

A derivada em relação a θ_j é dada por

$$\frac{\partial \ln(L(\theta,\beta))}{\partial \theta_j} = \sum_{i=1}^{I}\left[x_{ji}\left(\frac{1}{P_{ji}}\right)\left(\frac{\partial P_{ji}}{\partial \theta_j}\right) - (1-x_{ji})\left(\frac{1}{1-P_{ji}}\right)\left(\frac{\partial P_{ji}}{\partial \theta_j}\right)\right]. \quad (29)$$

Como

$$\frac{\partial P_{ji}}{\partial \theta_j} = P_{ji}(1-P_{ji}). \quad (30)$$

O que resulta em

$$\begin{aligned}\frac{\partial \ln(L(\theta,\beta))}{\partial \theta_j} &= \sum_{i=1}^{I}\left[x_{ji}(1-P_{ji}) - (1-x_{ji})P_{ji}\right] \\ &= \sum_{i=1}^{I}\left[x_{ji}-P_{ji}\right]. \quad (31)\end{aligned}$$

Desse modo, as equações de estimação são dadas por

$$b_i: \quad -\sum_{j=1}^{N}(x_{ji}-P_{ji})=0 \tag{32}$$

e

$$\theta_j: \quad \sum_{i=1}^{I}(x_{ji}-P_{ji})=0. \tag{33}$$

Denotando as somas das colunas e das linhas dos valores observados por

$$x_{.i}=\sum_{j=1}^{N}(x_{ji}) \qquad \text{e} \qquad x_{j.}=\sum_{i=1}^{I}(x_{ji}) \tag{34}$$

obtêm-se

$$x_{.i}=\sum_{j=1}^{N}P_{ji}=\sum_{j=1}^{N}\frac{1}{1+e^{-(\theta_j-b_i)}} \tag{35}$$

e

$$x_{j.}=\sum_{i=1}^{I}P_{ji}=\sum_{i=1}^{I}\frac{1}{1+e^{-(\theta_j-b_i)}}. \tag{36}$$

Esse é um caso especial no qual as equações pertencem à família exponencial e as equações de verossimilhança são estatísticas suficientes. Isto significa que essas estatísticas contém toda a informação relevante sobre o parâmetro. Desse modo, os valores observados (x_{ji}) são iguais aos valores gerados pelo modelo (FISCHER; MOLENAAR, 1995). Quando os dados estão completos, há $N+I$ equações, embora grupos de pessoas ou grupos de itens com as mesmas estatísticas suficientes resultem em equações idênticas.

Reescrevendo esse sistema, têm-se, para os parâmetros dos itens,

$$\begin{aligned} x_{.1} &= P_{11}+P_{21}+\cdots+P_{N1} \\ x_{.2} &= P_{12}+P_{22}+\cdots+P_{N2} \\ &\vdots \\ x_{.I} &= P_{1I}+P_{2I}+\cdots+P_{NI} \end{aligned} \tag{37}$$

e, para os parâmetros das habilidades,

$$\begin{aligned} x_{1.} &= P_{11} + P_{12} + \cdots + P_{1I} \\ x_{2.} &= P_{21} + P_{22} + \cdots + P_{2I} \\ &\vdots \\ x_{N.} &= P_{N1} + P_{N2} + \cdots + P_{NI}. \end{aligned} \tag{38}$$

Em aplicações reais, normalmente o número de itens I do teste é menor do que o número de pessoas N submetidas a ele. Nota-se que uma das equações relativas aos itens (eq. (37)) sempre pode ser escrita como combinação linear das equações dos parâmetros das pessoas (eq. (38)), o que resulta em um sistema com $N+I$ incógnitas e $N+(I-1)$ equações independentes, desse modo, com mais de uma solução. Esse fato gera o problema da não-identificabilidade do modelo. Demonstrações formais desse resultado podem ser conferidas em San Martin *et al.* (2009), San Martin e Rolin (2013) e Noventa *et al.* (2014). Para se obter solução única, é necessário impor alguma restrição, a usual é $\sum_i b_i = 0$, mas pode-se também exigir que a média dos parâmetros das pessoas seja igual a 0 ou, então, pode-se escolher alguma outra restrição conveniente (DE AYALA, 2000; FISCHER; MOLENAAR, 1995).

Na prática, comumente utiliza-se duas abordagens. Para a primeira, deve-se fixar a média da habilidade das pessoas (θ_j) na origem da escala e depois de cada passo da estimação, a média dos parâmetros é novamente centrada na origem. Para a segunda abordagem, deve-se fixar a média dos parâmetros dos itens (b_i) na origem e do mesmo modo, atualizar a média em zero depois de cada iteração. Embora estes dois métodos, provavelmente, produzam estimativas dos parâmetros diferentes, o significado relativo dos resultados não é afetado pela escolha do método de fixação da escala escolhido (DE AYALA, 2000).

Em relação ao efeito da falta de identificabilidade do modelo quando se utiliza o algoritmo de estimação JMLE, observa-se que primeiramente, as locações dos itens são estimadas utilizando-se estimativas provisórias das locações das pessoas. Então, a locação dos itens é centrada por meio da diferença entre as estimativas da locação das pessoas e a média dessas estimativas, esta é a iteração inicial ou a iteração [0] para a locação dos itens.

No passo seguinte, as locações das pessoas são estimadas utilizando-se as estimativas das locações dos itens calculadas no passo anterior. Subsequentemente, as locações dos itens são re-estimadas utilizando-se as novas

estimativas das locações das pessoas, e assim sucessivamente. Como, nesse processo, a estimação é uma melhoria da estimação alcançada no passo anterior, a média da métrica começa a se distanciar da origem à medida que as iterações avançam, é por isso que são necessárias atualizações da média e do centro das locações após cada passo da estimação, quando se utiliza o algoritmo JMLE (De Ayala, 2000). O algoritmo de estimação dos parâmetros pelo método JMLE para o modelo de Rasch para itens dicotômicos está descrito no Anexo B, Quadro 29.

Para determinar a solução das equações (32) e (33), pode-se utilizar o algoritmo de Newton-Raphson, que, ignorando-se covariância, para os parâmetros das habilidades θ, é

$$\theta_j^{(k+1)} = \theta_j^{(k)} - \left(\frac{\partial \ln(L(\theta,\beta))}{\partial \theta_j} \right) / \left(\frac{\partial^2 \ln(L(\theta,\beta))}{\partial \theta_j^2} \right) \tag{39}$$

em que θ^k é a estimativa inicial da locação da habilidade da pessoa e θ^{k+1} é a estimativa seguinte. As iterações são feitas até que a diferença entre valores consecutivos de θ seja demasiadamente pequena.

Similarmente, para os parâmetros dos itens β, a equação de Newton-Raphson é

$$b_i^{(k+1)} = b_i^{(k)} - \left(\frac{\partial \ln(L(\theta,\beta))}{\partial b_i} \right) / \left(\frac{\partial^2 \ln(L(\theta,\beta))}{\partial b_i^2} \right). \tag{40}$$

Derivando novamente a equação (28) em relação a b_i e a equação (31) em relação a θ_j, obtêm-se

$$\frac{\partial^2 \ln(L(\theta,\beta))}{\partial b_i^2} = -\sum_{j=1}^{N} P_{ji}(1-P_{ji}) \tag{41}$$

e

$$\frac{\partial^2 \ln(L(\theta,\beta))}{\partial \theta_j^2} = -\sum_{i=1}^{I} P_{ji}(1-P_{ji}). \tag{42}$$

Desse modo

$$b_i^{(k+1)} = b_i^{(k)} - \frac{\sum_{j=1}^{N}(x_{ji} - P_{ji}^k)}{\sum_{j=1}^{N} P_{ji}^k(1 - P_{ji}^k)} \tag{43}$$

e

$$\theta_j^{(k+1)} = \theta_j^{(k)} - \frac{\sum_{i=1}^{I}(x_{ji} - P_{ji}^k)}{-\sum_{i=1}^{I} P_{ji}^k(1 - P_{ji}^k)}. \tag{44}$$

Os passos para a estimação dos parâmetros para o modelo de Rasch para itens dicotômicos pelo método JMLE utilizando o algoritmo de Newton-Raphson são resumidos nos Quadros 29 e 30 dispostos no Anexo B.

3.4.3 Método de estimação JMLE para o modelo de Rasch para itens politômicos

A equação para o modelo de escala gradual na forma *logitos* é dada por

$$\ln\left[\frac{P_{jik}}{P_{ji(k-1)}}\right] = \theta_j - b_i - d_k. \tag{45}$$

Os parâmetros relacionados a um conjunto particular de examinandos e itens interagem para produzir cada uma das observações separadamente. Assim, a equação imediatamente satisfaz dois requisitos da objetividade. Para o primeiro, cada componente da situação de julgamento é caracterizado por um parâmetro independente dos outros parâmetros. Para o segundo, os parâmetros combinam aditivamente para obter as probabilidades das avaliações empíricas. Existe ainda um terceiro requisito para a objetividade das medidas relacionado às estimativas, com a exigência de que a estimativa de cada parâmetro é dependente somente do escore no qual ele participa. Isso significa que a soma do escore no qual o parâmetro participa é estatística suficiente para a estimação desse parâmetro.

Resumindo o modelo dado na equação (45) para a categoria k, elimi-

nando as probabilidades dos passos intermediários até a categoria 1, resulta em

$$\ln\left[\frac{P_{jik}}{P_{ji0}}\right] = k(\theta_j - b_i) - \sum_{s=1}^{k} d_s. \tag{46}$$

Desse modo, a probabilidade P_{jik} associada com a categoria k na forma exponencial é dada por:

$$P_{jik} = P_{ji0}\left[\exp\left(k(\theta_j - b_i) - \sum_{s=1}^{k} d_s\right)\right]. \tag{47}$$

É exigido que $\sum_{k=0}^{m} P_{jik} \equiv 1$. Somando-se ambos os membros da equação (47) para as categorias de 0 até m, obtem-se

$$1 = P_{ji0} + \sum_{k=1}^{m} P_{ji0}\left[\exp\left(k(\theta_j - b_i) - \sum_{s=1}^{k} d_s\right)\right] \tag{48}$$

o que resulta na fórmula para a probabilidade da categoria 0

$$P_{ji0} = 1/\left\{1 + \sum_{k=1}^{m} \exp\left[k(\theta_j - b_i) - \sum_{s=1}^{k} d_s\right]\right\}. \tag{49}$$

Substituindo esse resultado na equação (47), obtém-se a fórmula da probabilidade para as categorias $k = 1 \cdots m$

$$P_{jik} = \frac{\exp\left[k(\theta_j - b_i) - \sum_{s=1}^{k} d_s\right]}{1 + \sum_{r=1}^{m} \exp\left[r(\theta_j - b_i) - \sum_{s=1}^{r} d_s\right]}. \tag{50}$$

Se a dificuldade da categoria 0 é definida como sendo d_0, tanto o numerador quanto o denominador da equação (50) podem ser multiplicados por $\exp(-d_0)$, o que resulta em

$$P_{jik} = \frac{\exp\left[k(\theta_j - b_i) - \sum_{s=0}^{k} d_s\right]}{\sum_{r=0}^{m} \exp\left[r(\theta_j - b_i) - \sum_{s=0}^{r} d_s\right]}. \tag{51}$$

A equação (51) é a forma exponencial do modelo de Rasch de escala gradual.

3.4.4 Equações de estimação para o modelo de Rasch para itens politômicos

O método JMLE utiliza em seu algoritmo de estimativa incondicional a abordagem da probabilidade máxima, na qual é assumido que as diferenças entre as observações empíricas e os respectivos valores teóricos esperados, com base nos valores dos parâmetros reais não observáveis, estão normalmente distribuídos. Desse modo, as diferenças normalmente distribuídas podem ser testadas pelo exame das diferenças residuais entre as classificações observadas e os valores esperados com base nas estimativas dos parâmetros geradas pelo modelo (LINACRE, 1994).

O método JMLE será desenvolvido nesta seção para o modelo de Rasch de escala gradual, isto é, para ser aplicado a testes que possuem uma única escala de classificação para todos os itens. Para o modelo de crédito parcial, o algoritmo poderá ser facilmente estendido.

Sejam o vetor de habilidades dos N indivíduos denotado por $\theta = (\theta_1, \cdots, \theta_N)^T$, o conjunto dos parâmetros dos I itens denotado por $\beta = (b_1, \cdots, b_I)^T$ e cada uma das observações denotada por x_{ji}, indicando que a observação é proveniente da resposta do indivíduo j ao item i.

$$\text{Prob}(x_{ji}|\theta, \beta) = \frac{\exp\left[k(\theta_j - b_i) - \sum_{s=0}^{k} ds\right]}{\sum_{r=0}^{m} \exp\left[r(\theta_j - b_i) - \sum_{s=0}^{r} ds\right]}. \tag{52}$$

Por conveniência, para facilitar as manipulações, seja

$$G_k = \sum_{s=0}^{k} ds = d_0 + d_1 + \cdots + d_k$$

o que resulta

$$d_k = G_k - G_{k-1}.$$

Reescrevendo o modelo, com a restrição de que a soma das probabilidades de todas as categorias da escala de classificação é 1, e utilizando a notação $\text{Prob}(x_{ji}|\{\theta,\beta\}) = P_{jix}$ para designar a probabilidade do indivíduo j escolher a categoria x_{ji} no item i.

$$P_{jix} = \frac{\exp\left[x_{ji}\left(\theta_j - b_i\right) - G_{x_{ji}}\right]}{\sum_{k=0}^{m} \exp\left[k\left(\theta_j - b_i\right) - G_k\right]}. \tag{53}$$

A verossimilhança é dada por

$$L(\theta,\beta) = \prod_{j=1}^{N}\prod_{i=1}^{I}(P_{jix}). \tag{54}$$

A log-verossimilhança é

$$\begin{aligned}
\ln(L(\theta,\beta)) &= \sum_{j=1}^{N}\sum_{i=1}^{I} \ln\left\{\frac{\exp\left[x_{ji}\left(\theta_j - b_i\right) - G_{x_{ji}}\right]}{\sum_{k=0}^{m} \exp\left[k\left(\theta_j - b_i\right) - G_k\right]}\right\} \\
&= \sum_{j=1}^{N}\sum_{i=1}^{I}\left\{x_{ji}\left(\theta_j - b_i\right) - G_{x_{ji}}\right\} - \\
&- \sum_{j=1}^{N}\sum_{i=1}^{I}\left\{\ln\left[\sum_{k=0}^{m}\exp\left(k\left(\theta_j - b_i\right) - G_k\right)\right]\right\}.
\end{aligned} \tag{55}$$

A derivada parcial de primeira ordem da log-verossimilhança em relação a θ_j é

$$\frac{\partial \ln(L(\theta,\beta))}{\partial \theta_j} = \sum_{i=1}^{I} x_{jih} - \sum_{i=1}^{I} \frac{\sum_{k=0}^{m} k \exp\left(k\left(\theta_j - b_i\right) - G_k\right)}{\sum_{k=0}^{m} \exp\left(k\left(\theta_j - b_i\right) - G_k\right)}$$

$$= \sum_{i=1}^{I} x_{ji} - \sum_{i=1}^{I} \left[\sum_{k=0}^{m} k P_{jik}\right]. \tag{56}$$

As derivadas parciais da log-verossimilhança em relação ao parâmetro b_i é similar a esta. Observando-se a equação (55), verifica-se facilmente que a derivada em relação ao parâmetro b_i possui a mesma configuração, alterando-se apenas a soma que participa da formulação, que é aquela relacionada ao parâmetro considerado fixo na derivada.

Utilizando-se as notações

$$x_{j.} = \sum_{i=1}^{I} (x_{ji}) \qquad \text{e} \qquad x_{.i} = \sum_{j=1}^{N} (x_{ji}) \tag{57}$$

tem-se que as equações de estimação são

$$\theta: \qquad x_{j.} = \sum_{i=1}^{I} \left[\sum_{k=0}^{m} k P_{jik}\right] \tag{58}$$

e

$$\beta: \qquad x_{.i} = \sum_{j=1}^{N} \left[\sum_{k=0}^{m} k P_{jik}\right]. \tag{59}$$

O sistema resultante possui $N+I$ equações e $N+I$ parâmetros para serem estimados, além das locações das categorias dos itens, que, no modelo de escala gradual, são m (número de categorias), pois $d_0 = 0$ (as locações são comuns para todos os itens). Mas, uma entre as equações das habilidades das pessoas pode ser escrita como combinação linear das equações dos parâmetros dos itens, sendo necessária uma restrição para se obter solução única. Nesse caso, é usual escolher a restrição $\sum_i b_i = 0$ ou $\sum_j \theta_j = 0$, mas, como no caso do modelo de Rasch dicotômico, pode-se escolher outra restrição

conveniente.

Fazendo a derivada da log-verossimilhança também em relação ao parâmetro das categorias da escala de classificação ($G = G_1, G_2, \cdots, G_m$), obtem-se as equações de estimação para os passos de dificuldade:

$$G: \qquad x_k = \sum_{j=1}^{N} \sum_{i=1}^{I} \left(P_{jik}\right) \tag{60}$$

x_k ($k = 1, \cdots, m$) é a soma de todas as respostas na categoria k da escala de classificação. Neste caso, as probabilidades são somadas sobre todas as observações, pois a escala é comum para todos os itens.

Se alguma categoria k não é observada nos dados, então os seus parâmetros não são calculados e, nesse caso, $P_{jik} = 0$ para todo j e i.

Uma outra restrição é também exigida para as categorias da escala de classificação, usualmente é exigido que $\sum_{k=1}^{m} d_k = 0$. Desse modo, para o modelo de Rasch de escala gradual, é necessária uma restrição que pode ser relacionada com a dificuldade dos itens ou com a habilidade dos examinandos e mais duas relacionadas com as categorias da escala. Estas restrições resolvem o problema da não-identificabilidade do modelo:

$$\sum_{i=1}^{I} b_i = 0, \quad d_0 = 0 \quad \text{e} \quad \sum_{k=1}^{m} d_k = 0.$$

A derivada segunda da log-verossimilhança em relação à habilidade é

$$\begin{aligned}
\frac{\partial^2 \ln(L(\theta,\beta))}{\partial \theta_j^2} = & \\
= & -\sum_{i=1}^{I} \left[\frac{\sum_{k=0}^{m} k^2 \exp\left(k\left(\theta_j - b_i\right) - G_k\right)}{\sum_{k=0}^{m} \exp\left(k\left(\theta_j - b_i\right) - G_k\right)} - \left(\frac{\sum_{k=0}^{m} k \exp\left(k\left(\theta_j - b_i\right) - G_k\right)}{\sum_{k=0}^{m} \exp\left(k\left(\theta_j - b_i\right) - G_k\right)} \right)^2 \right] \\
= & -\sum_{i=1}^{I} \left[\sum_{k=0}^{m} k^2 P_{jik} - \left(\sum_{k=0}^{m} k P_{jik} \right)^2 \right].
\end{aligned} \tag{61}$$

Substituindo as derivadas de primeira e de segunda ordem da log--verossimilhança em relação à habilidade (equações (56) e (61)) na equação (39) (algoritmo de Newton Raphson), obtém-se a equação de estimação para

os parâmetros da habilidade (θ_j):

$$\theta'_j = \theta_j - \frac{x_{j.} - \sum_{i=1}^{I}\left[\sum_{k=0}^{m} kP_{jik}\right]}{\sum_{i=1}^{I}\left[\sum_{k=0}^{m} k^2 P_{jik} - \left(\sum_{k=0}^{m} kP_{jik}\right)^2\right]}. \tag{62}$$

As equações de estimação para os parâmetros dos itens (b_i) são similares a esta, embora possua sinal contrário.

$$b'_i = b_i - \frac{\sum_{j=1}^{N}\left[\sum_{k=0}^{m} kP_{jik}\right] - x_{.i}}{\sum_{j=1}^{N}\left[\sum_{k=0}^{m} k^2 P_{jik} - \left(\sum_{k=0}^{m} kP_{jik}\right)^2\right]}. \tag{63}$$

Uma aproximação para o erro padrão assintótico das estimativas de cada um dos parâmetros θ e β é dada pela raiz quadrada do inverso do denominador da equação de estimação respectiva.

Do mesmo modo, substituindo-se as derivadas de primeira e de segunda ordem da log-verossimilhança (eq. (55)) em relação às categorias da escala de classificação na equação de Newton Raphson (eq. (39)), obtém-se a equação de estimação para os passos cumulativos de dificuldade:

$$G'_k = G_k - \frac{\sum_{j=1}^{N}\sum_{i=1}^{I}\left(P_{jik}\right) - x_k}{\sum_{j=1}^{N}\sum_{i=1}^{I}\left(P_{jik} - (P_{jik})^2\right)}. \tag{64}$$

Os parâmetros que correspondem às categorias da escala de classificação são de natureza diferente dos outros dois parâmetros. Eles não são independentes e são impostas restrições mais fortes sobre os passos das estimativas do que as impostas para as estimativas dos outros parâmetros. Por exemplo, no caso de itens dicotômicos, a dificuldade do passo é definida para ser 0 (zero) e as etapas desaparecem das equações de estimação. No entanto, para escalas de avaliação em geral, há dois graus de liberdade a menos do que o número de categorias, e isso tem um efeito considerável sobre a covariância. No entanto, Linacre (1994) afirma que estudos de simulação indicam não

haver diferença significativa nas estimativas quando a covariância é ignorada.

Uma aproximação do erro padrão assintótico para G_k, ignorando a covariância, é dada por

$$\text{S.E.}(G_k) = \left(1/ \sum_{j=1}^{N} \sum_{i=1}^{I} \left(P_{jik} - (P_{jik})^2 \right) \right)^{1/2}.$$

As dificuldades dos passos relacionados com cada categoria da escala de classificação d_k são dadas por

$$d_k = G_k - G_{k-1}, \qquad d_0 \equiv 0. \tag{65}$$

Assumindo que os passos da escala de classificação são independentes e ignorando os efeitos da estrutura da escala de classificação, uma aproximação do erro padrão de d_k é

$$\begin{aligned} \text{S.E.}(d_k) &= \left((\text{S.E.}(G_k))^2 + (\text{S.E.}(G_{k-1}))^2 \right)^{1/2} \\ &= \left(1/ \sum_{j=1}^{N} \sum_{i=1}^{I} \left(P_{jik} - (P_{jik})^2 \right) + 1/ \sum_{j=1}^{N} \sum_{i=1}^{I} \left(P_{jik-1} - (P_{jik-1})^2 \right) \right)^{1/2} \end{aligned} \tag{66}$$

com S.E. $d_0 \equiv 0$.

Wright e Masters (1982 *apud* LINACRE, 1994), propõem que as estimativas do erro padrão de d_k sejam feitas diretamente no lugar de utilizar as estimativas do erro padrão de G_k. Desse modo, as estimativas do erro padrão de d_k são calculadas por

$$\text{S.E.}(d_k) = \left(1/ \sum_{j=k}^{N} \sum_{i=k}^{I} \sum_{h=k}^{H} \left(\sum_{k=0}^{m} (P_{jik}) - (\sum_{k=0}^{m} P_{jik})^2 \right) \right)^{1/2}. \tag{67}$$

Segundo Linacre (1994), a equação (66), em geral, produz erros das estimativas maiores do que os erros produzidos pela equação (67) em replicações idênticas.

3.4.5 Método de estimação JMLE para o modelo multifacetas de Rasch

A equação para o modelo multifacetas de Rasch de três facetas na forma *logitos* é dada por

$$\ln\left[\frac{P_{jihk}}{P_{jih(k-1)}}\right] = \theta_j - b_i - c_h - d_k. \tag{68}$$

Este modelo difere do modelo de Rasch de escala gradual (eq. (45)) apenas pelo acrécimo do termo referente à severidade dos avaliadores (c_h). Do mesmo modo procedido anteriormente, obtem-se a forma exponencial do modelo multifacetas de Rasch de três facetas:

$$P_{jihk} = \frac{\exp\left[k\left(\theta_j - b_i - c_h\right) - \sum_{s=0}^{k} d_s\right]}{\sum_{r=0}^{m} \exp\left[r\left(\theta_j - b_i - c_h\right) - \sum_{s=0}^{r} d_s\right]}. \tag{69}$$

3.4.5.1 Equações de estimação para o modelo multifacetas de Rasch

O método JMLE será desenvolvido nesta seção para o modelo multifacetas de Rasch de três parâmetros de escala gradual, isto é, para ser aplicado a testes que possuem uma única escala de classificação para todos os avaliadores em todos os itens. Para outras configurações desse modelo, o algoritmo poderá ser facilmente estendido.

Neste caso, o vetor de habilidades dos N indivíduos é denotado por $\theta = (\theta_1, \cdots, \theta_N)^T$, o conjunto dos parâmetros dos I itens é denotado por $\beta = (b_1, \cdots, b_I)^T$, o conjunto dos parâmetros do desempenho dos H avaliadores por $\gamma = (c_1, \cdots, c_H)^T$ e cada uma das observações denotada por x_{jih}, indicando que a observação é proveniente da resposta do indivíduo j ao item i devido ao julgamento do avaliador h.

$$\text{Prob}(x_{jih}|\theta,\beta,\gamma) = \frac{\exp\left[k(\theta_j - b_i - c_h) - \sum_{s=0}^{k} ds\right]}{\sum_{r=0}^{m} \exp\left[r(\theta_j - b_i - c_h) - \sum_{s=0}^{r} ds\right]}. \tag{70}$$

A verossimilhança é dada por

$$L(\theta,\beta,\gamma) = \prod_{j=1}^{N}\prod_{i=1}^{I}\prod_{h=1}^{H}(P_{jihx}). \tag{71}$$

A log-verossimilhança é

$$\begin{aligned}
\ln(L(\theta,\beta,\gamma)) &= \sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{h=1}^{H} \ln\left\{\frac{\exp\left[x_{jih}(\theta_j - b_i - c_h) - G_{x_{jih}}\right]}{\sum_{k=0}^{m}\exp\left[k(\theta_j - b_i - c_h) - G_k\right]}\right\} \\
&= \sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{h=1}^{H}\left\{x_{jih}(\theta_j - b_i - c_h) - G_{x_{jih}}\right\} - \\
&- \sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{h=1}^{H}\left\{\ln\left[\sum_{k=0}^{m}\exp\left(k(\theta_j - b_i - c_h) - G_k\right)\right]\right\}. \qquad (72)
\end{aligned}$$

De modo analogo ao modelo de Rasch de escala gradual, deriva-se log-verossimilhança em relação aos parâmetros θ_j, b_j e c_h.

Utilizando-se as notações

$$x_{j..} = \sum_{i=1}^{I}\sum_{h=1}^{H}(x_{jih}), \quad x_{.i.} = \sum_{j=1}^{N}\sum_{h=1}^{H}(x_{jih}) \quad \text{e} \quad x_{..h} = \sum_{i=1}^{I}\sum_{j=1}^{N}(x_{jih}) \tag{73}$$

tem-se que as equações de estimação são

$$\theta: \qquad x_{j..} = \sum_{i=1}^{I}\sum_{h=1}^{H}\left[\sum_{k=0}^{m} kP_{jihk}\right] \tag{74}$$

$$\beta: \qquad x_{.i.} = \sum_{j=1}^{N} \sum_{h=1}^{H} \left[\sum_{k=0}^{m} kP_{jihk} \right] \tag{75}$$

$$\gamma: \qquad x_{..h} = \sum_{j=1}^{N} \sum_{i=1}^{I} \left[\sum_{k=0}^{m} kP_{jihk} \right]. \tag{76}$$

Para o modelo multifacetas de Rasch de três facetas, o sistema resultante possui $N+I+H$ equações e $N+I+H$ parâmetros para serem estimados. Mas, uma entre as equações das habilidades das pessoas e uma entre as equações dos avaliadores podem ser escritas como combinações lineares das equações dos parâmetros dos itens, resultando em $N+I+H-2$ equações independentes, sendo necessárias neste caso, duas restrições para se obter solução única. São usuais as restrições $\sum_i b_i = 0$ e $\sum_h c_h = 0$, mas, podem-se escolher outras restrições convenientes.

Este modelo também exige a estimação das m locações das categorias dos itens, lembrando que as locações são comuns ao longo dos itens e que é exigido que $d_0 = 0$, além de uma outra restrição para as categorias da escala de classificação.

Desse modo, as rescrições que garantem a identificabilidade do modelo são:

$$\sum_{i=1}^{I} b_i = 0, \quad \sum_{h=1}^{H} c_h = 0, \quad d_0 = 0 \quad \text{e} \quad \sum_{k=1}^{m} d_k = 0.$$

As duas últimas restrições são estabelecidas para cada item i e cada avaliador h.

As derivadas de segunda ordem da log-verossimilhança (eq. (72)) em relação a cada um dos parâmetros são similares às obtidas para o modelo de Rasch de escala gradual.

Substituindo as derivadas de primeira e de segunda ordem da log--verossimilhança em relação à cada um dos parâmetros no algoritmo de Newton Raphson (eq. (39)), obtém-se as equações de estimação para os parâmetros da habilidade (θ_j), da dificuldade dos itens (b_i) e da severidade dos avaliadores (c_h):

$$\theta'_j = \theta_j - \frac{x_{j..} - \sum_{i=1}^{I}\sum_{h=1}^{H}\left[\sum_{k=0}^{m} kP_{jihk}\right]}{\sum_{i=1}^{I}\sum_{h=1}^{H}\left[\sum_{k=0}^{m} k^2 P_{jihk} - \left(\sum_{k=0}^{m} kP_{jihk}\right)^2\right]} \tag{77}$$

$$b'_i = b_i - \frac{\sum_{j=1}^{N}\sum_{h=1}^{H}\left[\sum_{k=0}^{m} kP_{jihk}\right] - x_{.i.}}{\sum_{j=1}^{N}\sum_{h=1}^{H}\left[\sum_{k=0}^{m} k^2 P_{jihk} - \left(\sum_{k=0}^{m} kP_{jihk}\right)^2\right]} \tag{78}$$

$$c'_h = c_h - \frac{\sum_{j=1}^{N}\sum_{i=1}^{I}\left[\sum_{k=0}^{m} kP_{jihk}\right] - x_{..h}}{\sum_{j=1}^{N}\sum_{i=1}^{I}\left[\sum_{k=0}^{m} k^2 P_{jihk} - \left(\sum_{k=0}^{m} kP_{jihk}\right)^2\right]}. \tag{79}$$

Uma aproximação para o erro padrão assintótico das estimativas de cada um dos parâmetros θ, β e γ é dada pela raiz quadrada do inverso do denominador da equação de estimação respectiva.

Do mesmo modo, substituindo-se as derivadas de primeira e de segunda ordem da log-verossimilhança (eq. (72)) em relação às categorias da escala de classificação na equação de Newton Raphson (eq. (39)), obtém-se a equação de estimação para os passos cumulativos de dificuldade:

$$G'_k = G_k - \frac{\sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{h=1}^{H}\left(P_{jihk}\right) - x_k}{\sum_{j=1}^{N}\sum_{i=1}^{I}\sum_{h=1}^{H}\left(P_{jihk} - (P_{jihk})^2\right)}. \tag{80}$$

Aproximações para erro padrão de d_k são similares às dadas pelas equações (66) e (67).

3.4.5.2 Dados faltantes e pontuação perfeita

Uma propriedade bastante útil dessas equações de estimação é a não obrigatoriedade de que o conjunto de observações seja completo, isto é, podem faltar algumas observações desde que as observações presentes sejam suficientemente varidas e também suficientemente interligadas para os resultados calculados serem inequívocos.

O efeito causado pelas respostas em falta é que a diminuição das informações sobre os parâmetros aumenta os erros padrão. Kruskal (1960 *apud* LINACRE, 1994) sugere estimar os parâmetros duas vezes, uma delas, omitindo as respostas em falta, e a outra, considerando as respostas em falta como respostas erradas. A comparação entre as estimativas resultantes e as análises das estatísticas de ajuste (Seção 3.5.1) fornecem a orientação quanto à melhor alternativa a ser adotada.

Os dados em falta podem fazer com que os subconjuntos se tornem disjuntos. Nesse caso, o significado das medidas pode não ser claro.

Se a pontuação marginal correspondente a qualquer um dos parâmetros é perfeita, isto é, igual a zero ou o valor máximo possível, as equações de estimação correspondentes a esse parâmetro não serão calculadas, pois tais estimativas seriam infinitas. Desse modo, as informações correspondentes a essas observações perfeitas são omitidas a partir do conjunto de dados no processo de estimativa dos parâmetros.

3.4.5.3 A origem das subescalas

A escala no qual os parâmetros do modelo multifacetas de Rasch são estimados é única, isto significa que os parâmetros de cada elemento das facetas incluídas no modelo são estimados sobre uma única escala. Desse modo, é possível obter comparações entre os elementos de uma mesma faceta e também entre elementos de facetas distintas.

Entretanto, como as análises dos elementos de cada faceta separadamente são importantes no contexto de uma avaliação, é usual denominar a escala de cada faceta por subescala, entendendo-se neste caso, que cada uma delas está alocada em uma mesma escala.

Em algumas áreas aplicadas, como a psicologia ou as ciências da natureza, a localização da origem do sistema de medição, muitas vezes, é uma questão de conveniência para o pesquisador. Não existe um único local correto para a origem da escala, que deve ser escolhido de acordo com as espe-

cificidades de cada aplicação. O local de origem de uma escala, como, por exemplo, para medir a habilidade do examinando em algum construto, não é determinado pela ausência de capacidade. Pelo contrário, é um certo nível de habilidade que pode ser convenientemente representado pelo número zero.

Observando a equação (11), pode-se modificar a estimativa de algum dos parâmetros, desde que as estimativas dos outros parâmetros também sofram modificações que tornem as equações equivalentes. Essa é uma propriedade conveniente quando se pretende igualar as escalas do teste. Desse modo, o posicionamento real do examinando, do item, do avaliador e das escalas de avaliação utilizadas dentro do quadro comum de referência é arbitrário.

Na equação (11), quando são fixadas as origens das subescalas de todos os parâmetros, menos de um deles, a origem deste último é forçada para assumir uma posição única sobre a escala, que é linear, assim como os valores dos parâmetros, em relação às respectivas origens, são combinados de modo a calcular as probabilidades específicas.

Por convenção, os locais da origem de cada subescala são escolhidos como sendo a média dos valores calibrados para os itens, para os avaliadores e para a escala de classificação. O local de origem para a subescala da habilidade dos examinandos é definido unicamente pelo modelo. Em termos algébricos, isso significa que

$$\sum_{i=1}^{I} b_s = 0; \quad \sum_{h=1}^{H} c_h = 0; \quad d_0 \equiv 0 \quad \text{e} \quad \sum_{k=1}^{m} d_k = 0. \tag{81}$$

Nessas equações, I é o número de itens, H é o número de avaliadores, d_0 é a categoria mais baixa e m é o número de categorias.

Algumas vezes não é conveniente utilizar o modelo em sua forma padrão, sendo mais vantajoso utilizar as subescalas com diferentes origens. Uma utilização comum para a mudança da origem da escala é quando se deseja comparar as habilidades dos examinandos provenientes de testes diferentes, que possuem itens em comum. O local da origem da subescala dos itens deve ser proveniente da média da calibração dos itens que os testes têm em comum.

A redefinição da origem para qualquer uma das subescalas necessita de um ajuste das origens em todas as outras subescalas das facetas participantes dos modelos. Essa mudança da origem das subescalas é feita neste trabalho, para o modelo três facetas dado pela equação (11). Para tanto, seja o modelo multifacetas em sua forma exponencial dado pela equação (51) e pelas condições (81). Para denotar a presença de uma translação nas ori-

gens nas subescalas, é utilizado o símbolo “asterisco” na notação das facetas, isto é,

$$P_{jihk} = \frac{\exp\left[k\left(\theta_{j^*} - b_{i^*} - c_{h^*}\right) - \sum_{s^*=0}^{k} d_{s^*}\right]}{\sum_{r=0}^{m} \exp\left[r\left(\theta_{j^*} - b_{i^*} - c_{h^*}\right) - \sum_{s^*=0}^{r} d_{s^*}\right]}. \tag{82}$$

Os termos $\exp(-d_0)$ e $\exp(-d_{0^*})$ ocorrem para $k = 0$ tanto no numerador quanto no denominador das equações (51) e (82), respectivamente. Além disso, as probabilidades são independentes de valores que possam ser atribuídos a qualquer um desses termos, isso significa que

$$\begin{aligned} P_{jih0} &= \frac{\exp(-d_0)}{\sum_{r=0}^{m} \exp\left[r\left(\theta_j - b_i - c_h\right) - \sum_{s=0}^{r} d_s\right]} \\ &= \frac{\exp(-d_0^*)}{\sum_{r=0}^{m} \exp\left[r\left(\theta_j^* - b_i^* - c_h^*\right) - \sum_{s^*=0}^{r} d_{s^*}\right]}. \end{aligned} \tag{83}$$

Desse modo, os denominadores são iguais.

$$\sum_{r=1}^{m} \exp\left[r\left(\theta_j - b_i - c_h\right) - \sum_{s=1}^{r} d_s\right] = \sum_{r=1}^{m} \exp\left[r\left(\theta_j^* - b_i^* - c_h^*\right) - \sum_{s^*=1}^{r} d_{s^*}\right]. \tag{84}$$

As equações (51) e (82) também devem produzir o mesmo valor para a probabilidade P_{jihk} uma vez que os termos de d_0 são cancelados, os denominadores são iguais pela equação (84) e os demais termos dos numeradores também devem ser iguais, assim,

$$k\left(\theta_j - b_i - c_h\right) - \sum_{s=1}^{k} d_s = k\left(\theta_j^* - b_i^* - c_h^*\right) - \sum_{s^*=1}^{k} d_{s^*}. \tag{85}$$

Como $\sum_{s=1}^{k} d_s = 0$ para a categoria m e dividindo ambos os membros da última equação por m

$$\theta_j - b_i - c_h = \theta_j^* - b_i^* - c_h^* - \sum_{s^*=1}^{m} d_{s^*}/m. \tag{86}$$

Substituindo a equação (86) na equação (85) para a categoria $k = 1$

$$d_1 = d_{1^*} - \sum_{s=1}^{m} d_{s^*}/m. \tag{87}$$

Considerando-se todas as categorias da escala em ordem ascendente, o resultado obtido é

$$d_k = d_{k^*} - \sum_{s=1}^{m} d_{s^*}/m \tag{88}$$

sendo o valor de d_0 independente do valor de d_{0^*}, como foi discutido anteriormente.

Para se obter a translação da origem dos outros parâmetros, basta considerar que, por convenção, o local de origem nas subescalas é a média do parâmetro considerado. Assim, o local de origem da subescala da dificuldade do item é a média da dificuldade do item, assim, para $i = 1 \cdots I$

$$b_i = b_{i^*} - \sum_{s=1}^{I} b_{s^*}/I. \tag{89}$$

O local de origem da subescala da severidade dos avaliadores é a média da severidade dos avaliadores, assim, para $h = 1 \cdots H$

$$c_h = c_{h^*} - \sum_{s=1}^{H} c_{s^*}/H. \tag{90}$$

Pela equação (86), a translação da habilidade dos examinandos é dada por

$$\theta_j = \theta_{j^*} + \sum_{s=1}^{N} b_{s^*}/N + \sum_{s=1}^{H} c_{s^*}/H - \sum_{s=1}^{m} d_{s^*}/m. \tag{91}$$

Do mesmo modo, podem-se obter outros conjuntos de locais de origens, que resultaram em estimativas dos parâmetros equivalentes.

3.5 ANÁLISES DOS DADOS

Nesta seção são intruduzidos indicadores estatísticos que resumem as informações sobre a variabilidade dentro de cada faceta, no contexto do modelo multifaceta de Rasch. Estes indicadores estão subdivididos em quatro grupos: Estatísticas de ajuste, Estatísticas de separação, Médias justas e médias observadas.

3.5.1 Estatísticas de ajuste

As análises de ajuste dos dados aos modelos de Rasch são primordiais para o sucesso da aplicação. Segundo Linacre (1994), diferenças entre os valores obtidos e os esperados de acordo com o modelo não indicam falha por parte do modelo, no entanto podem indicar que aqueles dados não suportam a construção das medidas no intervalo determinado. Nas aplicações práticas dos modelos de Rasch, as estatísticas de ajuste devem ser obtidas para cada parâmetro, e estes devem ser inspecionados para assegurar o sucesso da análise. Esse autor afirma ainda que análises apenas no nível global podem ser enganosas, porque as diferenças individuais entre elementos dentro do conjunto de dados são tão ameaçadoras ao sucesso das análises quanto as diferenças globais.

Na tentativa de corrigir as discrepâncias, dependendo de sua natureza e da motivação para a análise, pode-se modificar alguma intenção do pesquisador quanto ao modelo utilizado ou a remoção de observações aberrantes. Por exemplo, pode-se modificar a forma de utilização das escalas de classificação por parte dos avaliadores. Além disso, quanto aos dados, podem-se remover ou alterar observações aberrantes, como as pontuações por avaliadores inconsistentes (LINACRE, 1994).

Desse modo, as estatísticas de ajuste são centrais para a avaliação da qualidade dos dados utilizados para a obtenção das medidas e devem ser calculadas para cada uma das facetas especificadas no modelo. Essas medidas são as médias quadráticas (*MQ*): *MQ–Infit*, *MQ–Outfit*; e as médias quadráticas padronizadas (*MQZ*): *MQZ–Infit* e *MQZ–Outfit*.

A média quadrática *Infit*, também denominada média quadrática ponderada, é baseada no quadrado dos resíduos padronizados entre os dados observados e o que seria esperado com base no modelo. A média quadrática *Outfit*, ou média quadrática não ponderada, também é baseada no quadrado dos resíduos padronizados entre os dados observados e os dados esperados,

mas o quadrado dos resíduos padronizados não são ponderados quando somados através das observações.

As estatísticas de ajuste para o avaliador, por exemplo, referem-se ao grau em que um determinado avaliador está associado com pontuações inesperadas, resumidas sobre os examinandos e os critérios (itens). A média quadrática *Infit* é sensível a padrões de pontuações inesperadas (*inlying*), enquanto a média quadrática *outfit* é sensível às pontuações inesperadas individuais (*outliers*), neste caso basta uma resposta suficientemente inesperada ou previsível para esta medida assumir valores extremos (altos ou baixos) (Linacre, 2002c; Myford; Wolfe, 2004).

As estatísticas de ajuste para o grupo de avaliadores e também para os examinandos são estabelecidas na sequência conforme o trabalho de Eckes (2011). As estatísticas de ajuste para as outras facetas incluídas nas análises são estabelecidas de modo análogo a essas.

3.5.1.1 Estatísticas de ajuste para os examinandos

As estatísticas de ajuste para a faceta examinandos referem-se à extensão com que as suas pontuações estão associadas com atribuições discrepantes por parte dos avaliadores.

Analogamente ao que foi feito para a obtenção das estatísticas de ajuste para os avaliadores, o ajuste média quadrática (MQ) para o examinando j é definido por meio da média dos resíduos padronizados ao quadrado dos avaliadores $h = 1, \cdots, H$, e dos itens $i = 1, \cdots, I$, dada para cada examinando:

$$MQ_U(j) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{I} z_{jih}^2}{H \cdot I}. \tag{92}$$

Essa equação fornece o ajuste estatístico média quadrática não ponderada para o examinando j, denotada neste trabalho por MQ–*Outfit*. A estatística de ajuste média quadrática ponderada para o examinando j é dada pela equação:

$$MQ_w(j) = \frac{\sum_{h=1}^{H}\sum_{i=1}^{I} w_{jih} z_{jih}^2}{\sum_{h=1}^{H}\sum_{i=1}^{I} w_{jih}} \tag{93}$$

no qual

$$w_{jih} = \sum_{k=0}^{m} (k - e_{jih})^2 P_{ihk}(\theta_j). \tag{94}$$

w_{jih} é a variância da observação em relação aos valores esperados pelas condições do modelo de Rasch. A equação (93) fornece o ajuste estatístico média quadrática ponderada para o examinando j, denotada neste trabalho por MQ–*Infit*.

3.5.1.2 Estatísticas de ajuste para os avaliadores

As estatísticas de ajuste para a faceta avaliador referem-se à extensão com que as pontuações provenientes de um determinado avaliador estão associadas com as respostas inesperadas dos examinandos.

Para a definição das estatísticas de ajuste, conforme estabelecidas por Eckes (2011), considera-se o modelo apresentado na equação (11). A probabilidade do examinando j de receber uma classificação k ($k = 0, \cdots, m$) em relação ao critério utilizado pelo avaliador h para o item i é dada por:

$$P_{ihk}(\theta_j) = \frac{\exp\left[k(\theta_j - b_i - c_h) - \sum_{s=0}^{k} d_s\right]}{\sum_{r=0}^{m} \exp\left[r(\theta_j - b_i - c_h) - \sum_{s=0}^{r} d_s\right]}. \tag{95}$$

Por definição $d_0 = 0$, o denominador é um fator de normatização baseado na soma dos elementos do numerador.

Geralmente, as estatísticas de ajuste indicam o grau em que as classificações observadas se aproximam das classificações esperadas, que são valores gerados pelo modelo MFR. Seja x_{jih} a classificação observada para o examinando j dada pelo avaliador h no item i e e_{jih} o valor esperado para essa classificação com base nas estimativas dos parâmetros pelo modelo de Rasch. As diferenças entre as classificações observadas e esperadas podem ser expressas em termos dos resíduos padronizados:

$$z_{jih} = \frac{x_{jih} - e_{jih}}{\sqrt{w_{jih}}} \tag{96}$$

$$e_{jih} = \sum_{k=0}^{m} kP_{ihk}(\theta_j) \tag{97}$$

w_{jih} é dada pela equação (94).

Valores de resíduos padronizados grandes para os avaliadores individualmente podem indicar a ocorrência de inconsistências em suas classificações. Resíduos padronizados com valores absolutos maiores que 2 podem indicar desvios significativos nos dados do modelo de Rasch. Esses valores podem ser usados para indicar quais das classificações dadas pelos avaliadores (observadas) são mais propensas a serem classificações surpreendentes ou inesperadas (MYFORD; WOLFE, 2004).

Quando os resíduos padronizados são elevados ao quadrado e os resíduos padronizados quadrados são resumidos sobre as diferentes facetas e elementos diferentes dentro de uma faceta, são obtidos índices de ajuste dos dados do modelo. Essas estatísticas sumárias são denominadas de estatísticas de ajuste média quadrática (ECKES, 2011).

Para obter o ajuste estatístico média quadrática (MQ) para o avaliador h, é utilizada a média dos resíduos padronizados ao quadrado dos examinandos $j = 1, \cdots, N$, e dos itens $i = 1, \cdots, I$, avaliado por cada avaliador:

$$MQ_U(h) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{I} z_{jih}^2}{N \cdot I}. \tag{98}$$

A equação (98) fornece o ajuste estatístico média quadrática não ponderada para o avaliador h. A estatística de ajuste não ponderada é também denominada de *Outfit*. As medidas *Outfit* para os avaliadores são particularmente sensíveis a eventuais classificações inesperadas de um avaliador.

A soma dos valores de z ao quadrado para todos os avaliadores pode ser vista como uma distribuição qui-quadrado com $H - 1$ graus de liberdade (H é o número de avaliadores), sob a hipótese nula de que os avaliadores estão classificando de forma consistente (MYFORD; WOLFE, 2000).

Menos sensível às classificações inesperadas periféricas é a estatística de ajuste média quadrática ponderada para o avaliador h, dada pela equação:

$$MQ_w(h) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{I} w_{jih} z_{jih}^2}{\sum_{j=1}^{N}\sum_{i=1}^{I} w_{jih}} \tag{99}$$

em que w_{jih} é definido na equação (94).

A estatística de ajuste ponderada dada na equação (99) é também denominada *Infit*. A estatística *Infit* para os avaliadores fornece uma estimativa da consistência com que cada avaliador em particular utiliza a escala de avaliação através dos examinandos e dos critérios, ou seja, essa estatística é sensível ao acúmulo de classificações inesperadas (*Infit* é a abreviação de "informação estatística ponderada fit"). Por essa razão, a estatística *Infit* é frequentemente considerada mais importante do que a estatística *Outfit* para a avaliação do ajuste do modelo (ECKES, 2011; LINACRE, 2002a; MYFORD; WOLFE, 2004).

3.5.1.3 Interpretação das estatísticas de ajuste

Quando os dados são provenientes de situações reais de teste, não há um ajuste perfeito dos dados ao modelo de Rasch para a construção da medida. Desse modo, deve-se estipular o "tamanho" das diferenças que podem ser tolerados. Tradicionalmente, os testes de significância são utilizados para a tomada de decisões quanto a esses ajustes, mas esses testes são fortemente influenciados pelo tamanho da amostra. Apenas algumas exceções ocorridas para um conjunto de dados suficientemente grande poderia provocar a rejeição do modelo proposto ou dos dados. Nesses casos, é necessário determinar uma medida quantitativa do tamanho da discrepância entre o modelo estatístico e o conjunto de dados observados (WRIGHT; LINACRE, 1994; GUSTAFSON, 1980).

Segundo Wright e Linacre (1994), os valores esperados para as medidas médias quadráticas *MQ–Infit* e *MQ–Outfit* são próximos de 1, embora tais medidas estejam definidas no intervalo $(0,\infty)$. Valores de média quadrática maiores do que 1 indicam que os resultados são menos previsíveis do que o modelo de Rasch prevê. Já os valores de média quadrática menores do que 1 indicam que os resultados são mais previsíveis do que o modelo de Rasch prevê.

Para Myford e Wolfe (2004), quando se trata da variabilidade entre as

pontuações, os valores de *MQ* maiores do que 1 indicam haver maior variabilidade entre as pontuações atribuídas aos examinandos. Para valores de *MQ* menores do que 1, a pontuação atribuída pelos avaliadores é semelhante quanto aos graus de severidade, indicando pouca variação no padrão das classificações dos examinandos. Geralmente, valores de *MQ* maiores do que 1 são mais problemáticos do que valores dessa medida menores do que 1.

Linacre (2002c; 2014a) sugere como limite de controle para os valores das médias quadráticas o intervalo entre 0,50 e 1,50 e denomina os valores dessas médias nessa faixa como "produtivos para a medição". Wright e Linacre (1994) propõem valores para as estatísticas de ajuste no intervalo 0,8 e 1,2, mas alertam que esses limites de controle dependerão, em parte, da natureza e do propósito de cada avaliação em particular. Para esses pesquisadores não existem regras rígidas para o estabelecimento de limites superiores e inferiores para as estatísticas medidas *MQ*, no entanto alguns intervalos podem ser considerados razoáveis para essas medidas de ajuste.

Os índices de ajuste indicam o grau em que os dados observados estão de acordo com os dados esperados de acordo com o modelo utilizado. As grandes diferenças entre os dados observados e esperados, expressas como resíduos padronizados, indicam discrepâncias. De acordo com Linacre (2014), os resultados são satisfatórios quando cerca de 5% ou menos dos resíduos padronizados, em valores absolutos, são iguais ou superiores a 2, e cerca de 1% ou menos dos resíduos padronizados, em valores absolutos, são iguais ou superiores a 3. Uma interpretação das estatísticas *MQ* é dada por Write e Linacre (1994) no Quadro 13.

Quadro 13 – Interpretação das estatísticas de ajuste: Média quadrática

Média quadrática	Interpretação
>2,0	O sistema de medição está distorcido ou degradado.
1,5 – 2,0	Improdutivo para a construção da medida, mas não é degradante.
0,5 – 1,5	Produtivo para a medida.
<0,5	Menos produtivo para a construção da medida, mas não degradante. Pode produzir enganosamente boa confiabilidade e comparações.

Adaptado de: Wright e Linacre (1994)

3.5.2 Estatísticas de separação

Em avaliações, para se medir a dificuldade de itens ou a habilidade de pessoas, é necessário que seja possível a comparação entre os itens para a localização deles em uma escala. Os itens são localizados nessa escala de acordo com o grau de suas dificuldades. As pessoas são localizadas de acordo com o número de itens que foram capazes de responder corretamente. Os avaliadores são localizados na escala de acordo com o grau de severidade com que atribuíram pontuações para os itens e também para a habilidade das pessoas (MYFORD e WOFE, 2004).

Os itens localizados à esquerda na escala são mais fáceis do que aqueles localizados à direita; as pessoas localizadas à esquerda têm menor capacidade do que aquelas localizadas à direita na escala de habilidades, enquanto os avaliadores localizados à esquerda são mais complacentes quando comparados aos localizados mais à direita na escala, que são mais severos. É necessário localizar as pessoas e os itens ao longo da escala com precisão suficiente para ter respostas conclusivas. Por exemplo, se os itens (ou as pessoas) se encontram muito próximos uns dos outros ao longo da escala, pode não ser possível uma medição útil, pois a diferença entre os graus de dificuldade dos itens pode não ser suficiente para distinguir as pessoas quanto às suas habilidades. Entretanto, separação entre os itens (ou as pessoas) muito grande geralmente significa lacunas entre a dificuldade dos itens e a habilidade das pessoas, o que resulta em medidas imprecisas (WRIGHT; STONE, 1999).

Na verdade, a seleção de itens de um teste deve propiciar diferenciações relevantes entre o desempenho de pessoas distintas. A localização dos itens consiste na definição operacional da variável latente de interesse, enquanto a localização das pessoas é o resultado da aplicação da variável latente para a medição.

As estatísticas de separação das pessoas indica o quanto um conjunto de itens é capaz de separar entre as habilidades das pessoas que estão sendo medidas. As estatísticas de separação dos itens indicam o quanto uma amostra de pessoas é capaz de separar os itens utilizados no teste quanto às suas dificuldades.

As estatísticas de separação são calculadas para cada faceta especificada no modelo. Os valores dessas estatísticas variam entre 0 e 1, quanto mais próximas de 1 melhor a separação existente e mais precisa será a medição (WRIGHT; STONE, 1999).

Eckes (2011) estabelece quatro estatística de separação com foco na

faceta avaliadores. Na sequência, com base no trabalho desse autor, são definidas essas estatísticas e também são feitas extensões dessas estatísticas para a faceta examinandos. As estatísticas de separação para as outras facetas são estabelecidas de modo análogo.

3.5.2.1 Estatísticas de separação para os examinandos

As estatísticas de separação dos examinandos indicam o quanto um conjunto de itens é capaz de separar as habilidades das pessoas que estão sendo medidas.

A primeira estatística, denominada *índice de homogeneidade do examinando*, é um teste da hipótese nula: as medidas da habilidade dos examinandos na população são as mesmas para todos eles. Essa estatística é:

$$Q_J = \sum_{j=1}^{J} w_j(\hat{\theta}_j - \hat{\theta}_+)^2 \tag{100}$$

onde

$$\hat{\theta}_+ = \frac{\sum_{j=1}^{J} w_j \hat{\theta}_j}{\sum_{j=1}^{J} w_j} \qquad \text{e} \qquad w_j = \frac{1}{SE_j^2} \tag{101}$$

$\hat{\theta}_j$ é a estimativa do parâmetro habilidade do examinando j, e SE_j refere-se ao erro padrão associado com a estimativa do parâmetro habilidade do examinando j.

A estatística *taxa de separação dos examinandos* dá a propagação das medidas da habilidade dos examinandos em relação à precisão dessas medidas. Esse índice de separação é expresso como uma razão entre o desvio padrão "verdadeiro" das medidas da habilidade do examinando em relação à média de erro padrão da habilidade do examinando. Quanto mais próximo o valor dessa medida estiver de zero, mais semelhantes são as medidas das habilidades dos examinandos.

Para a definição da taxa de separação dos examinandos, é necessário, primeiramente, que seja definida a variância "verdadeira" da medida da habilidade do examinando:

$$SD_v^2(J) = SD_o^2(J) - MQE_J \tag{102}$$

$SD_o^2(H)$ é a variância observada da habilidade do avaliador, e MQE_J é a "média quadrática do erro das medidas", isto é, a média da variância das medidas dos examinandos:

$$MQE_J = \frac{\sum_{j=1}^{N} SE_j^2}{N} \tag{103}$$

assim, a variância "verdadeira" das medidas da habilidade dos examinandos é a variância obsevada dessas medidas ajustadas pelo erro das medidas. A razão entre a variação ajustada e a variância média quadrática do erro leva a:

$$G_J^2 = \frac{SD_v^2(J)}{MQE_J} \tag{104}$$

a raiz quadrada dos membros da equação (104) resulta no índice que é a razão de separação do examinando:

$$G_J = \frac{SD_v(J)}{\sqrt{MQE_J}}. \tag{105}$$

A estatística G_J indica a propagação das medidas da habilidade do examinando na unidade dos erros de medidas. Quanto maior for o valor de G_J, mais espalhados estão os examinandos na escala de classificação quanto ao traço latente.

Utilizando a razão de separação do examinando, pode-se calcular o índice de separação do examinando, que é o número de níveis estatisticamente diferentes de habilidade numa determinada amostra de indivíduos. Esses níveis são determinados por, pelo menos, três unidades de erro de medição. O índice de separação do examinando, também denominado *índice estrato*, é dado por:

$$J_J = \frac{4SD_v(J) + \sqrt{MQE_J}}{3\sqrt{MQE_J}} = \frac{4G_J + 1}{3}. \tag{106}$$

O valor do índice J_J fornece o número de grupos estatisticamente diferentes, no qual todo o grupo de examinandos é subdividido.

A *confiabilidade do índice de separação* fornece informações sobre a forma como os elementos são separados dentro do grupo. É calculada como

a razão da variância verdadeira das medidas da habilidade dos examinandos pela variância observada dessas medidas:

$$R_J = \frac{SD_v^2(J)}{SD_o^2(J)} = \frac{G_J^2}{1+G_J^2}. \tag{107}$$

R_J representa a proporção da variância das medidas da habilidade dos examinandos que não são provenientes de erros de medição. Essa medida fornece o quanto a habilidade dos examinandos do grupo é diferente. Valores de R_J próximos de zero indicam que os examinandos do grupo possuem habilidades semelhantes, enquanto valores próximos de 1 sugerem que os examinandos possuem graus de habilidade muito diferentes.

3.5.2.2 Estatísticas de separação para os avaliadores

A primeira estatística, denominada *índice de homogeneidade do avaliador*, fornece um teste da hipótese nula, na qual as medidas da severidade dos avaliadores na população são as mesmas para todos os avaliadores. Essa estatística é:

$$Q_H = \sum_{h=1}^{H} w_h(\hat{c}_h - \hat{c}_+)^2 \tag{108}$$

onde

$$\hat{c}_+ = \frac{\sum_{h=1}^{H} w_h \hat{c}_h}{\sum_{h=1}^{H} w_h} \qquad \text{e} \qquad w_h = \frac{1}{SE_h^2} \tag{109}$$

$\hat{c}_h$ é a estimativa do parâmetro severidade do avaliador h, e SE_h refere-se ao erro padrão associado com a estimativa do parâmetro severidade do avaliador h.

Os índices Q_h são distribuídos aproximadamente como a estatística qui-quadrado com $H-1$ graus de liberdade. Na prática, um valor significante de Q_h para um determinado grupo de avaliadores indica que as medidas de severidade de pelo menos dois dos avaliadores são significantemente diferentes (MYFORD; WOLFE, 2004). Note-se que Q_h é muito sensível ao número de avaliadores do grupo. Para grupos grandes, esse índice pode atin-

gir um nível de significância, apesar de as diferenças reais entre a severidade dos avaliadores serem pequenas.

Outra estatística de separação é a *taxa de separação dos avaliadores*. Essa estatística dá a propagação das medidas da severidade dos avaliadores em relação à precisão dessas medidas. Isso significa que quanto mais próximo o valor dessa medida estiver de zero, mais semelhantes são as medidas da severidade dos avaliadores. Especificamente, o índice de separação do avaliador é expresso como uma razão entre o desvio padrão "verdadeiro" das medidas da severidade do avaliador (isto é, o desvio padrão ajustado para o erro de medição) em relação à média de erro padrão da severidade do avaliador.

Para a definição da taxa de separação dos avaliadores, é necessário, primeiramente, que seja definida a variância "verdadeira" da medida da severidade do avaliador:

$$SD_v^2(H) = SD_o^2(H) - MQE_H \tag{110}$$

$SD_o^2(H)$ é a variância observada da severidade do avaliador, e MQE_H é a "média quadrática do erro das medidas", isto é, a média da variância das medidas dos avaliadores:

$$MQE_H = \frac{\sum_{h=1}^{H} SE_h^2}{H} \tag{111}$$

assim, a variância "verdadeira" das medidas da severidade dos avaliadores é a variância obsevada dessas medidas ajustadas pelo erro das medidas. A razão entre a variação ajustada e a variância média quadrática do erro resulta em:

$$G_H^2 = \frac{SD_v^2(H)}{MQE_H} \tag{112}$$

a raiz quadrada dos membros da equação (112) resulta no índice que é a razão de separação do avaliador:

$$G_H = \frac{SD_v(H)}{\sqrt{MQE_H}}. \tag{113}$$

A estatística G_H indica a propagação das medidas da severidade do avaliador na unidade dos erros de medidas. Quanto maior for o valor de G_H, mais espalhados estão os avaliadores na escala de severidade.

Utilizando a razão de separação do avaliador, pode-se calcular o *ín-*

dice de separação do avaliador, que é o número de níveis estatisticamente diferentes de severidade dos avaliadores numa determinada amostra de avaliadores. Esses níveis são determinados por, pelo menos, três unidades de erro de medição. O índice de separação do avaliador, também denominado *índice estrato* (camadas), é dado por:

$$J_H = \frac{4SD_v(H) + \sqrt{MQE_H}}{3\sqrt{MQE_H}} = \frac{4G_H + 1}{3}. \tag{114}$$

O valor do índice J_H fornece o número de grupos estatisticamente diferentes, no qual todo o grupo de avaliadores é subdividido. Por exemplo, um índice de separação próximo de 1 indicaria que todos os avaliadores estão pontuando de forma semelhante em relação à severidade.

A *confiabilidade do índice de separação* fornece informações sobre a forma como os elementos são separados dentro das facetas e pode ser calculada como a razão da variância verdadeira das medidas da severidade dos avaliadores pela variância observada dessas medidas:

$$R_H = \frac{SD_v^2(H)}{SD_o^2(H)} = \frac{G_H^2}{1 + G_H^2}. \tag{115}$$

R_H representa a proporção da variância das medidas da severidade dos avaliadores que não são provenientes de erros de medição. Essa medida fornece o quanto a severidade dos avaliadores do grupo é diferente. Valores de R_H próximos de zero indicam que os avaliadores do grupo estão pontuando de modo semelhante, enquanto valores próximos de 1 sugerem que os avaliadores estão pontuando com graus de severidade muito diferentes.

3.5.3 Médias justas e médias observadas

As médias justas e as médias observadas auxiliam na obtenção de uma interpretação entre as diferenças nas medidas das facetas e suas implicações. Essas medidas são estabelecidas na sequência conforme o trabalho de Eckes (2011) para as facetas avaliadores e examinandos. As equações para as outras facetas são análogas a essas.

3.5.3.1 Médias justas e observadas para os examinandos

De maneira análoga ao cálculo das médias justas para os avaliadores, as médias justas para os examinandos procuram compensar as diferenças entre a severidade dos avaliadores. Ou seja, para cada examinando, existe uma classificação esperada que seria obtida a partir de um avaliador com um nível médio de severidade. Entre os avaliadores, o grupo de referência para calcular esse nível médio de severidade é o grupo formado por todos os avaliadores incluídos na análise.

A média observada para o escore de cada examinando é obtida da classificação média que o examinando receberia por todos os avaliadores em todos os itens envolvidos na obtenção de cada classificação:

$$M_O(j) = \frac{\sum_{i=1}^{I}\sum_{h=1}^{H} x_{jih}}{I \cdot H}. \tag{116}$$

Para calcular a média justa, para o examinando j, as estimativas dos parâmetros de todos os elementos das outras facetas que participaram das análises, exceto para o parâmetro de proficiência do examinando, são definidas pelos seus valores médios. Para o modelo de três facetas utilizado neste experimento, a equação é:

$$\ln\left[\frac{P_{jk}}{P_{j(k-1)}}\right] = \theta_j - b_M - c_M - d_k \tag{117}$$

onde P_{jk} é a probabilidade de o examinando j de receber uma classificação na categoria k, $k = 0, \cdots, m$, de todos os avaliadores, em todos os itens; b_M e c_M são os valores da dificuldade média e da severidade média dos avaliadores, respectivamente. A média justa (ou pontuação esperada) para os examinando é dada por:

$$M_F(j) = \sum_{k=0}^{m} kP_{jk}. \tag{118}$$

De maneira análoga, são estabelecidas as fórmulas das médias justas e observadas das outras facetas incluídas na análise.

3.5.3.2 Médias justas e observadas para os avaliadores

A média observada para o avaliador h, dada por $M_O(h)$, é a média na qual esse avaliador pontuou na avaliação de todas as tarefas e de todos os examinandos participantes da avaliação:

$$M_O(h) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{I} x_{jih}}{N \cdot I} \tag{119}$$

onde x_{jih} é o valor observado para o examinando j no item i e atribuído pelo avaliador h.

Segundo Eckes (2011), quando se trata de médias observadas, é normal confundir a severidade do avaliador e a proficiência dos examinandos. Por exemplo, a média observada de um determinado avaliador é significativamente menor do que as médias observadas dos outros avaliadores. A ocorrência desse fato não possui um único motivo, então, pelo menos duas conclusões podem ser obtidas: a) o avaliador é mais severo do que os outros avaliadores, b) o avaliador pontuou um grupo de examinandos com menor habilidade.

A média justa é inserida no contexto das avaliações com itens de respostas construídas para resolver esse problema. Essa média para o avaliador h ajusta a média observada $M_O(h)$ 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.

Para calcular uma média justa para o avaliador h, as estimativas dos parâmetros de todos os elementos das outras facetas que participaram na produção dos escores, exceto para o parâmetro severidade do avaliador, são definidas como seus valores médios. A equação para o modelo de três facetas é:

$$\ln\left[\frac{P_{hk}}{P_{h(k-1)}}\right] = \theta_M - b_M - c_h - d_k. \tag{120}$$

Nessa fórmula, P_{hk} é a probabilidade de o avaliador h usar a categoria k, $k = 0, \cdots, m$, para todos os examinandos e para todos os itens; θ_M e b_M são as médias das medidas da habilidade do examinando e da dificuldade do item, respectivamente.

A média justa para o avaliador h, $M_J(h)$ é dada por:

$$M_J(h) = \sum_{k=0}^{m} kP_{hk}. \tag{121}$$

As médias justas permitem comparações mais justas entre a severidade dos avaliadores e o desempenho dos examinandos ao executarem a tarefa determinada.

3.6 ANÁLISES PARA A VALIDADE

Nesta seção são apontadas análises que devem ser feitas para aferir a qualidade de uma avaliação com itens de respostas construídas no contexto do modelo multifacetas de Rasch. Essas análises são baseadas no modelo multifacetas de Rasch de três facetas, sendo elas, a habilidade dos examinandos, a dificuldade dos itens e a severidade dos avaliadores, e são feitas, para os elementos das facetas, tanto no nível individual quanto no nível de grupo.

O Quadro 14 consiste em um *guia resumo* dessas análises, entretanto outras análises e procedimentos poderão ser adotados dependendo dos objetivos de cada avaliação em particular.

Quadro 14 – Análises para a validade no contexto do modelo multifacetas de Rasch

1. Análise do ajuste global dos dados ao modelo multifacetas de Rasch.
2. Análise visual do mapa das variáveis.
3. Resumo das estatísticas.
3. Interpretação da qualidade da escala de classificação.
4. Análises dos elementos da faceta Itens.
5. Análises dos elementos da faceta Examinandos.
6. Análises dos elementos da faceta Avaliadores.
7. Conclusão sobre o padrão de qualidade da avaliação.

Fonte: Autora

3.6.1 Ajuste global dos dados ao modelo multifacetas de Rasch

Por meio das respostas inesperadas, pode-se analisar o ajuste do modelo de acordo com as hipóteses do modelo. Segundo Linacre (2014a), os resultados da avaliação são satisfatórios quando cerca de 5% ou menos dos resíduos padronizados, em valores absolutos, são iguais ou superiores a 2 e cerca de 1% ou menos dos resíduos padronizados, em valores absolutos, são iguais ou superiores a 3.

3.6.2 Análise visual do mapa das variáveis

O mapa das variáveis é um recurso muito informativo para auxiliar na interpretação dos resultados da avaliação de modo geral, uma vez que esse mapa retrata todas as facetas da análise em um único quadro de referência. Esse recurso é de grande valia para facilitar comparações dentro e entre as várias facetas. Pode-se perceber, por exemplo, se algum elemento da avaliação apresenta um comportamento, em média, diferente do comportamento dos outros elementos do grupo. Nesse mapa, é possível também analisar, à primeira vista, se os avaliadores utilizaram as categorias de classificação como foi estabelecido originalmente na elaboração do teste.

3.6.3 Resumo das estatísticas

Por meio de um resumo dos principais índices, pode-se ter uma visão geral dos resultados da avaliação no contexto do modelo multifacetas de Rasch. Desse modo, para cada uma das facetas, podem-se obter os valores médios das estimativas e a precisão com que eles foram calculados, as difereça entre os elementos por meio dos índices de separação, o ajuste dos dados ao modelo por meio das estatísticas de ajuste e outras análises para o entendimento, de modo geral, da qualidade da avaliação.

3.6.4 Análises dos elementos da faceta Examinandos

Primeiramente devem-se analisar os valores das medidas média quadrática *MQ–Infit* e *MQ–Outfit*. A expectativa dessas medidas é para valores próximos de 1. Valores de média quadrática maiores do que 1 indicam que os

resultados são menos previsíveis do que o modelo de Rasch prevê, enquanto valores de média quadrática menores do que 1 indicam que os resultados são mais previsíveis do que o modelo de Rasch prevê.

Wright e Linacre (1994) consideram que valores para a média quadrática no intervalo entre 0,5 e 1,5 são produtivos para a medida, valores que estão fora desse intervalo devem ser analisados individualmente.

Linacre (2014a) recomenda uma análise da tabela de respostas inesperadas, que faz parte do contexto do modelo multifacetas de Rasch, uma vez que os valores grandes das estatísticas de ajuste podem corresponder a essas respostas, auxiliando na avaliação da consistência dos dados e dos processos envolvidos no teste. Por meio desses dados, é possível obter algumas informações sobre os elementos da avaliação, inclusive sobre os examinandos.

Valores do residual padronizado, em valores absolutos, muito acima de 1 podem significar que os indivíduos portadores desses índices se saíram de modo diferente do que era esperado pelo ajuste do modelo. Residual padronizado muito grande e positivo indica que o indivíduo se saiu melhor do que era esperado, ou seja, melhor do que a sua capacidade permite, indicando uma resposta ao acaso ou cópia. Valor alto e negativo indica que o indivíduo se saiu pior do que era esperado. Nesse caso é necessária uma análise cuidadosa, pois o problema pode ter sido causado pelo item, pelo avaliador ou mesmo por algum problema externo ao instrumento ou à avaliação (LINACRE, 2014a). Outras análises podem ser feitas no nível individual, para cada examinando em particular, dependendo dos objetivos da avaliação.

3.6.5 Análises dos elementos da faceta Avaliadores

As avaliações que necessitam do julgamento de avaliadores, especialmente as avaliações com itens abertos, possuem algumas questões consideradas críticas. Entre essas questões está a diferença entre a maneira com que os diversos avaliadores da equipe de correção dos testes pontuam as tarefas. A experiência dos avaliadores, a utilização de critérios de pontuação bem estabelecidos, o treinamento dos avaliadores, entre outros procedimentos, são tidos como fatores importantes na obtenção de bons índices de confiabilidade, mas também é necessário levar em conta as tendências dos avaliadores em julgamentos sistemáticos dos desempenhos avaliados e que causam variabilidade na pontuação dos examinandos.

Esse tipo de variabilidade na pontuação é geralmente associado com características dos avaliadores e não com o desempenho de examinandos. Isso significa que a variabilidade causada pelo avaliador pode consistir em

uma fonte importante de variância construto-irrelevante na pontuação das tarefas elaboradas pelos examinandos, prejudicando a medida do construto que o teste deve medir e, desse modo, ameaçando a validade e a imparcialidade da avaliação (ECKES, 2011; McNAMARA, 2000; MESSICK, 1989).

As tendências dos avaliadores em pontuações sistemáticas, que podem causar uma gama de diferentes tipos de erros nas classificações dos examinandos, são frequentemente abordadas nas pesquisas, pois identificar e determinar esses erros torna-se importante para assegurar a validade da avaliação.

Os efeitos mais discutidos causados por essas tendências dos avaliadores são:

1. ***Efeito da severidade***, que é a tendência dos avaliadores em avaliar de maneira muito exigente as tarefas elaboradas pelos examinandos. Os avaliadores portadores dessa tendência atribuem pontuações que são, em média, inferiores às pontuações atribuídas pelos outros avaliadores do grupo. Desse modo, os avaliadores severos subestimam o nível de desempenho do examinando em toda a escala de habilidades.

2. ***Efeito da complacência***. Análogo ao efeito da severidade, o efeito da complacência é tradicionalmente definido como a tendência do avaliador em atribuir pontuações que são, em média, mais elevadas do que as pontuações atribuídas pelos outros avaliadores do grupo. Estes avaliadores possuem a tendência em superestimar o nível de desempenho dos examinandos em toda a escala de habilidades.

3. ***Efeito de tendência central***, que é a tendência excessiva dos avaliadores de classificações iguais ou perto do ponto médio da escala, evitando, desse modo, classificações nos extremos da escala. A tendência central pode apresentar-se de formas diferentes. Em alguns casos, o avaliador pode ser capaz de avaliar com precisão examinandos cujos níveis de desempenho se encontram nos extremos da escala de habilidades, no entanto ele é incapaz de utilizar as categorias do meio da escala de forma consistente para diferenciar entre os desempenhos médios dos examinandos. Outras vezes, a tendência central pode manifestar-se como a incapacidade do avaliador em fazer distinções entre qualquer uma das categorias da escala e, assim, atribui pontuações semelhantes no meio da escala.

   Se muitos avaliadores da equipe são portadores dessa tendência, o problema pode estar relacionado com os critérios ou com a escala de pontuação, não com os avaliadores. Isso pode ocorrer se a escala de classificação possui muitas categorias, exigindo distinções minuciosas. Neste

caso, seria indicado revisão da escala utilizada, diminuindo o número de categorias para que as distinções entre os níveis fiquem mais evidentes.

4. ***Efeito de aleatoriedade***, que é definido como a tendência do avaliador em aplicar uma ou mais categorias da escala de maneira inconsistente com o modo com que os outros avaliadores aplicam a mesma escala. O avaliador que possui essa tendência é demasiadamente inconsistente no uso da escala, apresentando maior variabilidade aleatória do que o esperado na avaliação. Esse avaliador pode ter desenvolvido uma interpretação diferente do significado de uma ou mais categorias da escala, utilizando-as de forma diferente dos outros avaliadores da equipe. Em alguns casos, o avaliador pode não ter formação suficiente para ser capaz de fazer discriminações minuciosas e atribui as pontuações de forma aleatória e não confiável.

5. ***Efeito halo***, que é definido como a tendência dos avaliadores em atribuir pontuações semelhantes para todos os examinandos para o mesmo item. Isto é, os examinandos recebem pontuações semelhantes mesmo que os seus desempenhos tenham sido muito diferentes. Desse modo, diferentes desempenhos podem obter a mesma pontuação.

6. ***Efeito de viés***, que é comumente denominado efeito de severidade/complacência diferencial. Quando a amostra de examinandos é composta por grupos distintos separados por sexo, idade, cor, raça, escola, região, entre outros, o avaliador pode ter a tendência em agir de forma discriminatória, pontuando diferentemente os grupos que fazem parte da avaliação. O efeito de severidade diferencial do avaliador é definido como a tendência em atribuir pontuações a um determinado grupo, em média, menores do que as pontuações atribuídas pelos outros avaliadores a esse grupo. Analogamente, a tendência de efeito de complacência diferencial é definida como a tendência do avaliador em atribuir pontuações, em média, maiores a um grupo do que as pontuações atribuídas pelos outros avaliadores da equipe a esse grupo. Em ambos os casos, o avaliador mostra um comportamento discriminatório entre grupos participantes da avaliação causando viés nas avaliações desses grupos.

Esses e outros detalhes sobre essas tendências dos avaliadores podem ser conferidos nos trabalhos de Knock, Read e Randow (2007), Myford e Wolfe (2000, 2004) e Engelhard e Myford (2003).

Os Quadros de 15 a 18 fornecem resumos dos indicadores estatísticos para diagnóstico de quatro desses efeitos causados por tendências dos

avaliadores em pontuações sistemáticas, tanto no nível de grupo quanto no nível individual no contexto do modelo multifacetas de Rasch e das estatísticas citadas nas Seções 3.5.1, 3.5.2 e 3.5.3. Esses resumos foram elaborados baseando-se no trabalho de Myford e Wolfe (2004).

Quadro 15 – Estatísticas indicativas dos efeitos de severidade e complacência dos avaliadores

Indicadores no nível de grupo	**Diagnóstico**
Contagem de frequência do uso de cada uma das categorias.	Verificar se há uso excessivo das categorias dos extremos da escala.
Teste qui-quadrado fixo para os avaliadores*.	Se o teste for estatisticamente significante (p<0,05), a hipótese é falsa. Os avaliadores possuem níveis de severidade diferentes.
Taxa de separação dos avaliadores.	Quanto maior for essa medida, mais dispersos estão os avaliadores.
Índice de separação dos avaliadores (estrato).	Quando alto, indica níveis diferentes de severidade, se for igual a 1, todos compartilham da mesma medida de severidade.
Índice de confiabilidade dos avaliadores.	Valores entre 0 e 1, quanto mais perto de 1 mais significativas são as diferenças entre a severidade dos avaliadores.
Indicadores no nível individual	**Diagnóstico**
Distribuição das medidas da severidade dos avaliadores.	Procurar por avaliadores isolados do grupo de avaliadores no mapa de variáveis.
Medidas da severidade do avaliador.	Verificar se há avaliador com medida de severidade muito diferente da média das medidas dos avaliadores.
Medidas médias justas dos avaliadores.	Comparar a média justa do avaliador mais severo/complacente com a média justa de um avaliador padrão.
Contagem de frequência da utilização de cada uma das categorias por cada um dos avaliadores **.	Verificar se há avaliador utilizando excessivamente as categorias dos extremos da escala.

* hipótese: todos os avaliadores possuem o mesmo nível de severidade após correção do erro.
** Disponível para o modelo de crédito parcial.

Fonte: Autora

Quadro 16 – Estatísticas indicativas do efeito de tendência central dos avaliadores

Indicadores no nível de grupo	**Diagnóstico**
Contagem de frequência da utilização de cada uma das categorias.	Verificar se há uso excessivo das categorias centrais da escala.
Teste qui-quadrado fixo para os examinandos*.	Se o teste for estatisticamente significante (p<0,05), a hipótese é falsa. Não existe um efeito de tendência central no nível do grupo.
Taxa de separação dos examinandos.	Se pequeno, sugere efeito de tendência central para o grupo.
Índice de separação dos examinandos (estrato).	Se pequeno, sugere efeito de tendência central para o grupo.
Índice de confiabilidade dos examinandos.	Se pequeno (≈ 0), sugere efeito de tendência central para o grupo.
Indicadores no nível individual	**Diagnóstico**
Análise dos índices de ajuste média quadrática.	Analisar as pontuações observadas para os avaliadores com *MQ* fora do intervalo entre 0,5 e 1,5.
Tabela de valores inesperados.	Verificar se os valores observados do avaliador são mais próximos do centro da escala do que os valores esperados.
Contagem de frequência da utilização de cada categoria por cada um dos avaliadores**.	Verificar se há avaliador utilizando excessivamente as categorias do centro da escala.
Índices média quadrática *outfit* para as categoria da escala de classificação diferentes de 1,0**.	Verificar se a diferença entre as médias observadas e esperadas do avaliador é significativa.
Limiares das categorias da escala de classificação para cada avaliador**.	Verificar se há limiares: dispersos, em menor número, com ordens invertidas.
Curvas de probabilidade para cada avaliador**.	Os limiares das categorias, especialmente das do meio da escala, apresentam grande separação.

* hipótese: todas as pessoas possuem o mesmo nível de severidade após correção do erro.

** Disponível para o modelo de crédito parcial.

Fonte: Autora

Quadro 17 – Estatísticas indicativas do efeito de aleatoriedade dos avaliadores

Indicadores no nível de grupo	**Diagnóstico**
Teste qui-quadrado fixo para os examinandos*.	Se o teste for estatisticamente significante (p<0,05), a hipótese é falsa. Não há evidência de efeito de aleatoriedade.
Taxa de separação para os examinandos.	Uma taxa de separação baixa para os examinandos sugere um efeito de aleatoriedade no nível de grupo.
Índice de separação dos examinandos (estrato).	Um índice de separação baixo para os examinandos sugere um efeito aleatoriedade.
Índice de confiabilidade dos examinandos.	Valores baixos desse índice sugerem um efeito de aleatoriedade.
Indicadores no nível individual	**Diagnóstico**
Índices de ajuste média quadrática para os avaliadores.	Medidas *infit* e *outfit* significativamente maiores do que 1 podem ser indício de efeito de aleatoriedade para o avaliador.
Coeficiente de correlação ponto bisserial.	Medidas das correlações ponto bisserial inferiores às correlações dos outros avaliadores sugerem tendência de aleatoriedade.

* hipótese: todas as pessoas possuem o mesmo nível de severidade após correção do erro.

Fonte: Autora

Quadro 18 – Estatísticas indicativas do efeito de halo dos avaliadores

Indicadores no nível de grupo	**Diagnóstico**
Teste qui-quadrado fixo para os itens*.	Teste estatisticamente significante (p<0,05) significa que os itens são significativamente diferentes em termos de suas dificuldades e não há evidência de efeito de halo.
Taxa de separação para os itens.	Uma taxa de separação baixa para os itens sugere um efeito de halo no nível de grupo.
Índice de separação dos itens (estrato).	Um índice de separação baixo para os itens sugere um efeito halo no nível de grupo.
Índice de confiabilidade dos itens.	Valores desse índice baixos sugerem um efeito de halo.

Continua

Continuação

Indicadores no nível individual	Diagnóstico
Índices de ajuste média quadrática para os avaliadores.	Medidas *infit* e *outfit* fora do intervalo entre 0,5 e 1,5 podem ser indício de efeito de halo para o avaliador portador dessas medidas.
Valores médios observados e esperados.	Quando os valores observados e esperados do avaliador em questão são significativamente diferentes uns dos outros, pode haver evidência de efeito de halo.
Análises dos avaliadores com resultados de |t-Student|>2**.	Comparar as pontuações observadas e esperadas para os avaliadores em busca de vieses entre os avaliadores e as categorias.

* hipótese: todos os itens possuem o mesmo nível de severidade após correção do erro.

** Disponível para o modelo de crédito parcial.

Fonte: Autora

3.6.6 Análises dos elementos da faceta Itens

Devem-se verificar as medidas, em *logitos*, do grau de dificuldade dos itens e a localização de cada um deles na escala de habilidades. O ideal é que os itens estejam distribuídos por uma boa extensão da escala de habilidades para que possam discriminar pessoas com níveis de habilidade diferentes.

Também devem-se analisar as estatísticas de ajuste, como nas outras facetas. Valores de *MQ–Infit* e *MQ–Outfit* fora do intervalo entre 0,5 e 1,5 indicam que as medidas podem não ser adequadamente produtivas, não oferecendo um bom ajuste dos dados aos modelos de Rasch.

Como citado anteriomente, os valores observados podem diferir dos valores esperados calculados pelo modelo de Rasch, por vários motivos. Então, como sugerido por Linacre (1998), devem-se primeiramente corrigir contradições às medidas de Rasch, em seguida, diagnosticar pessoas e itens com comportamentos fora do padrão por meio das estatísticas de ajuste e então procurar por multidimensionalidade.

3.6.7 Interpretação da qualidade da escala

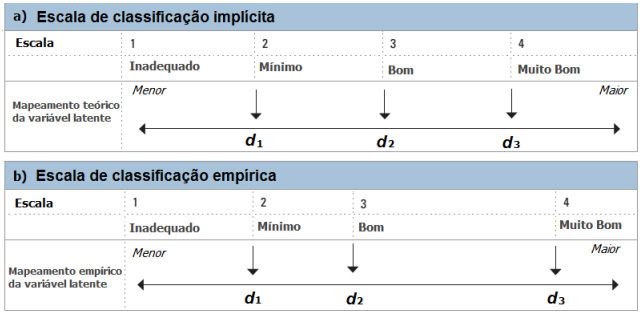
As escalas de classificação fornecem aos avaliadores um formato para que eles possam descrever seus julgamentos de acordo com critérios estabelecidos. Características de resposta típicas e exemplos são muitas vezes fornecidos durante o treinamento de avaliadores para ajudá-los a familiarizar-

-se com as diferenças de desempenho correpondentes a cada nível em uma escala de classificação. A qualidade da escala de classificação utilizada pelos avaliadores para julgar as tarefas elaboradas pelos examinandos é uma questão relevante para as avaliações que necessitam do julgamento de avaliadores (ENGELHARD, 2013).

Segundo Linacre (2002a), é produtivo que, no início das análises dos dados provenientes de testes, seja feita uma investigação sobre o funcionamento das categorias da escala de classificação. As observações em uma escala de avaliação são geralmente destinadas a capturar graus de habilidade em relação ao construto ou traço latente. Essas categorias devem obter, a partir das respostas, indicações inequívocas das localizações dos examinandos ao longo da escala de habilidades. A determinação da qualidade da escala de classificação e do conjunto de critérios utilizados na avaliação pode apoiar a suposição de unidimensionalidade psicométrica do teste.

Além disso, é importante examinar como as categorias de classificação foram interpretadas pelos avaliadores que julgaram as respostas dos examinandos. A Figura 7 destaca o mapa das categorias, que consiste em duas partes. Na parte (a), pode-se conferir a intenção que os elaboradores do teste tiveram quando definiram implicitamente a escala de classificação, alocando as categorias em intervalos igualmente espaçadas. Na parte (b), pode-se examinar, por meio de análise dos dados, o modo como os avaliadores realmente utilizaram as categorias da escala de classificação e se os intervalos estão ou não igualmente espaçados (ENGELHARD, 2013).

Figura 7 – Mapa das categorias de classificação



Fonte: Adaptado de Engelhard (2013)

Linacre (2002a) descreve um conjunto de diretrizes para examinar a qualidade das escalas de avaliação que utilizam os modelos de Rasch.

Engelhard (2013) utiliza esse mesmo conjuto para servir de guia na determinação do funcionamento das categorias da escala nas avaliações que necessitam de avaliadores.

1. *Direcionalidade:* É a orientação direcional das categorias, em sequência, da escala de classificação com a variável latente. Em outras palavras, quando a direção das categorias de uma escala de avaliação está orientada juntamente com a variável latente, espera-se que valores elevados nas observações correspondam a altas posições na variável latente. Os valores das medidas observadas e esperadas fornecem informações sobre a direcionalidade em uma escala de avaliação que podem ser utilizadas para apoiar inferências sobre a progressão da dificuldade implícita por categorias ordenadas.
2. *Monotonicidade:* É a progressão monotônica das categorias da escala de classificação. O aumentando de categorias da escala de classificação deve corresponder ao aumento das médias das medidas da habilidade dos examinandos em relação à variável latente dentro das categorias. Pode-se utilizar a média da localização da habilidade dos examinandos em todas as observações em cada categoria como indicativo de monotonia.
3. *Utilização da Categoria:* Deve-se observar a distribuição das observações nas categorias da escala de classificação. Quando a frequência de observações em todas as categorias não é igual, as categorias não podem indicar diferenças substantivas nas avaliações. Linacre (2002) sugere que as categorias com menos de 10 observações limitam a precisão e a estabilidade dessas estimativas. As categorias não observadas apresentam desafios significativos para a interpretação da escala de avaliação.
4. *Distribuição das classificações:* É o percentual de observações dentro das categorias da escala de classificação para uma determinada tarefa. Quando as classificações estão em conformidade com uma distribuição regular (uniforme, normal, unimodal, bimodais), pode-se verificar a distribuição das classificações. A presença de inclinação ou modalidade em gráficos de distribuições de classificação em todas as categorias pode ser usada para identificar rapidamente violação dessa diretriz.
5. *Ajuste da escala de classificação:* Essa diretriz está relacionada à ocorrência de valores inesperados das categorias de escala de classificação. Quando os dados se encaixam nos valores esperados pelos modelos de Rasch, um nível razoávelmente uniforme de aleatoriedade será observado.
6. *Ordem das categorias:* A localização dos coeficiente (limiares) da cate-

goria deve ter uma progressão ou desenvolvimento ao longo da variável latente. Além disso, para medidas invariantes, a capacidade dos examinandos sobre a escala de avaliação depende de uma sequência monótona da localização dos coeficientes de categoria.

7. *Localização das categorias.* A descrição precisa do desempenho dos examinandos em relação à variável latente deve corresponder à localização das categorias da escala de avaliação. Quando as categorias são distintas, cada uma delas descreve uma gama única de pessoas sobre a variável latente. Funções de informação pontiagudas fornecem evidência para localizações dos coeficientes de categoria distintas para todas as categorias da escala de classificação.

O Quadro 19 refere-se a um resumo das diretrizes e procedimentos principais para a verificação da qualidade da escala de classificação utilizada na avaliação e se o conjunto de dados está adequado para descrever a localização dos examinandos na escala de habilidades em relação ao construto.

Quadro 19 – Diretrizes: Qualidade das escalas de classificação

Diretrizes	**Questões**	**Determinação**
1. Direcionalidade	As categorias de classificação estão alinhadas com a variável latente?	Os valores das medidas observadas e esperadas devem estar alinhados.
2. Monotonicidade	A habilidade dos examinandos em relação à variável latente aumenta juntamente com as categorias de classificação?	Observar a média da localização da habilidade dos examinandos em cada categoria.
3. Uso das categorias	Existem observações suficientes por categoria?	Categorias com menos de 10 observações limitam a precisão e a estabilidade das estimativas.
4. Distribuição das classificações	Qual é a distribuição das observações em todas as categorias?	Observar se as classificações ocorrem em uma distribuição regular (uniforme, normal, unimodal, bimodais)
5. Ajuste da escala de classificação	O ajuste da escala de classificação para o modelo de Rasch é suficientemente bom?	Observar se os valores das estatísticas de ajustes *MQ–Infit* e *MQ–Outfit* estão próximos de 1.
6. A ordem das categorias	As localizações dos limites das categorias refletem a ordem pretendida?	Verificar se a localização dos coeficientes da categoria possui uma sequência monótona ao longo da escala da variável latente.
7. Localização das categorias	As localizações dos limites das categorias são distintas?	Verificar se as distâncias entre os limiares das categorias são maiores do que 1,4 *logitos*.

Fonte: Autora

4 METODOLOGIA DE PESQUISA

Apresentam-se, neste capítulo, a metodologia de pesquisa e os procedimentos metodológicos empregados para o desenvolvimento deste trabalho.

Segundo Gil (2008), para um conhecimento ser considerado científico, é necessário identificar as operações mentais e técnicas que possibilitam a sua verificação, ou seja, determinar o método para se alcançar esse conhecimento. Método é comumente definido como o caminho para se chegar a determinado fim, o conjunto de procedimentos intelectuais e técnicos adotados para se atingir o conhecimento (GIL, 2008; PACHECO *et al.*, 2007; LAKATOS; MARCONI, 2007; PRODANOV; FREITAS, 2013). A metodologia consiste em uma disciplina dedicada em compreender, avaliar e aplicar os métodos disponíveis para a realização de uma pesquisa científica. Por meio da metodologia, são possíveis a coleta e o processamento de informações, com a finalidade da resolução de problemas ou respostas à investigação (PRODANOV; FREITAS, 2013).

4.1 MÉTODOS DE ABORDAGEM

Os métodos de abordagem para pesquisas conduzidas por meio de raciocínio lógico podem ser classificados em dedutivo, indutivo e hipotético-dedutivo (PRODANOV; FREITAS, 2013; GIL, 2008; CHALMERS, 2000). Esses métodos oferecem ao pesquisador normas destinadas a distinguir entre os objetivos científicos e os não científicos e propõem procedimentos lógicos a serem seguidos no processo da investigação científica que possibilitam, entre outras, a determinação do alcance de sua investigação, das regras envolvidas para a explicação dos fatos e da validade de suas generalizações (PRODANOV; FREITAS, 2013; GIL, 2008).

Os métodos dedutivo e indutivo procedem-se inversamente um ao outro. O dedutivo parte de princípios reconhecidos como verdadeiros e indiscutíveis e, em virtude de sua lógica, possibilita chegar a conclusões de maneira formal, é o método usualmente empregado nas ciências exatas. O indutivo parte do particular e somente após o trabalho de coleta e análise de dados faz as devidas generalizações. No método indutivo, a generalização deve ser consequência da observação de casos concretos suficientemente confirmadores da realidade (GIL, 2008).

As finalidades dessas duas abordagens são distintas. Conforme Lakatos e Marconi (2007), o método dedutivo tem o propósito de explicar

o conteúdo das premissas e o método indutivo tem o objetivo de ampliar o alcance dos conhecimentos.

O método indutivo foi alvo de muitas críticas por pesquisadores do século XX. Karl Popper (1935 *apud* GIL, 2008) afirmou que o método indutivo não se justifica, pois a indução parte da observação de "alguns" fatos isolados para generalizá-los para "todos". Para tanto seria necessário que a quantidade de observações atingisse o infinito, o que nunca poderia ocorrer. Além disso, a indução apoia-se na demonstração sobre a tese que se pretende demonstrar (GIL, 2008), em outras palavras, a indução é justificada nela própria.

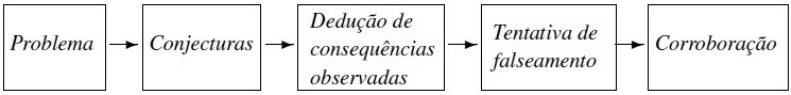
O método hipotético-dedutivo proposto por Popper consiste no seguinte: "falseabilidade de afirmações universais [leis e teorias] pode ser deduzida de afirmações singulares disponíveis" (CHALMERS, 1995).

Kaplan (1972 *apud* Gil, 2008), define esse método do seguinte modo:

> O cientista, através de uma combinação de observação cuidadosa, hábeis antecipações e intuição científica, alcança um conjunto de postulados que governam os fenômenos pelos quais está interessado, daí deduz ele as consequências por meio de experimentação e, dessa maneira, refuta os postulados, substituindo-os, quando necessário, por outros, e assim prossegue.

Gil (2008) apresenta o método hipotético-dedutivo utilizando-se do quadro exposto na Figura 8 a seguir.

Figura 8 – método hipotético-dedutivo



Fonte: (GIL, 2008)

Segundo Gil (2008), o problema de pesquisa surge quando os conhecimentos existentes para a explicação de um fenômeno são falhos ou insuficientes. Então são formuladas conjecturas ou hipóteses das quais são deduzidas as consequências que deverão ser testadas ou falseadas. Enquanto, no método dedutivo, a preocupação principal consiste em confirmar a hipótese, no método hipotético-dedutivo, procuram-se evidências empíricas para derrubá-la. Quando não se conseguem evidências capazes de falsear a hipótese, ela mostra-se válida e tem-se a sua corroboração, mesmo que provisoriamente

(GIL, 2008). Popper (1935 *apud* CHALMERS, 2000) considera que o conhecimento tem um caráter provisório e dinâmico, uma vez que as teorias são criadas para superar teorias que apresentaram problemas anteriormente, isto é, a falha de uma teoria implicará na proposição de outra, com maior poder explicativo (CHALMERS, 2000).

Com base nessas descrições, a parte prática deste trabalho é caracterizada pelo ponto de vista hipotético-dedutivo, pois parte-se da hipótese de que os resultados das avaliações são analisados de maneira mais robusta no contexto do modelo multifacetas de Rasch do que o seria pelos métodos tradicionais, possibilitando a detecção de pontos problemáticos no nível individual dos elementos participantes da avaliação, o que pode resultar em intervenções e resoluções desses problemas.

O modelo multifacetas de Rasch refere-se à aplicação de uma série de ferramentas de medição que visam proporcionar análises mais minuciosas de avaliações compostas por múltiplas variáveis.

4.1.1 Procedimentos técnicos

Os procedimentos técnicos são métodos com o objetivo de garantir a objetividade e a precisão no estudo dos fatos ou fenômenos. Segundo Gil (2008), esses métodos visam proporcionar ao investigador orientações referentes a obtenção, processamento e validação dos dados obtidos na investigação.

Esta pesquisa utiliza instrumento de avaliação com itens que possibilitam a avaliação do traço latente, que é a habilidade da expressão escrita por meio de ferramentas estatísticas para análise dos dados, o que a caracteriza como uma pesquisa de abordagem quantitativa (LAKATOS; MARCONI, 2009).

4.1.2 Classificação da pesquisa

As pesquisas podem ser classificadas quanto à natureza e quanto ao objetivo. Sob o ponto de vista de sua natureza, a pesquisa pode ser básica ou aplicada. Esta pesquisa tem como objetivo o conhecimento sobre a variabilidade da habilidade da expressão escrita além de indicar métodos para a elaboração, aplicação e validação de avaliações em larga escala com itens de respostas construídas, por isso, em relação à sua natureza, caracteriza-se como uma pesquisa aplicada, pois é orientada à geração de conhecimentos

com o propósito de aplicá-los para essas finalidades.

O presente trabalho é caracterizado como pesquisa exploratória, pois envolve um levantamento bibliográfico minucioso com a finalidade de proporcionar claro entendimento sobre os problemas e processos adotados nessas avaliações. Também é classificado como pesquisa descritiva, pois procura descrever certas características do desempenho das pessoas ao desenvolverem as atividades propostas e o estabelecimento de relações entre algumas variáveis envolvidas. Além disso, utiliza-se de técnicas padronizadas para a coleta de dados, sendo elas os itens descritos pela tarefa, os critérios de avaliação e as escalas. Assim, esta pesquisa é classificada como exploratória e descritiva.

4.2 DESCRIÇÃO DO PROCEDIMENTO METODOLÓGICO

Esta pesquisa constitui-se de duas partes, uma teórica e uma prática. A parte teórica é caracterizada pelo levantamento de referências bibliográficas e pelo estabelecimento dos procedimentos essenciais em cada uma das etapas demandadas para concepção, elaboração, aplicação, pontuação, análises, entre outros, de avaliações em larga escala com itens de respostas construídas. A parte prática consiste em um estudo, no qual são utilizadas as respostas à prova de redação do concurso público para provimento de vagas da Polícia Militar do Estado do Paraná aplicado em fevereiro de 2010 pela Coordenadoria de Processos Seletivos da Universidade Estadual de Londrina (COPS/UEL). A prova de redação desse concurso em particular foi escolhida por conter dois textos elaborados pelos candidatos e que foram pontuados em cinco competências cada, denominadas de itens, segundo as técnicas para correção estabelecidas neste trabalho de modo a assegurar a confiabilidade de pontuação e proporcionar análises sobre a qualidade da avaliação no contexto do modelo multifacetas de Rasch.

Com o intuito de responder à questão de pesquisa desta tese, foram utilizados os modelos multifacetas de Rasch de duas facetas, habilidade dos examinandos (faceta 1) e dificuldade dos itens (faceta 2) e o de quatro facetas, habilidade dos examinandos (faceta 1), dificuldade das tarefas (faceta 2), severidade dos avaliadores (faceta 3) e dificuldade dos itens (faceta 4). O modelo de duas facetas não considera os efeitos causados pelos avaliadores e se reduz a um dos modelos de Rasch clássicos para itens politômicos, o de escala gradual de Andrich (eq. (5)) e o de crédito parcial de Masters (eq. (8)).

Para o modelo multifacetas de Rasch de quatro facetas, foram implementados tanto o modelo de escala gradual, quanto o modelo de crédito parcial. Este último foi utilizado em duas formulações distintas; a primeira per-

mite que a estrutura da escala de classificação possa variar de um item para outro possibilitando análises sobre a qualidade das escalas de classificação utilizadas pelos avaliadores. A segunda formulação permite que a estrutura da escala de avaliação possa variar entre os avaliadores o que possibilita análises individuais de cada avaliador e do modo como ele atribuiu as pontuações. Assim, cada um dos modelos possibilitou um tipo específico de análises, incluindo também, estudos sobre a estrutura da escala de avaliação na qual os itens de cada uma das tarefas foram julgados. Estas e outras configurações do modelo multifacetas de Rasch encontram-se descritas no Capítulo 3.

A implementação de ambos os modelos, de duas facetas e de quatro facetas propiciaram comparações entre a classificação dos examinandos quando são considerados ou não os efeitos causados pelos avaliadores.

O Quadro 20 apresenta uma sistematização dos modelos utilizados nesta aplicação.

Quadro 20 – Modelos multifacetas de Rasch utilizados na aplicação prática

Modelos multifacetas de escala gradual (Andrich)			
Duas facetas	$\ln\left[\frac{P_{jik}}{P_{ji(k-1)}}\right]$	$=$	$\theta_j - b_i - d_k$
Quatro facetas	$\ln\left[\frac{P_{jiphk}}{P_{jiph(k-1)}}\right]$	$=$	$\theta_j - b_i - t_p - c_h - d_k$
Modelos multifacetas de crédito parcial (Masters)			
Duas facetas	$\ln\left[\frac{P_{jik}}{P_{ji(k-1)}}\right]$	$=$	$\theta_j - b_i - d_{ik}$
Quatro facetas	$\ln\left[\frac{P_{jiphk}}{P_{jiph(k-1)}}\right]$	$=$	$\theta_j - b_i - t_p - c_h - d_{ik}$
Quatro facetas	$\ln\left[\frac{P_{jiphk}}{P_{jiph(k-1)}}\right]$	$=$	$\theta_j - b_i - t_p - c_h - d_{hk}$

Fonte: Autora

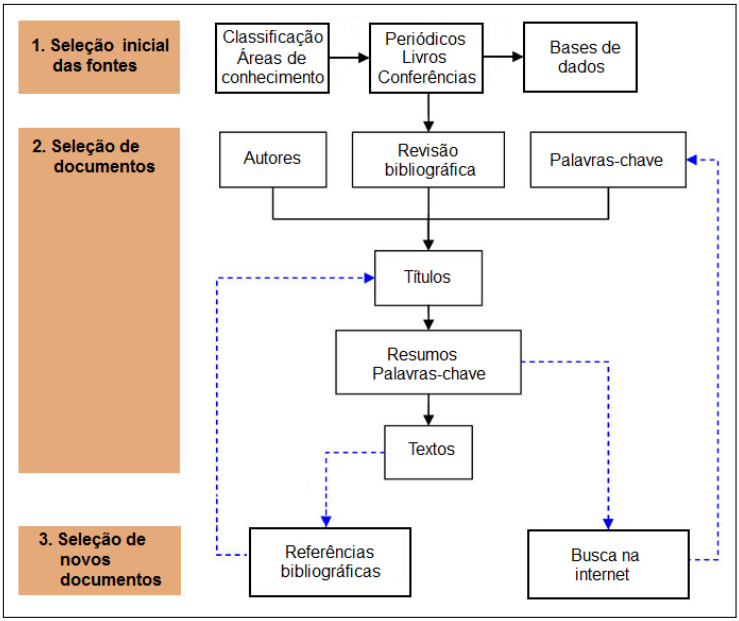
As facetas são a habilidade dos j indivíduos denotada por θ_j, a dificuldade dos i itens denotada por b_i, a dificuldade das p tarefas denotada por t_p e a severidade dos h avaliadores denotada por c_h. O tamanho do passo de

dificuldade é denotado por d_k, indicando que a escala de classificação não varia entre os itens, d_{ik} que a escala de avaliação varia entre os itens e d_{hk} que a escala de avaliação varia entre os avaliadores.

4.3 PROCEDIMENTOS ADOTADOS NA PESQUISA BIBLIOGRÁFICA

Para a parte teórica, é feito um levantamento bibliográfico sistemático sobre as avaliações que necessitam do julgamento de avaliadores para pontuar as tarefas elaboradas pelos examinandos, especificamente sobre as avaliações com itens de respostas construídas, utilizando-se parcialmente da técnica proposta por Villas *et al.* (2008), que consiste em três estágios, conforme ilustrado na Figura 9: (1.) a seleção das fontes de dados iniciais, (2.) a seleção de documentos e (3.) a seleção de novos documentos.

Figura 9 – Método de busca bibliográfica



Fonte: Adaptado de Villas *et al.* (2008)

Na primeira etapa, fez-se uma seleção inicial das fontes de dados, que consistem em livros, teses, pesquisas divulgadas em eventos científicos e as bases de dados que fazem parte do portal de periódicos da CAPES, sendo

estas últimas as fontes principais para a obtenção de artigos científicos. As áreas consideradas nas pesquisas são: Humanas, Ciências Sociais, Engenharias, Psicometria e Multidisciplinar.

Foram pesquisadas as seguintes bases de dados:

Academic Search Premier (EBSCO)	Scielo
Cambridge University Press	ScienceDirect (Elsevier)
SAGE Journals Online	Scopus

As pesquisas de documentos nessas fontes, primeiramente, foram desenvolvidas utilizando-se palavras-chave e também algumas combinações entre elas. O esquema descrito no Quadro 21 traz as principais palavras-chave utilizadas e um resumo das buscas realizadas de acordo com algumas combinações entre essas.

Quadro 21 – Esquema de busca por palavras-chave

Palavra chave	**e/ou (*and/or*)**	**e/ou (*and/or*)**
avaliação	vestibular	ENEM
avaliação escrita		
itens abertos		
redação		
multifacetas de Rasch		
many-facet Rasch	*assessment*	*test*
performance		
writing proficiency		
large-scale		
high-stakes		
rater-mediated		
validade	*performance assessment*	*writing assessment*
confiabilidade		
reliability		
rubrics		
scoring		
comparability		
rater variability		
rater tendency		

Fonte: Autora

As teorias envolvidas nas avaliações com itens abertos tiveram um desenvolvimento grande a partir dos anos de 1950, por esse motivo, para a obtenção de artigos seminais, não houve limite quanto ao tempo na primeira fase dessas pesquisas.

A partir desses resultados, foi utilizado o gerenciador de pesquisas acadêmicas gratuito *"Mendeley Desktop"*, para a exclusão de documentos repetidos e uma primeira triagem por título, palavras-chave e resumos. Esse procedimento está ilustrado na Figura 9, estágio 2.

Uma segunda triagem, mais minuciosa, foi feita com o auxílio do *software* livre "Docear", uma suíte desenvolvida para procurar, organizar e criar literatura acadêmica por meio de mapas mentais. Essa etapa consistiu na leitura dos textos e na classificação por palavras-chave, relevância para o trabalho, do capítulo ou seção da tese na qual o texto se insere, entre outros. Essa classificação é importante porque facilita novos acessos aos documentos durante o desenvolvimento do trabalho.

Para a diagramação da tese foi utilizado o Programa LaTeX com a classe de formatação de teses desenvolvida para o Programa de Pós-graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina, por Moreto (2009), denominada "pgeeltex".

Além disso, os programas Mendeley e Docear permitem a geração automática, na linguagem LaTeX, da bibliografia utilizada conforme estilo predefinido pelo usuário.

A terceira fase das pesquisas consistiu em buscar as referências relevantes para o trabalho citadas nos artigos já acervados e também em algumas buscas por palavras-chave na internet aberta. Durante todo o desenvolvimento do trabalho, novas pesquisas foram feitas nas bases, limitando-se as buscas apenas para o ano atual em questão, para a inclusão de documentos recentes.

4.4 INSTRUMENTO DE AVALIAÇÃO

A coleta de dados para a pesquisa foi realizada por meio de dois itens de respostas construídas aplicados no concurso público para provimento de vagas da Polícia Militar do Estado do Paraná elaborado pela COPS/UEL, ocorrido em fevereiro de 2010. A COPS/UEL – Coordenadoria de Processos Seletivos da Universidade Estadual de Londrina é o órgão dessa universidade responsável pela elaboração e aplicação dos concursos vestibulares da própria instituição e de outras instituições de menor porte e também pela prestação de serviços em concursos e testes seletivos, atendendo às necessidades de

seleção de profissionais do setor público e privado.

A primeira fase desse concurso público constou de uma prova escrita de conhecimentos, de caráter eliminatório e classificatório, contendo 48 itens objetivos de múltipla escolha e 2 itens com respostas construídas. Os itens objetivos foram elaborados abrangendo conteúdos do ensino médio das disciplinas: Língua Portuguesa, Matemática, Estatuto da Criança e do Adolescente (ECA), Ciências da natureza, Ciências humanas e os dois itens com respostas construídas abrangendo a área de conhecimentos gerais.

Os dois itens de respostas construídas no teste de conhecimentos gerais da avaliação foram concebidos para avaliar a capacidade de expressão escrita dos candidatos. Os aspectos desse construto que foram previstos no edital do concurso para serem avaliados são:

1. Observância das normas de ortografia, pontuação, concordância, regência e flexão.
2. Paragrafação, estruturação de períodos, coerência e lógica na exposição das ideias.
3. Pertinência da exposição relativamente ao tema e à ordem de desenvolvimentos propostos.

Esses dois itens de respostas construídas, na forma de tarefas de escrita, concebidos com temas que fazem parte dos conteúdos das disciplinas previstas no edital do concurso e, além disso, tais temas pertencem ao universo da profissão para a qual o concurso foi destinado. As duas tarefas estão expostas no Anexo A.

Como, no Brasil, o nível médio de ensino é estruturado a partir da matriz de competências e habilidades definidas pelo PCN – Ensino Médio, tanto o ENEM como os exames vestibulares e os concursos de nível médio de ensino devem ser estruturados seguindo diretrizes estabelecidas nesse documento. Desse modo, para a aplicação prática nesse trabalho, os textos escritos pelos participantes da avaliação foram pontuados de acordo com uma adaptação da matriz de referência para a redação do ENEM, divulgada no *Guia do Participante* (BRASIL, 2013). Essa lista de habilidades e competências estabelecida para a redação do ENEM se enquadra perfeitamente nas definições dessa aplicação prática, uma vez que os critérios para pontuação do ENEM também consistem em pontuação analítica, os construtos avaliados e o nível de ensino são os mesmos.

A principal adaptação feita às competências e habilidades estabelecidas para a prova de redação do ENEM se refere ao tipo de texto: no ENEM, o participante deve desenvolver o tema dentro dos limites estruturais do texto

dissertativo-argumentativo e, na avaliação dessa aplicação, o tipo de texto é dissertativo.

A pontuação analítica foi utilizada nesta aplicação prática pois permite identificar separadamente qualidades específicas do texto incluindo um maior nível de detalhes nas informações, clareza quanto aos tópicos que estão sendo medidos, facilidade na interpretação da relação entre o que está sendo medido e as pontuações correspondentes, além de ser mais fácil o treinamento dos avaliadores com o objetivo de se obter um nível razoável de confiabilidade. Cada uma das duas tarefas foi subdividida em 6 itens (competências) em uma escala que varia de 1 a 6 pontos.

As competências avaliadas e as competências avaliadas juntamente com os níveis da escala de avaliação utilizadas na aplicação prática deste trabalho, encontram-se descritas nos Apêndices A e B. Já um resumo dos critérios utilizados encontra-se no Apêndice B.1.

Essa avaliação ocorreu no estado do Paraná e suas provas foram aplicadas em 5 cidades do Estado: Cascavel, Curitiba, Foz do Iguaçu, Londrina e Maringá. A finalidade da avaliação foi a de selecionar pessoas para ocupar vagas de trabalho da Polícia Militar do Estado do Paraná. Tanto as questões objetivas quanto as questões abertas foram de caráter eliminatório, devendo o candidato ter obtido, no mínimo, 50% de acertos em cada uma delas. Pela maior complexidade e maior custo para a pontuação dos itens abertos, só foram corrigidas as questões dos candidatos que atingiram a pontuação mínima exigida nas questões objetivas, isto é, responderam corretamente a pelo menos 24 itens. Desse modo, o número de candidatos que tiveram suas questões abertas corrigidas foi de 17.112.

Infelizmente para esta aplicação prática, não foi possível corrigir novamente as respostas de todos esses candidatos, então foram separados desse montante os dois itens respondidos por 350 candidatos. As respostas dos candidatos foram selecionadas para representar todos os níveis de proficiência alcançados pelos examinandos na pontuação original do concurso.

4.5 TREINAMENTO DOS AVALIADORES

Para a formação do grupo de avaliadores, foi elaborado um evento junto à Pro-reitoria de Extensão Universitária da Universidade Estadual de Londrina (UEL) intitulado "*Oficina para correção de redações*". Este evento teve a coordenação e a participação em todo o processo de dois professores doutores do departamento de letras da UEL, experientes na correção de reda-

ções do vestibular e de outros concursos promovidos pela COPS/UEL e foi dirigido a professores de Língua Portuguesa do ensino médio da rede estadual de ensino, a alunos do último ano de graduação em letras da UEL e a alunos dos programas de pós-graduação do departamento de letras também da UEL. A carga horária destinada para o evento foi de 20 horas, distribuídas em seis encontros que ocorreram aos sábados, no período matutino, dos meses de outubro e novembro de 2013. Inicialmente 60 alunos se inscreveram, depois da primeira aula, na qual foram esclarecidos os objetivos da oficina, alguns não continuaram, resultando em 42 alunos efetivos.

Os participantes da oficina assistiram a aulas sobre alguns fundamentos da avaliação, como seus propósitos conforme descritos na Seção 2.2 e sobre os princípios essenciais para uma avaliação eficiente com ensinamentos sobre as questões de validade e confiabilidade descritos na Seção 2.3. Além do caráter formativo, essas aulas foram consideradas importantes para que todos os participantes entendessem a seriedade com a qual as pontuações devem ser atribuídas e que as avaliações causam consequências na vida das pessoas. Essas aulas ocuparam dois encontros, assim, apenas no terceiro encontro é que efetivamente teve início o treinamento para as pontuações das redações.

O treinamento de 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. Os tipos de abordagens comumente utilizadas encontram-se descritas na Seção 2.5.2.3 deste trabalho.

O treinamento dos avaliadores consistiu no seguinte:

1. A equipe de coordenação da correção escolheu seis textos escritos pelos participantes, sendo três para cada item da avaliação. Foram separados textos em três níveis de desempenho, *muito bem escrito*, *escrito medianamente* e *muito mal escrito*.

2. Foram distribuídas cópias dos critérios de avaliação (Apêndice B.1) aos participantes e esses foram explicados pelos professores coordenadores que também esclareceram os procedimentos e as eventuais dúvidas que surgiram.

3. O treinamento para o uso dos critérios foi feito primeiramente para o item aberto de número 1 (Anexo A). Cópias do texto escolhido como referência para o *muito bem escrito* foram distribuídas aos participantes da oficina e todos o corrigiram. Os participantes relataram as notas que atribuíram para

cada uma das cinco competência, uma de cada vez, em voz alta. As notas discrepantes tiveram suas pontuações esclarecidas por meio de discussões no grupo sobre os motivos da atribuição de tais pontuações. Desse modo sendo sanadas as dúvidas geradoras de diferenças nas pontuações. As notas consideradas discrepantes são aquelas com diferença maior do que 1 ponto. Os mesmos procedimentos foram repetidos para os outros textos de referência, o *escrito medianamente* e o *muito mal escrito*.

4. Após as sessões de pontuação para a correção dos textos escritos pelos candidatos em resposta ao item de número 1, cujos procedimentos estão relatados na próxima seção, os procedimentos de número três desse treinamento foram repetidos para o item de número 2 do teste (Anexo A).

4.6 PONTUAÇÃO DO TESTE

Na ocasião do concurso público para provimento de vagas da Polícia Militar do Estado do Paraná, um total de 17.112 participantes tiveram seus dois itens com respostas abertas pontuados para as suas classificações efetivas. Nesse experimento, no entanto, foram separadas inicialmente para a pontuação as respostas de apenas quatrocentos (400) participantes. Este número foi estabelecido por razão do tempo disponível para as pontuações e pela falta de experiência da equipe de avaliadores, lembrando que cada participante elaborou dois textos resultando um total de 800 (400×2) respostas para pontuar. Desse total, algumas respostas foram descartadas porque um dos ensaios estava em branco. Além disso, alguns ensaios não foram pontuados porque o número de sessões para a pontuação foi insuficiente. A diversidade quanto à formação e à experiência dos avaliadores resultou em lentidão nas pontuações, além de um número significante de notas discrepantes. Desse modo foram efetivamente pontuadas setecentas (700) respostas elaboradas por trezentos e cinquenta (350) participantes. As sessões para a pontuação dos ensaios ocorreram do seguinte modo:

1. Os textos produzidos pelos candidatos ao concurso para o item foram separados de 10 em 10 e colocados em envelopes, cada um desses envelopes foi distribuído para um dos avaliadores que corrigiu os 10 textos. Cada envelope, cujos textos foram todos corrigidos pelo primeiro avaliador, recebeu a notação "I–" para indicar a correção de número 1 e, na frente dessa notação, o avaliador que fez essa correção anotou o seu número de identificação. Após todos os envelopes terem sido corrigidos pela primeira vez,

o procedimento foi repetido para a segunda correção. Desta vez, a notação "II–" indicou a segunda correção e o avaliador correspondente anotou o seu número de identificação.

2. Problemas ou dúvidas pontuais que surgiram durante as sessões de correção foram resolvidos por um dos avaliadores coordenadores em particular com o avaliador portador do problema. Problemas que ocorreram repetidamente, como respostas anormais, foram resolvidos pelos coordenadores da correção e comunicados verbalmente e no quadro de avisos ao grupo.

3. A pontuação final foi a média aritmética dessas pontuações. A acuracidade da pontuação entre os avaliadores foi monitorada pela diferença entre as pontuações atribuídas a cada texto; quando essa diferença foi maior do que 3 pontos na nota final, que pode chegar a 30 pontos, ou de 2 pontos em alguma das competências, que pode atingir o valor máximo de 6 pontos, foi detectada uma discrepância. Os textos que receberam notas discrepantes foram separados novamente em envelopes com dez textos e estes foram corrigidos novamente por um novo avaliador.

4. Quando, após a terceira correção, ainda persistiu uma discrepância, o texto foi corrigido novamente pelos professores coordenadores da oficina.

Após a primeira correção, foram detectadas 98 notas discrepantes atribuídas a questão de número 49, equivalendo a 28% das correções a esta questão. Para a questão de número 50, ocorreram apenas 49 notas discrepantes, 14% do total das correções a esta questão. Esta diferença entre as notas discrepantes das duas questões se deu devido à inexperiência dos avaliadores que foram aprimorando os seus desempenhos durante as sessões de correção. Ressaltando que a questão de número 49 de todos os participantes foi corrigida em primeiro lugar, só após esta etapa ter sido finalizada, a questão de número 50 de todos os participantes foi pontuada. Após estes itens com notas discrepantes terem sido submetidas novamente à uma terceira correção, ainda resultaram 18 notas discrepantes da questão de número 49 (5,14%) e 8 notas discrepantes da questão de número 50 (2,6%). Estas questões foram então pontuadas pelos coordenadores da oficina de correção das redações.

4.7 ANÁLISES DOS DADOS

Para as análises dos dados foi utilizado o *software* comercial *Facets* versão 3.71.4 (LINACRE, 2014b). O programa foi utilizado para estimar a

proficiência individual de cada examinando (faceta 1), a dificuldade das tarefas (faceta 2), a severidade com que cada avaliador julgou as tarefas elaboradas pelos examinandos (faceta 3) e a dificuldade dos itens (faceta 4). Ainda foram feitos estudos sobre a escala de avaliação na qual os itens de cada uma das tarefas foram avaliados. Foram implementados tanto o modelo de escala gradual quanto o modelo de crédito parcial. A descrição dos métodos e modelos utilizados para a geração dos dados encontram-se descritos na Seção 4.2 que contém também uma sistematização dos modelos utilizados no Quadro 20.

Todas as facetas são centradas na origem da escala *logitos*, exceto a faceta examinandos. Foram utilizados os critérios de convergência padrão do programa, ou seja, o procedimento de estimação é JMLE (*Joint Maximum Likelihood Estimation*), também conhecido como UCON (*Unconditional estimation algorithm*) ou *incondicional máxima verossimilhança*, em português. Esse método de estimação encontra-se descrito na Seção 3.4.5. 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 *logitos*. Mais detalhes sobre os processos de estimação podem ser verificados na Seção 3.4.

O processo de estimação para o modelo multifacetas de escala gradual de quatro facetas terminou automaticamente após 213 iterações. Para o modelo multifacetas de crédito parcial, também de quatro facetas, no qual a estrutura da escala de classificação varia de um item para outro, o processo de estimação terminou automaticamente após 211 iterações enquanto para o modelo no qual a estrutura da escala de avaliação varia entre os avaliadores, o processo terminou após 342 iterações. Já, para o modelo de escala gradual clássico, de duas facetas, o processo de estimação terminou com 92 iterações. Neste caso, foi considerada a dificuldade dos itens de cada uma das tarefas.

5 RESULTADOS

O objetivo deste capítulo é realizar análises dos resultados provenientes da pontuação dos dois itens de respostas construídas elaboradas pelos candidatos ao concurso público.

Com a utilização do modelo multifacetas de Rasch de quatro facetas (eq. (11)), é analisada a confiabilidade da pontuação proveniente dos avaliadores e são identificados os avaliadores portadores de tendências em pontuações sistemáticas responsáveis pela geração de erros nas pontuações. A dificuldade dos itens e a estrutura da escala de avaliação são também estudadas.

5.1 ANÁLISE DO AJUSTE GLOBAL DOS DADOS AO MODELO MFR

Uma análise geral dos dados pode ser feita por meio das respostas não esperadas calculadas pelo programa *Facets* (LINACRE, 2014b) a partir das hipóteses do modelo. Ao todo, houve 7.604 respostas válidas, isto é, as respostas utilizadas para a estimativa dos parâmetros do modelo. Apenas 363 respostas, número equivalente a 4,8%, tiveram seus resíduos padronizados em valores absolutos iguais ou maiores do que 2. Entre estas, 23 respostas, ou o equivalente a 0,3% do total, foram associadas com resíduos padronizados em valores absolutos iguais ou superiores a 3. Esses resultados, tomados em conjunto, indicam um ajuste satisfatório dos dados ao modelo. Na sequência são apresentadas estatísticas mais detalhadas para avaliar o ajuste dos dados ao modelo multifacetas de Rasch.

A Figura 10 exibe o mapa das variáveis, segundo o modelo multifacetas de escala gradual (equação (11)), no qual é utilizada uma única escala de classificação para todos os avaliadores em todos os itens. Esse mapa representa as calibrações de todas as quatro facetas: a habilidade dos examinandos (“Examinandos”), a dificuldade das tarefas (“Tarefa”), a severidade dos avaliadores (“Avaliador”) e a dificuldade dos itens (“Item”), além da localização dos limiares entre as categorias da escala de classificação de seis pontos utilizada pelos avaliadores para pontuar as tarefas elaboradas pelos examinandos. O mapa das variáveis é um recurso muito informativo, fornecido pelo programa *Facets*, para auxiliar na interpretação dos dados de saída do programa, retratando todas as facetas da análise em um único quadro de referência. Esse recurso é de grande valia para facilitar comparações dentro e entre as várias facetas.

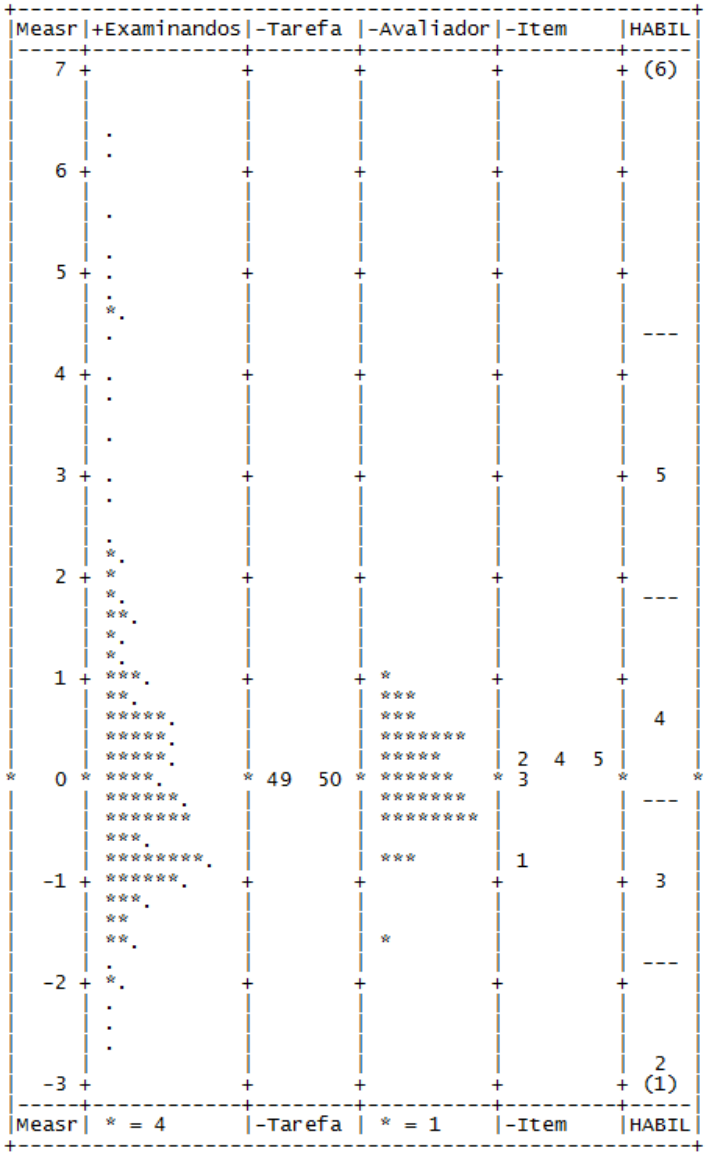
Nesse mapa, todas as medidas são dadas na mesma escala em *logitos* exibida na primeira coluna. A segunda coluna apresenta a distribuição das medidas da habilidade dos participantes do exame, na qual, cada asterisco representa 4 indivíduos e cada ponto, algum número menor do que 4. O sinal "+" que antecede a palavra "Examinandos" significa que as habilidades dos examinandos são distribuídas no gráfico de acordo com a orientação positiva, isto é, quanto maior a pontuação (escore), maior é a medida, nesse caso, a habilidade do indivíduo. Desse modo, na parte superior da coluna estão os indivíduos com maior habilidade, e os de menor habilidade estão representados pelas marcações na parte de baixo da coluna. As medidas dos desempenhos dos examinandos variam entre -2,63 e 6,46 *logitos*, embora a maior concentração de indivíduos ocorra entre -2,0 e 2,0 *logitos*. A média das medidas da habilidade dos participantes é de M = 0,22 e o desvio padrão é de SD = 1,52, com erro padrão de 0,05. A precisão é dada em termos do erro padrão, isto significa que quanto menor for o erro padrão, maior será a precisão das medidas.

As tarefas relatadas na terceira coluna possuem orientação negativa, significando que maior escore corresponde a uma menor medida. Neste estudo, as tarefas obtiveram níveis de dificuldade parecidos. A questão de número 49 obteve 0,04 *logitos* e a questão de número 50 obteve -0,04 *logitos*, com erro padrão de 0,02.

Na quarta coluna, está a distribuição do desempenho dos avaliadores quanto à severidade. Cada asterisco corresponde a um avaliador, e essa faceta possui orientação negativa, significando que o avaliador mais severo atribui notas menores enquanto o mais complacente, notas mais elevadas, assim, maior escore implica em menor medida. Os asteriscos na parte de baixo da coluna representam os avaliadores mais complacentes, enquanto os mais severos estão localizados na parte superior da coluna. As medidas da severidade dos avaliadores variam entre -1,54 e 0,94 *logitos* e a média é 0,0, com erro padrão de 0,5 *logitos*.

A quinta coluna mostra a dificuldade dos itens, que varia entre -0,74 e 0,29 *logitos* e a média é 0,0. Embora existam diferenças entre a dificuldade dos itens, estes não ocupam um intervalo amplo na escala de habilidades, estão todos localizados perto da origem.

Figura 10 – Mapa das variáveis – Modelo: Escala gradual



Fonte: Linacre (2014b)

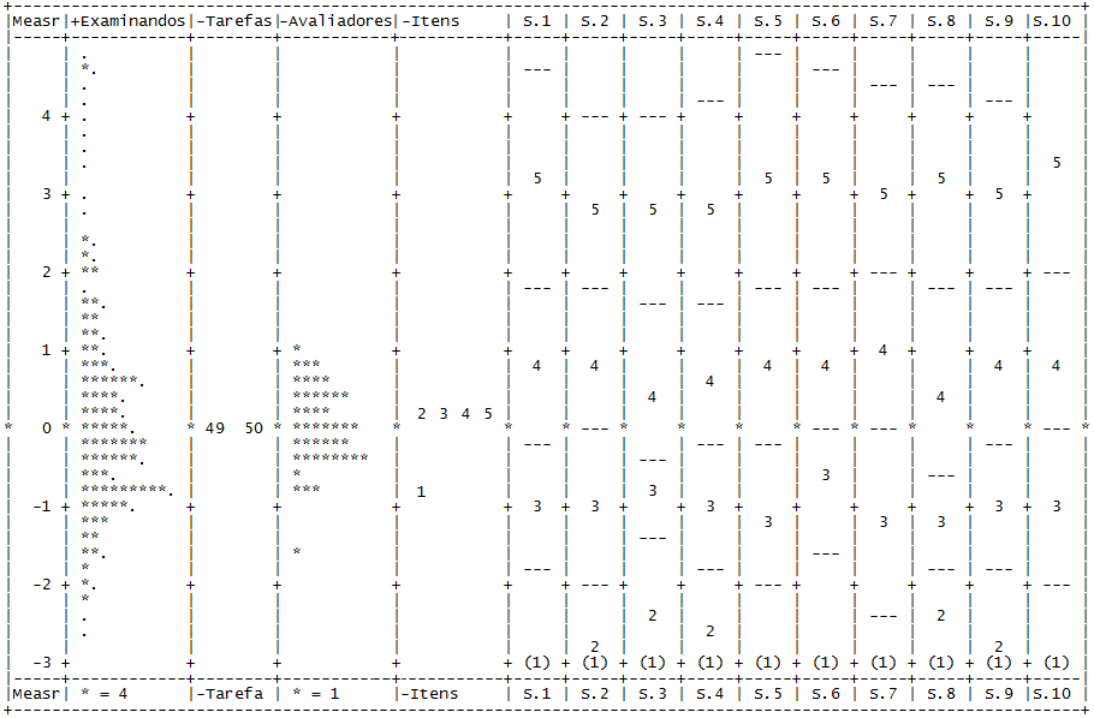
Outro ponto interessante é que a faixa de distribuição dos avaliadores está limitada entre, aproximadamente, -1,5 e 1 (*logitos*), muito estreita se comparada com a distribuição dos examinandos. Esse fato se dá porque foi exigida a concordância entre as notas dos avaliadores nas seções de pontuação.

Quanto à dificuldade das tarefas e dos itens, quanto mais difíceis eles forem menores serão os seus escores, nesse caso o escore é proporcional ao número de respostas corretas. As tarefas 49 e 50 possuem dificuldades equivalentes. Os itens 2, 4 e 5 são equivalentes quanto às suas dificuldades, enquanto o item de número 1 é o mais fácil. A última coluna representa a escala de 1 a 6 na qual os ensaios foram pontuados. Nota-se que as categorias são de comprimentos diferentes, significando que cada categoria corresponde a diferentes "quantidades" do traço latente. Essa é a distribuição das categorias ao longo da escala, do modo como os avaliadores as utilizaram. Outras discussões sobre esse assunto podem ser conferidas na Seção 5.5.

A Figura 11 exibe parte do mapa das variáveis representando as calibrações de todas as quatro facetas incluídas no modelo multifacetas de Rasch, segundo o modelo de crédito parcial. As facetas incluídas nesse modelo são as mesmas que fizeram parte do modelo de escala gradual, isto é, a habilidade dos examinandos, a dificuldade das tarefas, a severidade dos avaliadores e a dificuldade dos itens. A interpretação desse mapa de variáveis é semelhante ao da Figura 10, no entanto, o modelo multifacetas de Rasch para escala de crédito parcial não impõe uma escala de classificação fixa para todos os itens, ao contrário, cada item possui a sua própria estrutura de escala de classificação (veja o Quadro 20). Desse modo, nas colunas de 6 a 15 são incluídas as escalas de avaliação de cada item de cada tarefa, resultando em 10 escalas distintas, pertencendo as escalas S.1,$\cdots$,S.5 aos cinco itens da tarefa 49 e as escalas S.6,$\cdots$,S.10 aos itens da tarefa 50. Outras informações e as interpretações dessas escalas de avaliação serão tratadas com detalhes na Seção 5.5.

Para melhor exibição desta figura, são exibidas neste mapa apenas as informações que possuem medidas entre -3 e 5 *logitos*, o que não prejudica as interpretações pois a maior parte das medidas se encontram nesta faixa da escala.

Figura 11 – Mapa das variáveis – Modelo: Escala de crédito parcial



Fonte: Linacre (2014b)

5.1.1 Resumo dos resultados

As estatísticas expostas na Tabela 1 são referentes à habilidade dos examinandos, à dificuldade das tarefas, à severidade dos avaliadores e à dificuldade dos itens estimadas pelo programa *Facets* (LINACRE, 2014b) segundo o modelo de escala gradual do modelo multifacetas de Rasch. As estatísticas de separação, quando apresentam resultados significativos, indicam diferenças entre os elementos dentro de cada uma das facetas em toda a variável latente (ENGELHARD; WIND, 2013) e referem-se à reprodutibilidade das medidas (LINACRE, 2014a). Em geral, as medidas apresentadas neste estudo possuem valores altos para a confiabilidade de separação das facetas Examinandos (0,96), Avaliadores (0,94) e Itens (0,99). A faceta Tarefas apresenta a confiabilidade de separação com valor um pouco menor, 0,64. Além disso, essas medidas são significativas com probabilidade $p < 0,05$.

A confiabilidade das estatísticas de separação provenientes do programa *Facets* para indivíduos é comparável à confiabilidade do coeficiente alfa de Cronbach (ENGELHARD; WIND, 2013).

Tabela 1 – Resumo das análises estatísticas – Modelo: Escala gradual

	Examinandos	Tarefas	Avaliadores	itens
Medidas básicas				
Média	0,22	0,0	0,0	0,0
Desvio padrão	1,55	0,04	0,51	0,38
Número	350	2	44	5
Média quadrática (MQ)– Infit				
Média	0,98	1,0	1,0	1,00
Desvio padrão	0,44	0,04	0,26	0,17
Média quadrática (MQ)– Outfit				
Média	0,99	1,00	1,0	1,0
Desvio Padrão	0,43	0,03	0,26	0,15
Estatísticas de separação				
Taxa de separaração	5,01	1,35	3,88	10,88
Confiabilidade de separação	0,96	0,64	0,94	0,99
Estrato	7,01	2,13	5,5	14,85
Qui-quadrado (χ^2)	6697,1*	5,6*	785,3*	565,4*
Graus de liberdade	349	1	43	4
* p<0,05				

Fonte: Dados da pesquisa

Neste estudo, a média quadrática de todas as medidas estão de acordo com as sugestões de Wright e Linacre (1994) (Seção 3.5.1) embora, para a faceta examinandos, as medidas *MQ–Infit* e *MQ–Outfit* possuam desvios padrão maiores do que os desvios padrão dessas medidas das outras facetas. Esses são quase duas vezes maiores do que os desvios para a faceta avaliadores que, por sua vez, também são maiores do que os desvios padrão dessas medidas das outras facetas. Isso sugere que os valores das estatísticas de ajuste para cada um dos examinandos individualmente podem ter valores fora da faixa de valores produtivos, distorcendo as medidas. O mesmo pode ocorrer com a faceta avaliadores.

A Tabela 2 fornece um resumo das estatísticas básicas provenientes do modelo multifacetas de escala de crédito parcial. Essas estatísticas referem-se à habilidade dos examinandos, à dificuldade das tarefas, à severidade dos avaliadores e à dificuldade dos itens.

Tal como relatado para o modelo multifacetas de escala gradual, as diferenças gerais entre os examinandos, os itens, as tarefas e os avaliadores são significativas, com p <0,05 indicando que os elementos são dispersos dentro de cada faceta em toda a escala de habilidades. Os valores das estatísticas *infit* e *outfit* são semelhantes aos obtidos para o modelo de escala gradual, sugerindo um bom ajuste dos dados ao modelo de crédito parcial.

5.2 MEDIDA DA HABILIDADE DOS EXAMINANDOS

As medidas da habilidade dos examinandos não diferiram significativamente segundo os dois modelos multifacetas de Rasch implementados neste estudo, o de escala gradual e o de crédito parcial, ambos com quatro facetas. Entretanto, as medidas da habilidade dos examinandos obtidas pelo modelo de escala gradual de duas facetas, habilidade dos examinandos e dificuldade dos itens, que resulta no modelo original de Andrich, foram significativamente diferentes das medidas obtidas pelos modelos multifacetas de Rasch com quatro facetas.

Desse modo, nesta seção é feita uma breve análise da classificação dos examinandos segundo os modelos de escala gradual de duas facetas e de quatro facetas, entretanto, para as outras análises sobre a faceta habilidade dos examinandos, são utilizados os dados resultantes do modelo de escala gradual de quatro facetas.

Tabela 2 – Resumo das análises estatísticas – Modelo: Escala de crédito parcial

	Examinandos	Tarefas	Avaliadores	itens
Medidas básicas				
Média	0,26	0,0	0,0	0,0
Desvio padrão	1,58	0,02	0,51	0,44
Número	350	2	44	5
Média quadrática (MQ)– Infit				
Média	1,00	1,01	1,01	1,01
Desvio padrão	0,46	0,00	0,27	0,08
Média quadrática (MQ)– Outfit				
Média	1,00	1,02	1,02	1,02
Desvio padrão	0,48	0,00	0,28	0,10
Outras medidas				
Taxa de separaração	5,06	1,55	3,88	12,57
Confiabilidade de separação	0,96	0,71	0,94	0,99
Estrato	7,08	2,40	5,56	17,10
Qui-quadrado (χ^2)	6885,4*	6,8*	794,7*	768,7*
Graus de liberdade	349	1	43	4
* p<0,05				

Fonte: Dados da pesquisa

A Tabela 3 traz um resumo das medidas da habilidade de alguns examinandos obtidas com a implementação do modelo de escala gradual com duas facetas que considera apenas a habilidade dos examinandos e a dificuldade dos itens, não levando em conta o desempenho dos avaliadores.

Entre as medidas estão as dos examinandos de menor habilidade, de maior habilidade e de alguns com habilidades intermediárias, apresentadas em ordem crescente, de cima para baixo. Desse modo, o examinando de número 1770, com medida da habilidade -2,23 *logitos*, é o que se saiu pior no exame, enquanto o de número 17916, com a medida da habilidade 5,71 *logitos*, foi o mais bem sucedido no teste. A segunda coluna exibe o escore total alcançado pelo examinando. O teste é composto por duas tarefas com cinco itens cada uma, que foram corrigidas, pelo menos, por dois avaliadores distintos. Cada item pode receber pontuações (escores) que variam de 1 a 6, desse modo, o escore mínimo que um examinando que elaborou as duas tarefas pode alcançar é de 20 pontos (2 tarefas $\times$ 5 itens $\times$ 1 ponto $\times$ 2 avaliadores = 20), sendo 10 pontos de cada avaliador, enquanto o escore máximo possível é de 120 pontos (2 tarefas $\times$ 5 itens $\times$ 6 pontos $\times$ 2 avaliadores = 120), sendo 60 pontos de cada avaliador. Entretanto, quando as duas pontuações

de alguma das tarefas apresentam discrepância, esta tarefa é corrigida por um outro avaliador, e neste caso o escore recebido pelo examinando é maior. Embora nos modelos de Rasch clássicos, um maior escore corresponda a uma maior medida da habilidade, neste caso, por haver a mediação de avaliadores, este fato pode não ser verdadeiro.

Tabela 3 – Resumo das medidas dos examinandos – Modelo: Escala gradual de duas facetas

		Número		Erro	MQ		Média		
Examin.	Escore	pontuação	Habilid.	padrão	*Infit*	*Outfit*	observ.	justa	Classif.
1770	22	10	-2,23	0,43	1,13	1,05	2,20	2,19	350
2617	44	20	-2,23	0,30	0,71	0,73	2,20	2,19	349
3023	44	20	-2,23	0,30	0,57	0,58	2,20	2,19	348
373	58	25	-2,02	0,26	0,67	0,66	2,32	2,31	345
765	59	25	-1,95	0,26	0,89	0,91	2,36	2,35	343
2516	48	20	-1,89	0,29	0,35	0,35	2,40	2,39	342
...	...	...	...	...	...	...	...	...	...	...
2477	52	20	-1,57	0,28	0,77	0,76	2,60	2,59	334
2228	71	20	-0,14	0,28	0,71	0,76	3,55	3,56	188
596	71	20	-0,14	0,28	1,25	1,34	3,55	3,56	185
25	72	20	-0,07	0,28	0,48	0,50	3,60	3,61	177
2582	92	25	0,06	0,26	1,62	1,60	3,68	3,69	157
2305	93	25	0,13	0,26	1,04	1,08	3,72	3,73	147
302	101	25	0,69	0,27	4,47	4,47	4,04	4,05	93
...	...	...	...	...	...	...	...	...	...	...
36	84	20	1,00	0,31	1,30	1,30	4,20	4,20	69
945	85	20	1,09	0,31	1,00	1,03	4,25	4,25	65
1675	87	20	1,30	0,32	0,46	0,46	4,35	4,35	60
2007	109	25	1,32	0,29	1,09	1,08	4,36	4,36	55
2192	91	20	1,72	0,33	0,44	0,43	4,55	4,55	41
...	...	...	...	...	...	...	...	...	...	...
17652	111	20	4,29	0,41	0,74	0,71	5,55	5,56	10
23499	112	20	4,47	0,43	0,91	0,96	5,60	5,61	6
9666	114	20	4,86	0,47	0,83	0,77	5,70	5,71	4
34425	116	20	5,37	0,55	0,91	0,84	5,80	5,81	3
35883	116	20	5,37	0,55	0,91	0,84	5,80	5,81	2
17916	117	20	5,71	0,62	0,96	0,91	5,85	5,86	1

Fonte: Dados da pesquisa

Observa-se na terceira coluna o número de itens que foram pontuados. Normalmente seriam 20 pontuações, isto é, 10 itens (2 tarefas $\times$ 5 itens = 10) que foram corrigidos por dois avaliadores distintos. Entretanto, algu-

mas tarefas foram pontuadas pela terceira vez, somando-se então mais cinco pontos, além disso, alguns examinandos elaboraram apenas uma das tarefas, recebendo menos pontos.

Para exemplificar, o examinando de número 2516 no primeiro bloco da Tabela 3 possui escore 48, número de pontuação 20 e medida da habilidade de -1,89 *logitos*. O escore corresponde à soma das notas que o examinando recebeu em cada item de cada tarefa e o número de pontuação, ao número de vezes que ele obteve uma nota. Isto significa, que esse examinando recebeu para as duas tarefas com cinco itens cada uma, de dois avaliadores 20 pontuações (2 tarefas $\times$ 5 itens $\times$ 2 avaliadores = 20). Já o examinando de número 373 possui escore 58, número de pontuação 25 e medida da habilidade de -2,02 *logitos*. Embora o escore desse examinando seja maior do que a do outro, a medida da sua habilidade é menor. Isto porque o escore deste é resultante de número de pontuação 25, ele teve uma das tarefas corrigidas pela terceira vez, recebendo mais 5 pontuações, uma para cada item. Esses fatos justificam a variação do número de pontuação de 10 até 25 para os examinandos destacados nessa tabela, e também o motivo para que um maior escore possa corresponder a uma menor medida.

A precisão das medidas, dada por seus desvios padrão, é apresentada na quarta coluna da Tabela 3. As medidas que foram calculadas com a utilização de um número menor de dados são menos precisas, por exemplo, o examinando que elaborou apenas uma tarefa e foi pontuado por dois avaliadores obteve 10 pontos, o indivíduo de número 1770 (primeiro bloco) tem essa condição e a estimativa de sua habilidade tem precisão 0,43 *logitos*, enquanto o examinando que elaborou as duas tarefas, uma delas pontuada por dois avaliadores e a outra por três, obteve 25 pontos e a medida de sua habilidade é mais precisa. Os examinandos de números 373 e 765 (primeiro bloco) obtiveram 25 pontos e a estimativa de suas habilidades tem precisão de 0,27 *logitos*. Na última coluna consta a classificação que cada examinando obteve entre os 350 indivíduos do grupo. O primeiro da lista (n°. 2032) foi o que se saiu pior no exame, ficando na última posição, enquanto o de número 17916, o mais bem sucedido do teste, obteve a primeira colocação.

A média justa varia de 2,19 *logitos*, para o examinando de menor habilidade, até 5,86 *logitos*, para o de maior habilidade, isso significa que, após a correção de erros das medidas, a habilidade dos examinandos varia cerca de 3,7 pontos da escala. Como a escala aplicada nesse experimento é de seis pontos, de 1 a 6, essa variação equivale a quase 4/5 de sua extensão. Variações grandes na escala para a habilidade dos examinandos são esperadas.

A Tabela 4 traz um resumo das medidas da habilidade dos mesmos

examinandos exibidas na Tabela 3, mas dessa vez calibrados segundo o modelo multifacetas de escala gradual de quatro facetas.

Tabela 4 – Resumo das medidas dos examinandos – Modelo: Escala gradual de quatro facetas

Examin.	Escore	Número pontuação	Habilid.	Erro padrão	MQ *Infit*	MQ *Outfit*	Média observ.	Média justa	Classif.
3023	44	20	-2,63	0,31	0,61	0,60	2,20	2,07	350
2617	44	20	-2,43	0,31	0,82	0,80	2,20	2,17	349
373	58	25	-2,31	0,27	0,97	0,95	2,32	2,23	348
765	59	25	-2,24	0,27	0,76	0,77	2,36	2,27	347
2516	48	20	-2,16	0,3	0,28	0,27	2,40	2,32	345
1770	22	10	-2,06	0,44	1,16	1,13	2,20	2,37	344
...	...	...	...	...	...	...	...	...	...
2477	52	20	-1,64	0,29	0,99	1,02	2,60	2,61	332
596	71	20	-0,06	0,29	1,43	1,53	3,55	3,61	173
2228	71	20	-0,04	0,29	0,74	0,80	3,55	3,62	171
2305	93	25	-0,02	0,27	0,88	0,92	3,72	3,63	169
2582	92	25	0,00	0,27	1,61	1,57	3,68	3,64	168
25	72	20	0,01	0,29	0,54	0,54	3,60	3,64	167
302	101	25	0,72	0,28	4,39	4,39	4,04	4,02	90
...	...	...	...	...	...	...	...	...	...
36	84	20	1,57	0,32	1,33	1,33	4,20	4,43	51
2007	109	25	1,59	0,30	0,89	0,87	4,36	4,44	50
2192	91	20	1,60	0,34	0,51	0,50	4,55	4,44	49
1675	87	20	1,62	0,33	0,37	0,37	4,35	4,45	48
945	85	20	1,68	0,33	1,11	1,14	4,25	4,47	46
...	...	...	...	...	...	...	...	...	...
17652	111	20	5,03	0,42	0,72	0,69	5,55	5,69	8
23499	112	20	5,21	0,43	0,92	0,99	5,60	5,73	5
9666	114	20	5,61	0,47	0,79	0,72	5,70	5,80	4
34425	116	20	6,13	0,55	0,92	0,84	5,80	5,87	3
35883	116	20	6,13	0,55	0,92	0,84	5,80	5,87	2
17916	117	20	6,46	0,62	0,95	0,87	5,85	5,91	1

Fonte: Dados da pesquisa

Observam-se variações significativas nas medidas de alguns examinandos. O examinando de menor habilidade nesse caso, é o de número 2032, com medida -2,63, cerca de 0,40 *logitos* mais baixa do que a obtida pelo modelo de duas facetas. A classificação do examinando de maior habilidade, segundo o modelo de duas facetas, o de número 17916, continua sendo a de primeiro lugar segundo o modelo de quatro facetas, entretanto a medida

de sua habilidade foi estimada em cerca de 0,75 *logitos* mais alta do que a estimada com o modelo anterior.

Em termos gerais, a classificação dos examinandos segundo o modelo de escala gradual com quatro facetas foi modificada em relação à classificação obtida com o modelo de duas facetas em mais de 5 posições para 235 examinandos, o que equivale a 67,14% dos 350 indivíduos avaliados. Desses, 46,28% tiveram suas colocações modificadas em mais de 10 posições e 24% em mais de 20 posições.

Os examinandos que possuem habilidades intermediárias sofreram modificações maiores em suas classificações do que aqueles com habilidades extremamente altas ou extremamente baixas. Esse fato pode ser observado comparando-se a última coluna das Tabelas 3 e 4. Por exemplo, o examinando de número 2305, no segundo bloco, estava na posição 147 e passou para a posição 169, isso significa que, se for considerado o desempenho dos avaliadores, a colocação desse indivíduo passa a ser 22 posições abaixo. Já o examinando de número 36 estava na posição 69 e passou para a posição 51, subindo 18 posições em relação aos 350 indivíduos do grupo.

Outras análises sobre a faceta habilidade dos examinandos, feitas na sequência, são resultantes das estimativas dos parâmetros segundo o modelo multifacetas de Rasch de escala gradual de quatro facetas.

A Tabela 5 exibe um resumo do número de examinandos e porcentagens cujas medidas *MQ–infit* se enquadram em cada uma das categorias sugeridas por Wright e Linacre (1994) (Quadro 13) para que as medidas sejam satisfatórias para a construção do sistema de medição.

Tabela 5 – Resumo das estatísticas de ajuste (*infit*) para os examinandos – Modelo de escala gradual de quatro facetas

Média quadrática (MQ–*infit*)	Número de examinandos	Porcentagem de examinandos
>2,0	3	0,86%
1,5 – 2,0	37	10,57%
0,5 – 1,5	284	81,14%
<0,5	26	7,4%

Fonte: Dados da pesquisa

Observa-se nessa tabela que três entre os 350 examinandos avaliados (cerca de 0,86%) tiveram índices MQ–*infit* superiores a 2, evidenciando a presença de desajustes em suas pontuações e indicando que o sistema de medição pode estar distorcido ou degradado, sendo necessárias outras investigações para o entendimento correto sobre a habilidade desses examinandos, a

pontuação a eles atribuídas e os fatores que geraram esses desajustes. No entanto, a maior parte dos 350 examinandos possui os valores MQ–*infit* dentro dos padrões sugeridos por Wright e Linacre (1994) como evidência de que as medidas são produtivas. Apenas 37 indivíduos (10,57%) possuem a média quadrática *infit* no intervalo entre 1,5 e 2,0, o que não é muito bom, mas essas medidas não são degradantes para o sistema de medição, e 284 pessoas, cerca de 81,14%, dentro do intervalo entre 0,5 e 1,5, evidenciando que as medidas estão de acordo com o modelo. Entre os 350 examinandos, 26 (cerca de 7,4%) deles tiveram índices MQ–*infit* inferior a 0,5. Esses resultados sugerem que esses examinandos podem ter recebido pontuações muito semelhantes ou idênticas em todos os 10 itens do teste.

A Tabela 6 traz um resumo das medidas que possuem os maiores valores de *MQ–Infit*. Esses dados consistem nos 0,86% problemáticos, com *MQ–Infit*>2.

Tabela 6 – Maiores valores de *MQ–Infit*

		Número		Erro	MQ		Média	
Examin.	Escore	pontuação	Habilidade	padrão	*Infit*	*Outfit*	observ.	justa
302	101	25	0,72	0,28	4,39	4,39	4,04	4,02
1226	68	20	-0,66	0,29	2,53	2,51	3,40	3,24
2494	119	25	2,20	0,31	2,39	2,36	4,76	4,70

Fonte: Dados da pesquisa

O examinando de número 302 tem os maiores valores para *MQ–Infit* = 4,39 e *MQ–Outfit* = 4,39. Esses valores são muito maiores do que os valores esperados (1,0). Segundo Wright e Linacre (1994), para valores assim, deve haver mais "ruído" do que informações estatísticas úteis. Linacre (2014a) recomenda um olhar atento para a tabela de respostas não esperadas, uma vez que os valores dessa tabela podem corresponder a valores grandes das estatísticas de ajuste. Desse modo, é possível localizar os dados problemáticos. O critério nessa aplicação prática para as respostas serem consideradas não esperadas é que o valor absoluto do residual padronizado |*MQZ*| seja maior ou igual a três. Desse modo, foram detectadas 18 respostas não esperadas pelo modelo de escala gradual de duas facetas, 23 pelo modelo de escala gradual de quatro facetas e 40 pelo modelo de crédito parcial de quatro facetas, o que equivale a 0,2%, 0,3% e 0,52% do total de respostas válidas, respectivamente.

Por meio desses dados, é possível obter algumas informações sobre os examinandos, os itens e os avaliadores. Por exemplo, se o valor do residual padronizado *MQZ* é muito alto e positivo, significa que o indivíduo se saiu melhor do que o valor esperado, isto é, melhor do que a sua capacidade per-

mite, indicando que ele pode ter acertado a resposta ao acaso. Se o valor for alto e negativo, isso indica que era esperada uma pontuação maior do que a obtida por ele. Nesse caso é necessária uma análise cuidadosa, pois o problema pode ter sido causado pelo item, pelo avaliador ou mesmo por alguma variável externa ao instrumento ou à avaliação.

Para exemplificar, na Tabela 7, são exibidas as 23 respostas não esperadas calculadas pelo modelo multifacetas de quatro facetas para escala gradual.

Tabela 7 – Respostas não esperadas – Modelo multifacetas: escala gradual

Examinando	Categoria	Escore	V. esperado	Residual	*MQZ*	Tarefa	Item	Avaliador
2494	2	2	4,7	-2,7	-4,2	49	2	26
302	1	1	3,9	-2,9	-4,1	50	3	57
768	1	1	3,8	-2,8	-3,9	50	3	56
1027	1	1	3,8	-2,8	-3,8	50	3	36
757	2	2	4,4	-2,4	-3,5	50	3	10
2582	1	1	3,6	-2,6	-3,4	49	4	1
1137	5	6	3,4	2,6	3,3	49	3	58
1156	2	2	4,3	-2,3	-3,3	50	1	64
1332	2	2	4,2	-2,2	-3,3	50	1	30
922	6	6	3,6	2,4	3,2	50	3	36
2083	6	6	3,5	2,5	3,2	50	4	58
2089	2	2	4,2	-2,2	-3,2	50	5	58
2381	2	2	4,2	-2,2	-3,2	50	1	18
2807	2	2	4,2	-2,2	-3,2	49	1	18
718	2	2	4,1	-2,1	-3,1	49	1	3
1027	2	2	4,2	-2,2	-3,1	49	3	2
1923	2	2	4,2	-2,2	-3,1	49	3	31
2224	2	2	4,1	-2,1	-3,1	50	1	24
2484	2	2	4,2	-2,2	-3,1	50	3	17
58	5	5	2,6	2,4	3,0	49	1	49
1690	1	1	3,4	-2,4	-3,0	49	3	57
2257	3	3	4,9	-1,9	-3,0	49	1	38
2622	2	2	4,1	-2,1	-3,0	49	3	58

Fonte: Dados da pesquisa

O maior residual padronizado em valor absoluto é de 4,2 e ocorre para o examinando de número 2494. É a observação mais diferente do esperado nesses dados. O escore obtido pelo examinando é muito pequeno (2) em

comparação com o valor esperado (4,7). O sinal negativo de *MQZ* indica que ele se saiu pior do que era esperado. O indivíduo n°. 2494 corresponde a uma medida de 2,39 para *MQ–Infit* (Tabela 6), confirmando que deve haver problemas relacionados com esses dados ou com alguma situação deflagrada durante os procedimentos do teste. Já o examinando de número 1137 possui residual padronizado 3,3. O escore obtido por ele no item 3 da tarefa 49 foi 6, muito maior do que o valor esperado de apenas 3,4. Esse fato indica algum tipo de problema com esse escore, como, por exemplo, a resposta foi copiada do colega, ele acertou a resposta ao acaso ou então os avaliadores pontuaram erradamente.

5.3 CONFIABILIDADE ENTRE AVALIADORES

As análises nesta seção referem-se à qualidade da pontuação atribuída pelos avaliadores às tarefas elaboradas pelos examinandos, desse modo, essas análises referem-se à faceta Avaliadores. Para tanto são feitos estudos com o intuito de detectar tendências dos avaliadores em pontuações sistemáticas que podem causar uma gama de diferentes tipos de erros nas classificações dos examinandos. Neste trabalho busca-se a identificação de erros nas pontuações causados por quatro dessas tendências: efeito de severidade/complacência, efeito de tendência central, efeito de aleatoriedade e efeito de halo.

A busca e a identificação dos avaliadores com essas tendências são feitas de acordo com os Quadros 15 a 18 que fornecem resumos dos indicadores estatísticos normalmente utilizados para essa finalidade no contexto do modelo multifacetas de Rasch. Esses estudos são feitos tanto no nível de grupo quanto no nível individual.

5.3.1 Estudos no nível de grupo

A forma como os avaliadores utilizaram cada uma das categorias da escala (Tabela 8) demonstra que, em geral, os avaliadores, como um grupo, não mostraram tendências de severidade nem de complacência, pois não há uso excessivo das categorias dos extremos da escala (MYFORD; WOLFE, 2004). As categorias que receberam maior número de observações foram as categorias do centro da escala de habilidades.

Nota-se que parece não haver uma tendência generalizada para se ca-

racterizar o efeito da tendência central. A distribuição das pontuações ocorre de maneira espelhada em todas as categorias, sendo menores nas categorias dos extremos 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).

Tabela 8 – Resumo da utilização das categorias da escala de avaliação

Categoria					
1	2	3	4	5	6
1%	13%	28%	37%	16%	4%

Fonte: Dados da pesquisa

O teste do qui-quadrado com hipótese nula de que as medidas de severidade dos avaliadores não são significativamente diferentes (ou seja, que todos os avaliadores compartilham a mesma medida quanto à severidade, após a correção do erro de medição) indica resultados estatisticamente significativos com o valor do qui-quadrado em 785,3 com 43 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.

No entanto, segundo Myford e Wolfe (2004), é importante ressaltar que o teste do qui-quadrado corrigido para o avaliador é muito sensível ao tamanho da amostra. Em muitas aplicações do modelo MFR, o teste qui--quadrado pode ser estatisticamente significativo, mesmo que a variação real entre avaliadores quanto aos níveis de severidade seja pequena.

O teste qui-quadrado com hipótese nula de que todos os examinandos possuem o mesmo nível de desempenho tem valor qui-quadrado de 6697,1 com 349 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. Esse fato sugere que não existe um efeito de tendência central no nível de grupo para os avaliadores. Os resultados também indicam não haver evidências de efeito de aleatoriedade no nível do grupo.

A taxa de separação dos avaliadores, dada por G_H (eq. (113)), é um índice que indica a dispersão das medidas da severidade dos avaliadores em relação à precisão dessas medidas. O valor para esse índice de 3,88 significa que as diferenças entre os níveis de severidade dos avaliadores são quase quatro vezes maiores do que o erro dessas medidas, não sugerindo um efeito de tendência central no nível de grupo para esses avaliadores.

O índice de separação dos avaliadores (eq. (114)) é um indicador do

número de níveis estatisticamente diferentes nos quais os avaliadores estão distribuídos quanto aos seus níveis de severidade. Especificamente esse índice representa a variação "verdadeira" em unidades da variância do erro. Na Tabela 1, esse índice é denominado *estrato* e seu valor é de 5,5, isso sugere que há cerca de cinco e meio estratos estatisticamente diferentes de níveis de severidade entre os avaliadores do grupo.

Analisando o índice de separação para os examinandos (estrato), pode-se verificar se a pontuação sofre o efeito de tendência central no nível de grupo, já que esse índice indica o número de níveis estatisticamente distintos para o desempenho dos examinandos. Esse índice para os examinandos é de 7,01, sugerindo que há cerca de sete estratos estatisticamente diferentes para o desempenho dos examinandos. Portanto não há indícios de efeito de tendência central nem de aleatoriedade para o grupo de avaliadores.

A confiabilidade do índice de separação dos avaliadores fornece informações sobre a forma como os avaliadores são separados quanto aos seus níveis de severidade. É uma medida da difusão das medidas da taxa de separação dos avaliadores (G_H) em relação à precisão dessas medidas e reflete as variações indesejadas entre os níveis de severidade dos avaliadores. O valor da confiabilidade de separação é de 0,94. Isso sugere que, em média, os avaliadores dessas tarefas exercem níveis de severidade significativamente muito diferentes. O ideal é que os valores desse índice sejam pequenos, perto de zero, sugerindo que os avaliadores podem ser intercambiáveis, exercendo níveis de severidade semelhantes (ENGELHARD, 2013; MYFORD; WOLFE, 2004; ENGELHARD; MYFORD, 2003).

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 entre os padrões de desempenho. A confiabilidade de separação dos examinandos é 0,96. Um índice que assume valores entre 0 e 1 demonstra que os avaliadores puderam distinguir de forma confiável entre os níveis de desempenhos avaliados. Portanto, esse indicador não sugere um efeito tendência central nem de aleatoriedade para o grupo de avaliadores.

Teste qui-quadrado para os itens em 6697,1 com $p<0,05$ significa que os itens são significativamente diferentes em termos de suas dificuldades, não sugerindo efeito de halo para os avaliadores. As estatísticas elevadas da taxa de separação, confiabilidade do índice de separação e estrato confirmam que não há indícios de tendência a efeito de halo no nível de grupo para os avaliadores.

Com essas observações, pode-se concluir que esse grupo de avaliadores não apresenta tendências aos efeitos de severidade/complacência, central, halo ou aleatoriedade no nível de grupo.

5.3.2 Estudos no nível individual

Os resultados detalhados das medidas de cada avaliador individualmente calibradas de acordo com o modelo multifacetas de Rasch de crédito parcial, no qual a estrutura de categorias da escala de classificação varia com os itens (equação (11)), são apresentados na Tabela 9.

Os dados dessa tabela são ordenados de acordo com a medida da severidade dos avaliadores, do mais severo para o mais complacente. Na coluna à direita da severidade, é informado o erro padrão, isto é, indica a precisão com que cada uma dessas medidas foi estimada. Essa medida varia dependendo do número de pontuações (segunda coluna) no qual as medidas são baseadas, quanto maior o número de pontuações utilizadas para uma estimativa, menor seu erro padrão. Para exemplificar, o avaliador de número 65 pontuou 350 itens (68 tarefas com 5 itens cada uma), a medida de sua severidade, -0,07 *logitos*, foi estimada com a maior precisão do grupo (0,07). O avaliador de número 53 pontuou apenas 3 tarefas, com cinco itens cada uma, e obteve a medida da severidade (0,71 *logitos*) com menor precisão do grupo (0,34).

Geralmente as estatísticas de ajuste indicam o grau com que as classificações observadas estão de acordo com as classificações esperadas geradas pelo modelo. As medidas das médias quadráticas, expostas nas colunas 5 e 6, calculadas de acordo com as equações (98) e (99), 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. Normalmente, para avaliar o ajuste do modelo, as medidas *infit* são consideradas mais importantes do que as medidas *outfit* (LINACRE, 2014a; MYFORD; WOLFE, 2002).

A maioria dos avaliadores (86%) teve suas medidas *infit* no intervalo entre 0,5 e 1,5, produtivo para as medidas. Apenas 6,8% delas estão entre 1,5 e 2,0, com dados improdutivos para a construção da medida, mas não degradante, e 4,5% dessas medidas são menores do que 0,5. Entretanto, nenhum avaliador teve medida *infit* maior do que 2, o que indicaria que o sistema de medição poderia estar degradado (Quadro 13) (WRIGHT; LINACRE, 1994).

Os avaliadores de números 18, 27, 36, 40, 49 e 58 possuem as medidas *infit* e *outfit* entre 1,31 e 1,73 *logitos*, as maiores do grupo. As pontuações provenientes desses avaliadores mostram-se inconsistentes, sugerindo

que eles podem não ter sido capazes de diferenciar de forma confiável entre os níveis de desempenho, em vez disso, esses avaliadores podem ter atribuído pontuações aleatórias para muitos examinandos.

Já os avaliadores de números 42, 44, 47, 51, 53 e 66 são os que possuem os menores valores para as médias *infit* e *outfit* do grupo, que estão entre 0,43 e 0,75 *logitos*, sugerindo que esses avaliadores podem apresentar tendência central ou de halo. A tendência central ocorre quando o avaliador atribui pontuações próximas do ponto médio para muitos examinandos, já a tendência de halo ocorre quando os avaliadores não são capazes de diferenciar de forma confiável entre traços conceitualmente distintos, atribuindo pontuações semelhantes a uma variedade de traços para muitos examinandos (MYFORD; WOLFE, 2004; ECKES, 2011). Mesmo assim, é importante salientar que medidas da MQ significativamente menores do que 1 não necessariamente indicam a ocorrência desses efeitos nas pontuações, para tanto são necessárias outras análises.

A sétima e a oitava colunas da Tabela 9 referem-se às medidas denominadas média observada, que é a pontuação média atribuída por cada avaliador, e média justa, que é a pontuação média de cada avaliador ajustada para o desvio da média dos avaliadores da amostra. Ao comparar as médias justas dos avaliadores, é possível identificar quais possuem uma tendência em utilizar as escalas de avaliação de uma forma diferente dos outros avaliadores, ou seja, atribuem as pontuações, em média, superiores ou inferiores aos outros avaliadores.

Tabela 9 – Medidas dos avaliadores – Modelo: crédito parcial

	Número		Erro	MQ		Média		Correlação
Avaliador	pontuação	Sever.	padrão	*Infit*	*Outfit*	observ.	justa	Pt bisser.
27	50	0,94	0,18	1,53	1,65	2,96	3,22	0,43
50	85	0,85	0,14	0,92	0,90	3,08	3,26	0,52
49	255	0,78	0,08	1,45	1,46	2,97	3,31	0,63
53	15	0,71	0,34	0,49	0,50	2,87	3,35	0,69
36	150	0,67	0,11	1,41	1,49	3,25	3,38	0,54
67	285	0,59	0,10	1,01	1,00	4,89	3,43	0,50
43	205	0,56	0,09	0,73	0,72	3,10	3,44	0,70
40	100	0,51	0,13	1,51	1,51	3,08	3,47	0,62
66	300	0,49	0,08	0,64	0,65	3,61	3,48	0,58
6	180	0,47	0,10	0,84	0,88	3,17	3,50	0,67
59	45	0,46	0,19	1,06	1,01	2,96	3,51	0,70
68	280	0,44	0,10	0,91	0,93	4,93	3,51	0,67
51	180	0,36	0,10	0,71	0,72	3,32	3,57	0,59

continua

continuação

Avaliador	Número pontuação	Sever.	Erro padrão	MQ *Infit*	MQ *Outfit*	Média observ.	Média justa	Correlação Pt bisser.
24	195	0,32	0,09	1,05	0,99	3,29	3,58	0,43
3	300	0,23	0,08	1,26	1,27	3,42	3,64	0,51
28	200	0,22	0,09	1,05	1,04	3,64	3,65	0,77
64	205	0,21	0,09	0,94	0,94	3,43	3,63	0,56
57	185	0,20	0,10	1,23	1,23	3,37	3,65	0,57
18	145	0,09	0,11	1,38	1,41	3,53	3,71	0,64
37	145	0,01	0,11	0,86	0,84	3,53	3,75	0,72
56	100	0,00	0,13	1,12	1,20	3,29	3,74	0,64
42	55	-0,06	0,20	0,47	0,47	3,98	3,77	0,66
65	340	-0,07	0,07	0,90	0,94	3,55	3,80	0,43
47	295	-0,08	0,08	0,66	0,67	3,55	3,79	0,70
1	250	-0,09	0,09	1,12	1,12	3,57	3,79	0,58
38	195	-0,12	0,10	1,21	1,21	3,69	3,82	0,50
7	50	-0,13	0,20	0,86	0,89	3,82	3,85	0,50
2	180	-0,15	0,10	1,05	1,04	3,75	3,85	0,59
48	210	-0,16	0,09	0,84	0,85	3,60	3,84	0,62
29	235	-0,22	0,09	0,87	0,87	3,38	3,89	0,51
30	90	-0,25	0,14	1,11	1,11	3,60	3,89	0,56
17	140	-0,31	0,11	0,91	0,92	3,70	3,92	0,75
21	205	-0,34	0,09	0,80	0,80	3,57	3,93	0,83
33	135	-0,38	0,12	0,69	0,64	3,69	3,96	0,75
26	265	-0,39	0,08	0,96	1,01	3,76	3,97	0,63
14	100	-0,42	0,14	1,04	1,04	4,03	3,97	0,75
31	180	-0,47	0,10	1,07	1,06	3,83	4,01	0,41
54	205	-0,48	0,10	1,07	1,09	3,99	4,01	0,75
44	155	-0,49	0,11	0,73	0,75	3,90	3,98	0,80
58	190	-0,50	0,10	1,71	1,71	3,90	4,03	0,56
32	149	-0,70	0,12	1,16	1,15	4,22	4,12	0,30
19	215	-0,84	0,10	1,14	1,14	3,97	4,17	0,69
10	105	-0,89	0,14	0,93	0,98	3,83	4,17	0,61
41	55	-1,53	0,20	0,96	0,90	4,33	4,53	0,42

Fonte: Dados da pesquisa

5.3.2.1 Efeito de tendência de severidade e complacência

No nível individual, para verificar se existe alguma evidência quanto aos efeitos de severidade ou complacência, os pesquisadores Myford e Wolfe (2004) (Quadro 15) sugerem primeiramente uma análise visual do mapa das

variáveis, modelo de escala gradual (Fig. 10) para verificar a distribuição das medidas da severidade dos avaliadores ao longo da escala de habilidades. Em seguida, devem-se analisar as medidas da severidade dos avaliadores para perceber se existem avaliadores com medidas de severidade muito diferentes da média das medidas dos avaliadores.

No mapa das variáveis, verifica-se que os avaliadores estão distribuídos de acordo com o seu grau de severidade aproximadamente no intervalo entre -1,5 e 1,0 *logitos*. O avaliador considerado o mais complacente do grupo obteve medida -1,53 *logito*, com precisão de 0,20 *logitos*, e foi o único avaliador com medida de severidade menor do que -1,0 *logitos*, indicando que ele atribuiu pontuações, em média, mais elevadas do que os outros avaliadores do grupo. O avaliador mais severo obteve medida de 0,94, com precisão de 0,18 *logitos*. Esse avaliador atribuiu pontuações, em média, menores do que os outros avaliadores do grupo.

A comparação entre os níveis médios de severidade dos avaliadores pode não ser suficiente para determinar se um avaliador é mais severo ou mais complacente do 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 por esse avaliador são mais baixas. O mesmo pode ocorrer para o avaliador cuja média das pontuações é mais alta comparada às médias dos outros avaliadores do grupo. Esse fato não é suficiente para estabelecer que este avaliador é mais complacente. Para isso, é necessário determinar se a habilidade dos examinandos, cujas tarefas ele pontuou, não eram, de fato, maiores do que a habilidade dos outros examinandos da amostra.

No contexto do modelo multifacetas de Rasch, pode-se ter acesso às médias justas para cada avaliador. Essa média ajusta a média observada 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. Ao comparar as médias justas dos avaliadores, podem-se identificar os avaliadores com tendência a utilizar as escalas de classificação de uma forma mais severa ou mais branda em comparação com os outros avaliadores do grupo.

Na Tabela 9, o avaliador de número 27, o mais severo do grupo, teve uma média justa de 3,22 *logitos*, enquanto o de número 41, o mais complacente, obteve uma média justa de 4,53 *logitos*. Isso sugere que o avaliador de número 27 atribuiu pontuações, em média, 1,31 pontos menores do que o avaliador de número 41. A diferença entre os níveis de severidade dos dois

avaliadores é maior do que uma categoria da escala.

O modelo multifacetas de escala de crédito parcial, no qual cada avaliador utiliza a sua própria estrutura de escala (equação (17)), permite verificar a frequência com que cada avaliador utilizou as categoria da escala de para verificar se há uso excessivo das categorias dos extremos da escala.

A Tabela 10 exibe a frequência com que os três avaliadores, considerados os mais severos, utilizaram cada uma das categorias da escala de classificação (segunda coluna). Nessa tabela são informadas também, para cada categoria, as médias observadas e esperadas (terceira e quarta colunas), a média quadrática *outfit* (quinta coluna), a locação ($c_{hk} = c_h + d_{hk}$) e as distâncias entre uma locação e a locação anterior ($|c_{hk} - c_{h(k-1)}|$) (sexta e sétima colunas).

Tabela 10 – Estatísticas do uso das categorias: Avaliadores portadores de tendência de severidade

Categ.	CONTAGEM Categoria	MÉDIA Observ.	Esper.	MQ *Outfit*	c_{hk}	$\|c_{hk} - c_{h(k-1)}\|$
Avaliador número 27						
1	5 (10%)	-0,94	-1,27	1,1		
2	12 (24%)	-0,82	-0,81	1,0	-1,92	
3	14 (28%)	-0,28	-0,37	1,1	-0,74	1,18
4	18 (36%)	-0,16	0,01	1,6	-0,43	0,31
5	1 (2%)	0,75	0,37	0,9	3,08	3,51
6						
Avaliador número 49						
1	23 (9%)	-2,22	-2,02	0,9		
2	91 (36%)	-1,47	-1,58	1,1	-3,19	
3	49 (19%)	-1,14	-1,06	1,5	-0,70	2,49
4	57 (22%)	-0,19	-0,37	1,0	-0,89	0,19
5	32 (13%)	0,58	0,92	1,4	0,80	1,69
6	3 (1%)	2,15	2,21	1,1	3,99	3,19
Avaliador número 50						
1						
2	27 (32%)	-1,61	-0,39	0,7		
3	31 (36%)	-0,47	-0,72	1,1	-1,18	
4	20 (24%)	-0,17	-0,15	0,8	0,01	1,19
5	7 (8%)	0,15	0,39	1,2	1,18	1,17
6						

fonte: Dados da pesquisa

O avaliador número 27 praticamente não utilizou as categorias mais

altas da escala, 5 e 6, atribuindo 98% das pontuações às categorias 1, 2, 3 e 4. As médias observadas desse avaliador variaram entre -0,94 e 0,75. Conforme foi destacado anteriormente, quanto menores as pontuações atribuídas pelo avaliador, maior é a medida da severidade. O avaliador de número 49 atribuiu 64% às três primeiras categorias e suas médias observadas variaram de -2,22 a 2,15, referindo-se as médias maiores à pontuação de apenas 14% dos itens nas duas categorias mais altas da escala. O avaliador de número 50 atribuiu 92% de suas pontuações às categorias 2, 3, e 4 e não utilizou a pontuação mais alta da escala.

A Tabela 11 refere-se aos avaliadores considerados mais complacentes, que atribuíram pontuações, em média, mais altas do que os outros avaliadores do grupo.

Tabela 11 – Estatísticas do uso das categorias: Avaliadores portadores de tendência de complacência

Categ.	CONTAGEM Categoria	MÉDIA Observ.	MÉDIA Esper.	MQ *Outfit*	c_{hk}	$\lvert c_{hk} - c_{h(k-1)} \rvert$
Avaliador número 19						
1	1 (0%)	-2,31	-1,50	0,6		
2	20 (9%)	-1,13	-0,99	0,8	-4,24	
3	38 (18%)	-0,54	-0,46	0,7	-1,37	2,87
4	85 (40%)	0,36	0,17	1,2	-0,97	0,40
5	67 (31%)	1,09	1,13	1,2	0,83	1,80
6	4 (2%)	5,12	5,38	1,1	5,76	4,93
Avaliador número 32						
1						
2	13 (9%)	-1,51	-1,47	0,9		
3	20 (13%)	-1,12	-0,99	0,6	-1,67	
4	49 (33%)	-0,25	-0,36	1,2	-1,59	0,08
5	55 (37%)	0,64	0,72	1,0	0,00	1,59
6	12 (8%)	2,97	2,80	0,8	3,26	3,26
Avaliador número 41						
1						
2	2 (4%)	-1,49	-0,97	0,6		
3	10 (18%)	-0,69	-0,71	0,9	-2,46	
4	17 (31%)	-0,35	0,28	0,6	-1,05	1,41
5	20 (36%)	0,56	0,55	0,7	-0,09	0,96
6	6 (11%)	5,41	5,11	0,8	3,59	3,68

fonte: Dados da pesquisa

O avaliador de número 19 atribuiu apenas 9% de sua pontuação às categorias 1 e 2, restando 91% da pontuação às categorias mais altas da escala. O mesmo pode ser observado quanto às pontuações do avaliador de número 32. Já o avaliador de número 41 atribuiu 78% de suas pontuações às três categorias mais altas da escala (4, 5 e 6). As médias observadas para esses avaliadores são mais elevadas do que as médias observadas para os avaliadores mais severos.

As análises da frequência com que cada avaliador utilizou cada categoria da escala de classificação confirmam a tendência desses avaliadores em pontuações sistematicamente severas ou brandas.

5.3.2.2 Efeito de tendência central

De acordo com Myford e Wolfe (2004), quando os avaliadores possuem medidas média quadrática, *infit* e *outfit*, significativamente diferentes de 1,0, eles podem apresentar tendência central.

Mesmo assim, esses pesquisadores orientam cautela na determinação da tendência de efeito central ao interpretar os índices de ajuste. A tendência central frequentemente está associada a medidas MQ menores do que 1. No entanto, algumas vezes os índices de ajustes para o avaliador que exibe essa tendência poderão ser maiores do que 1. Os pesquisadores 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.

Quando se utiliza o modelo de escala de crédito parcial (equação (17)), é possível verificar como cada avaliador atribuiu as pontuações. Os avaliadores de números 29, 42, 44, 47, 48 e 66 estão entre os que possuem índices das médias quadráticas menores do que 0,88, sugerindo que as pontuações atribuídas por eles diferem pouco dos valores esperados para essas pontuações. Analisando a frequência com que eles utilizaram cada uma das categorias, pode-se concluir quais apresentam tendência em pontuações nas categorias centrais da escala. A frequência com que esses avaliadores atribuíram as pontuações são exibidas na Tabela 12 juntamente com outros índices que podem auxiliar nas análises.

Na Tabela 12, observam-se alguns avaliadores que apresentam tendência em pontuações nas categorias centrais da escala de avaliação. O avaliador de número 29 utilizou as categorias 3 e 4 em 80% das suas pontuações,

enquanto o avaliador de número 42 atribuiu 100% de suas pontuações às categorias 3, 4 e 5, sendo 58% apenas à categoria 4, o mesmo comportamento do avaliador de número 44, que utilizou as categorias 3 e 4 em suas pontuações 79% das vezes, sendo 50% delas utilizadas na categoria 4. O avaliador de número 47 utilizou as categorias 3 e 4 em 81% de suas pontuações, e o avaliador de número 66 utilizou essas categorias em 82% do total das suas pontuações, o que sugere um efeito de tendência central para esses avaliadores. Já o avaliador de número 51 apresentou as pontuações um pouco mais espalhadas, mas, mesmo assim, apresenta pontuações nas categorias 2, 3 e 4 89% das vezes, provavelmente também apresente efeito de tendência central em suas pontuações, apesar de mais leve.

Tabela 12 – Estatísticas do uso das categorias: Avaliadores portadores de tendência central

Categ.	CONTAGEM Categoria	MÉDIA Observ.	MÉDIA Esper.	MQ *Outfit*	c_{hk}	$\lvert c_{hk} - c_{h(k-1)} \rvert$
Avaliador número 29						
1	3 (1%)	-0,59	-0,66	1,0		
2	29 (12%)	-0,21	-0,19	1,0	-2,70	
3	94 (40%)	0,39	0,35	1,1	-1,10	1,60
4	93 (40%)	0,96	1,01	1,0	0,68	1,78
5	16 (7%)	1,82	1,69	0,9	3,12	2,44
6						
Avaliador número 42						
1						
2						
3	12 (22%)	-1,37	-0,96	0,7		
4	32 (58%)	0,06	-0,02	0,5	-1,44	
5	11 (20%)	0,19	0,75	0,9	1,44	2,88
6						
Avaliador número 44						
1						
2	2 (1%)	-1,68	-1,25	0,8		
3	45 (29%)	-0,88	-0,83	0,9	-4,16	
4	78 (50%)	-0,18	-0,19	1,1	-1,08	3,08
5	27 (17%)	1,04	0,91	0,9	1,36	2,44
6	3 (2%)	2,09	2,42	1,0	3,88	2,52

continua

continuação

Categ.	CONTAGEM Categoria	MÉDIA Observ.	Esper.	MQ *Outfit*	c_{hk}	$\lvert c_{hk} - c_{h(k-1)} \rvert$
Avaliador número 47						
1						
2	29 (10%)	-2,43	-2,30	0,9		
3	102 (35%)	-1,62	-1,56	0,9	-3,16	
4	137 (46%)	-0,86	-0,89	0,9	-1,53	1,63
5	26 (9%)	0,09	-0,11	0,8	1,16	2,69
6	1 (0%)	0,18	0,64	1,2	3,54	2,38
Avaliador número 51						
1	1 (1%)	-0,84	-1,41	1,1		
2	35 (19%)	-0,87	-0,82	0,9	-4,67	
3	70 (39%)	-0,37	-0,18	0,6	-1,21	3,46
4	56 (31%)	0,68	0,57	0,6	0,41	1,62
5	15 (8%)	1,72	1,29	0,5	2,26	1,85
6	3 (2%)	2,42	1,89	0,6	3,21	0,95
Avaliador número 66						
1	1 (0%)	-0,50	-2,01	1,9		
2	20 (7%)	-1,48	-1,15	0,8	-4,57	
3	119 (40%)	-0,33	-0,31	0,9	-2,52	2,05
4	125 (42%)	0,89	0,93	0,8	0,22	2,74
5	26 (9%)	2,59	2,22	0,5	3,18	2,96
6	9 (3%)	3,28	2,99	0,7	3,70	0,52

fonte: Dados da pesquisa

Analisando as pontuações atribuídas por cada avaliador, aparentemente observa-se que 18 deles, aproximadamente 40%, são portadores de tendência central. Considerando que cada um desses avaliadores utilizou as categorias mais baixas (1 e 2) juntamente com as mais altas (5 e 6) em menos do que 30% de suas pontuações, o restante das pontuações, no mínimo 70% delas, foi utilizado nas duas categorias do meio da escala (3 e 4).

Desses 18 avaliadores que atribuíram as suas pontuações nas categorias 3 e 4 mais de 70% das vezes, 14 possuem as medidas *infit* entre 0,43 e 0,93 e 4 deles possuem as medidas *infit* no intervalo entre 1,05 e 1,38.

A locação ou limiar das categorias também são úteis para a detecção ou confirmação de efeito de tendência central para os avaliadores individualmente. Os limiares são os pontos nos quais a probabilidade de atribuir pontuações às categorias adjacentes são iguais (LINACRE, 2014a).

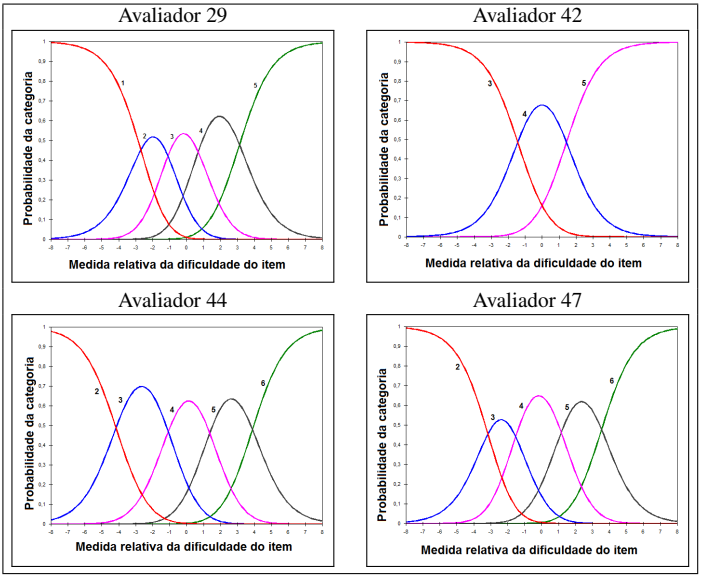
Se um avaliador apresenta um efeito de tendência central, os limiares das categorias na escala de classificação serão dispersos, com pouca utiliza-

ção das categorias nos extremos da escala (MYFORD; WOLFE, 2004). A última coluna da Tabela 12 exibe a locação das categorias e a distância entre duas locações consecutivas. As distâncias entre as locações das categorias adjacentes da escala, resultantes da utilização do modelo de escala gradual (Tabela 19), estão entre 1,16 e 2,49. Observa-se na Tabela 12 que as locações das categorias resultantes das pontuações desses avaliadores estão mais afastadas do que as citadas para a maioria das categorias adjacentes, além disso, eles não utilizam demasiadamente as categorias 1, 2 e 6.

Algumas vezes, além da pouca utilização das categorias nos extremos da escala, pode haver inversão na ordem das categorias para os avaliadores que apresentam efeito de tendência central, isto é, os limiares não aumentam monotonicamente nos extremos da escala. Esse fato não foi constatado para nenhum desses avaliadores.

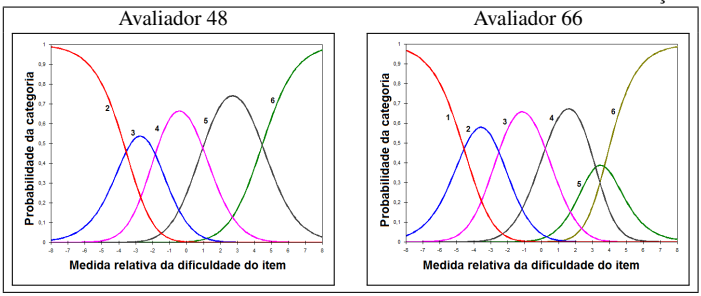
As curvas de probabilidade das categorias para cada um dos avaliadores contêm informações que podem auxiliar na detecção de efeito de tendência central. A Figura 12 exibe o gráfico dessas curvas para alguns avaliadores, obtidos com a utilização do modelo MFR de crédito parcial (equação (17)).

Figura 12 – Curvas de probabilidade das categorias – Modelo: Crédito parcial



continua

continuação



fonte: Linacre (2014b)

A escala de habilidades é dada no eixo horizontal em *logitos*, enquanto a probabilidade de se observar cada uma das categorias é dada no eixo vertical. As curvas são traçadas para cada uma das categorias da escala de classificação. Para a detecção do efeito de tendência central, deve-se olhar se as categorias de escala de classificação estão bastante separadas na escala e se as categorias formam picos distintos em suas curvas. Em geral, quando um avaliador apresenta tendência central, a probabilidade de observar pontuações nas categorias centrais da escala é maior, o que resulta em uma grande separação dos limiares das categorias, especialmente nas categorias do meio da escala (MYFORD; WOLFE, 2004).

A presença de efeito de tendência central para um determinado avaliador também pode ser confirmada por meio da média quadrática *outfit* para as categorias Tabela 12, coluna 5. Alguns desses valores são diferentes do valor esperado (1,0), indicando problemas nas pontuações. Para cada categoria da escala são estimadas duas medidas para o desempenho dos avaliadores, a média observada e a média esperada. A média esperada é o desempenho previsto pelo modelo para o avaliador para cada uma das categorias da escala. Quando as medidas de desempenho observado e esperado estão próximas uma da outra para o avaliador em uma determinada categoria da escala de classificação, o índice *outfit* para essa categoria estará próximo de 1. Quanto maior a discrepância entre as medidas esperadas e observadas para o desempenho do avaliador, maior será o valor do índice *outfit* para a categoria da escala. Desse modo, valores *outfit* consideravelmente maiores do que 1 para as categorias podem sugerir um efeito de aleatoriedade e não de efeito central (MYFORD; WOLFE, 2004).

Os avaliadores de números 18, 27, 36, 40, 49 e 58 possuem índices médias quadráticas maiores do que 1,3, indicando um maior desajuste entre os valores observados e esperados em suas pontuações. Após investigação sobre

a frequência com que esses avaliadores utilizaram cada categoria da escala de classificação, foi possível concluir que eles não apresentam tendência a pontuações nas categorias centrais da escala. Na sequência serão feitas outras análises sobre as pontuações desses avaliadores e sobre os prováveis motivos para os índices de ajuste de suas medidas serem maiores do que 1.

5.3.2.3 Efeito de aleatoriedade

O avaliador portador do efeito de aleatoriedade utiliza a escala de classificação de modo diferente do modo com que os outros avaliadores do grupo a utilizam. Esse avaliador pode ter desenvolvido uma interpretação diferente do significado de uma ou mais categorias da escala em relação aos traços ou, então, o avaliador pode não ser capaz de fazer distinções finas entre os traços avaliados para empregar as categorias da escala adequadamente e atribui as pontuações de forma aleatória e não confiável.

Os avaliadores com índices das médias quadráticas *infit* e *outfit* significativamente maiores do que 1 podem mostrar um efeito de aleatoriedade em suas classificações, uma vez que esses índices indicam o acordo acumulado entre as pontuações observadas e esperadas.

Para eliminar a possibilidade de diagnóstico errado para o efeito de aleatoriedade, uma vez que outras tendências também exibem as estatísticas de ajuste maiores do que um, devem-se comparar as correlações ponto bisserial desses avaliadores com as de outros avaliadores. Se a correlação bisserial de um avaliador é consideravelmente menor do que as dos outros avaliadores, é porque as suas pontuações tendem a ser em uma ordem diferente das pontuações dos outros avaliadores.

Para facilitar a observação, a Tabela 13 reproduz da Tabela 9 as medidas dos avaliadores com maiores valores das médias quadráticas *infit* e *outfit* e que, ao mesmo tempo, possuem as correlações ponto bisserial menores do que as dos outros avaliadores, indicando que esses avaliadores podem apresentar tendência ao efeito de aleatoriedade.

Com a utilização do modelo de crédito parcial, no qual a estrutura da escala pode variar entre os avaliadores, é possível ter acesso aos índices de ajuste que estabelecem a consistência com que cada avaliador utilizou a escala de avaliação para todas as categorias. Os avaliadores que mostram um efeito de aleatoriedade em suas classificações terão as medidas médias quadráticas *infit* e *outfit* significativamente maiores do que 1, sugerindo que eles não foram capazes de diferenciar entre os desempenhos dos examinandos ao

longo da escala de classificação, atribuindo pontuações aparentemente aleatórias para muitos examinandos.

Tabela 13 – Possíveis avaliadores portadores de tendência de aleatoriedade

Avaliador	Número pontuação	Severidade	Erro padrão	MQ *Infit*	MQ *Outfit*	Média observada	Média justa	Correlação bisserial
18	145	0,11	0,11	1,38	1,42	3,53	3,7	0,5
27	50	0,94	0,18	1,46	1,46	2,96	3,2	0,43
36	150	0,62	0,11	1,31	1,31	3,25	3,4	0,56
57	185	0,17	0,10	1,21	1,22	3,37	3,66	0,5
58	190	-0,49	0,10	1,72	1,73	3,9	4,02	0,55
59	45	0,47	0,19	1,10	1,10	2,96	3,49	0,41

fonte: Dados da pesquisa

Para uma análise mais detalhada da identificação de avaliadores portadores do efeito de aleatoriedade, a Tabela 14 apresenta o modo como os avaliadores apontados na Tabela 13 pontuaram as tarefas elaboradas pelos examinandos.

Tabela 14 – Estatísticas do uso das categorias: Avaliadores portadores de tendência de aleatoriedade

Categ.	CONTAGEM Categoria	MÉDIA Observ.	MÉDIA Esper.	MQ *Outfit*	c_{hk}	$\lvert c_{hk} - c_{h(k-1)} \rvert$
Avaliador número 18						
1	5 (3%)	-0,91	-1,09	1,1		
2	16 (11%)	-0,38	-0,72	1,6	-2,07	
3	45 (31%)	-0,42	-0,30	0,9	-1,55	0,52
4	59 (41%)	0,16	0,16	0,9	-0,34	1,21
5	16 (11%)	0,50	0,64	1,2	1,70	2,04
6	4 (3%)	1,45	1,10	0,7	2,26	0,56
Avaliador número 27						
1	5 (10%)	-0,94	-1,27	1,1		
2	12 (24%)	-0,82	-0,81	1,0	-1,92	
3	14 (28%)	-0,28	-0,37	1,1	-0,74	1,18
4	18 (36%)	-0,16	0,01	1,6	-0,43	0,31
5	1 (2%)	0,75	0,37	0,9	3,08	3,51
6	0					

continua

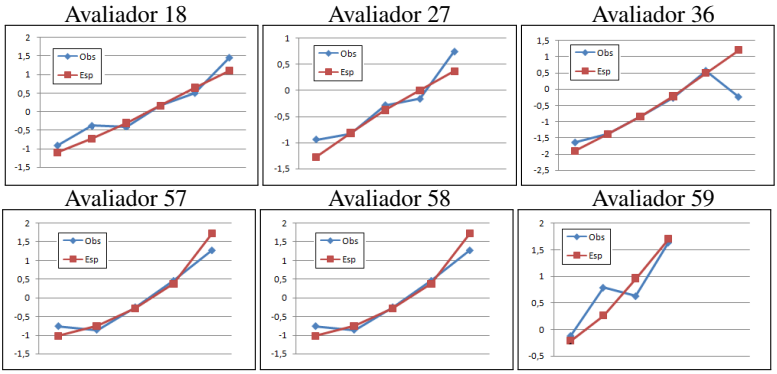
continuação

Categ.	CONTAGEM Categoria	MÉDIA Observ.	Esper.	MQ *Outfit*	c_{hk}	$\lvert c_{hk} - c_{h(k-1)} \rvert$
Avaliador número 36						
1	10 (7%)	-1,63	-1,90	1,6		
2	24 (16%)	-1,37	-1,38	1,0	-2,52	
3	50 (33%)	-0,84	-0,84	1,1	-1,85	0,67
4	51 (34%)	-0,27	-0,22	1,2	-0,56	1,29
5	14 (9%)	0,56	0,50	0,9	1,43	1,99
6	1 (1%)	-0,24	1,20	1,7	3,50	2,07
Avaliador número 57						
1	5 (3%)	0,22	-0,93	2,2		
2	24 (13%)	0,06	-0,46	1,8	-2,27	
3	72 (39%)	-0,01	0,14	0,9	-1,28	0,99
4	65 (35%)	0,93	1,08	1,2	0,68	1,96
5	19 (10%)	2,31	2,16	0,9	2,87	2,19
6	0					
Avaliador número 58						
1						
2	27 (14%)	-0,76	-1,02	1,4		
3	36 (19%)	-0,86	-0,75	0,9	-1,18	
4	73 (38%)	-0,25	-0,28	1,1	-1,24	0,06
5	37 (19%)	0,46	0,38	1,5	0,71	1,95
6	17 (9%)	1,28	1,73	1,4	1,71	1,00
Avaliador número 59						
1	2 (4%)	-0,12	-0,21	1,0		
2	12 (27%)	0,79	0,26	2,7	-1,78	
3	17 (38%)	0,63	0,96	0,7	0,25	2,03
4	14 (31%)	1,64	1,70	1,1	1,54	1,29
5	0					
6	0					

Fonte: Dados da pesquisa

A Figura 13 traz uma ilustração gráfica das médias observadas e esperadas dos avaliadores com comportamento tendencioso para o efeito de aleatoriedade. Observa-se um certo desacordo entre as médias observadas e esperadas para algumas categorias, indicando um nível de aleatoriedade nessas pontuações.

Figura 13 – Médias observadas e esperadas: Tendência de aleatoriedade



Fonte: Autora

Para a identificação de avaliadores portadores de tendências em pontuações sistemáticas, é necessário estabelecer limites para diferenciar comportamentos que serão considerados normais dos considerados anormais. Neste estudo, foram considerados portadores de tendência à aleatoriedade os avaliadores com médias quadráticas maiores do que 1,1 e a correlação ponto bisserial menor do que 0,56. Nesses parâmetros, entre os 44 avaliadores, 6 deles foram considerados portadores da tendência em pontuações aleatórias.

5.3.2.4 Efeito de halo

Se os índices de dificuldade das categorias diferem pouco, então as médias esperadas para as pontuações também diferem pouco. Nesse caso, os avaliadores que exibem a tendência de efeito de halo atribuem pontuações quase iguais para todas as categorias da escala. Consequentemente, as médias observadas e esperadas não diferem muito, o que resulta em valores de médias quadráticas *infit* e *outfit* significativamente menores do que 1. Tal tendência sugere que esses avaliadores não são capazes de diferenciar entre categorias conceitualmente diferentes atribuindo pontuações semelhantes a muitos examinandos. Mesmo assim, índices de ajuste menores do que 1 podem não significar tendência a efeito de halo (MYFORD; WOLFE, 2004).

Alternativamente, quando os índices de dificuldade das categorias variam, as médias esperadas das pontuações mostram uma maior variabilidade. Desse modo, as pontuações dos avaliadores que possuem tendência a efeito

de halo serão muito diferentes das pontuações esperadas, uma vez que essa tendência é caracterizada por pontuações quase iguais para muitos examinandos para cada categoria. Esse fato resultará em índices de ajuste significativamente maiores do que 1,0. Em ambos os casos, devem-se inspecionar as pontuações observadas para os avaliadores que possuem índices de ajuste diferentes de 1 (MYFORD; WOLFE, 2004).

Para determinar se um avaliador possui tendência a efeito de halo, pode-se contar quantas vezes esse avaliador utilizou cada categoria da escala de classificação e, então, deve-se determinar qual é a porcentagem na qual as pontuações desse avaliador 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* competências. Essa análise é fornecida pelo programa *Facets* e indica o grau com que as pontuações elaboradas por um determinado avaliador para determinada competência diferem dos valores esperados pelo modelo.

Esse índice de viés da interação é calculado pela estatística *t-Student* com hipótese de que não há viés maior do que o erro de medição, com número de graus de liberdade igual à contagem observada menos 1. Quando o número de observações é grande, a medida *t-Student* se aproxima de uma distribuição normal com média 0 e desvio padrão 1. A *t-Student* é utilizada para resumir o teste de significância estatística do tamanho do viés. As estatísticas de ajuste não informam sobre a existência ou não de viés, mas elas auxiliam a determinar se os desajustes nos dados são explicados por viés ou por outras causas (LINACRE, 2014a).

A maioria das medidas de viés são pequenas e estatisticamente insignificantes. Consideram-se as medidas de vieses quando o índice t-Student em valor absoluto for maior do que 2.

Neste estudo, a estatística *t-Student* é utilizada para identificar os avaliadores que apresentam alguma inconsistência em suas pontuações. As análises são as seguintes: se a medida *t-Student* é maior do que 2, o avaliador foi mais severo do que o esperado para pontuar a competência determinada; se a medida *t-Student* é menor do que -2, o avaliador foi mais complacente do que o esperado para pontuar a competência particular. Entretanto, para determinar se esse desajuste é resultado de tendência a efeito de halo, é necessário examinar as pontuações observadas e esperadas para os avaliadores identificados pelos índices *t-Student* (LINACRE, 2014a; MYFORD; WOLFE, 2004).

A Tabela 15 traz as medidas dos vieses das interações avaliadores *versus* itens para os avaliadores cujas medidas apresentaram vieses, com índices

t-Student maiores do que 2 em valores absolutos, em pelo menos 3 dos 5 itens. As colunas 1 e 2 da tabela exibem as médias observadas e esperadas, a terceira coluna o número de vezes que o avaliador atribuiu pontuação a cada item e a quarta coluna, a diferença entre as médias observada e esperada dividida pelo número de vezes que o avaliador pontuou cada item.

Tabela 15 – Análise dos vieses

Média		Núm.	(Obs-Esp)/		Erro			MQ			
obs.	esp.	Pont.	N. pont.	Vies	padrão	t-Stud.	Prob.	*Infit*	*Outf.*	Item	dific.
Número do avaliador: 24 Graus de liberdade: 38											
133	145,28	39	-0,31	-0,55	0,21	-2,64	0,0121	1,4	1,4	1	-0,74
130	121,68	39	0,21	0,36	0,21	1,71	0,0945	0,3	0,3	2	0,29
116	127,57	39	-0,3	-0,49	0,21	-2,39	0,0220	1,0	1,0	3	0,04
142	124,26	39	0,45	0,78	0,21	3,64	0,0008	1,2	1,2	4	0,18
121	123,20	39	-0,06	-0,09	0,21	-0,46	0,6517	0,6	0,6	5	0,23
Número do avaliador: 49 Graus de liberdade: 50											
218	173,56	51	0,87	1,64	0,20	8,07	0,0000	1,1	1,1	1	-0,74
149	143,01	51	0,12	0,21	0,18	1,12	0,2700	1,0	1,1	2	0,29
136	150,34	51	-0,28	-0,5	0,19	-2,64	0,0111	1,1	1,1	3	0,04
127	146,2	51	-0,38	-0,68	0,19	-3,54	0,0009	1,3	1,3	4	0,18
128	144,89	51	-0,33	-0,6	0,19	-3,13	0,0030	1,1	1,1	5	0,23
Número do avaliador: 65 Graus de liberdade: 67											
287	268,91	68	0,27	0,54	0,18	3,06	0,0032	0,8	0,9	1	-0,74
247	230,38	68	0,24	0,43	0,16	2,65	0,0099	0,9	0,9	2	0,29
235	240,13	68	-0,08	-0,13	0,16	-0,83	0,4106	1,3	1,3	3	0,04
218	234,67	68	-0,25	-0,42	0,16	-2,67	0,0095	0,6	0,6	4	0,18
220	232,91	68	-0,19	-0,33	0,16	-2,07	0,0426	0,6	0,6	5	0,23
Número do avaliador: 68 Graus de liberdade: 55											
305	290,55	56	0,26	0,94	0,27	3,43	0,0011	0,9	0,9	1	-0,74
283	270,10	56	0,23	0,63	0,23	2,75	0,0080	0,9	1,0	2	0,29
263	275,47	56	-0,22	-0,57	0,21	-2,72	0,0087	0,7	0,7	3	0,04
270	272,48	56	-0,04	-0,11	0,21	-0,53	0,5948	0,6	0,6	4	0,18
259	271,51	56	-0,22	-0,55	0,21	-2,69	0,0095	0,7	0,8	5	0,23

Fonte: Dados da pesquisa

Comparando-se as médias observadas e esperadas do avaliador de número 24, mostradas nas colunas 1 e 2 da Tabela 15, observa-se que esse avaliador atribuiu pontuações mais baixas do que a esperada para os itens 1, 3 e 5 e mais altas do que as esperadas para os itens 2 e 4. Os itens de números 2 e 5 foram os considerados mais difíceis, por isso receberam pontuações mais baixas dos avaliadores, enquanto o de número 1 foi considerado o mais fácil,

por isso recebeu pontuações mais elevadas. Esse avaliador atribuiu pontuações mais baixas ao item mais fácil, o de número 1, e mais alta ao item um pouco mais difícil, o de número 4, enquanto a média das pontuações dos outros avaliadores do grupo era por pontuações mais baixas para esse item. Os itens de número 2 e 5 não apresentam vieses para esse avaliador (p>0,01).

O avaliador de número 49 atribuiu pontuações mais baixas do que as esperadas para os itens 3, 4 e 5 e pontuações mais elevadas do que as esperadas para os itens 1 e 2. A última coluna da Tabela 5 traz as medidas da dificuldade dos itens. Esse avaliador atribuiu pontuações mais elevadas do que era esperado para o item de número 1 e pontuações mais baixas do que a média das pontuações dos outros avaliadores para os itens de números 3, 4 e 5. O item 2 não apresenta viés uma vez que a probabilidade é maior que 0,01.

O avaliador de número 65 atribuiu pontuações mais elevadas do que as esperadas para as competências de números 1, 2 e 3 e mais baixas do que as esperadas para as competências 1 e 2. Esse avaliador atribuiu pontuações elevadas para o item mais difícil (de número 2), enquanto a tendência dos outros avaliadores era por pontuações mais baixas e atribuiu pontuações baixas para um item considerado fácil (de número 3), contrariando a tendência de pontuações elevadas dos outros avaliadores. Esses avaliadores tendem a atribuir pontuações de modo diferente dos outros avaliadores para as mesmas características. Tal fato sugere que esses avaliadores possuem uma tendência a efeito de halo.

5.4 ANÁLISES DOS ELEMENTOS DAS FACETAS TAREFAS E ITENS

A calibração das tarefas e dos itens pelos modelos multifacetas de Rasch de escala gradual e de crédito parcial obtiveram índices semelhantes, diferindo muito pouco. Por isso, alguns dos índices destas facetas serão exibidos apenas para o modelo de escala gradual.

A Tabela 16 relata as medidas da faceta Tarefas segundo o modelo multifacetas de Rasch de escala gradual. As medidas *MQ–Infit* e *MQ–Outfit* são próximas do valor esperado 1,0, indicando que os dados se encaixam no modelo. As tarefas foram consideradas equivalentes em relação à dificuldade, pois a diferença de suas medidas é muito pequena. A de número 50 foi considerada um pouco mais fácil, -0,04 *logitos*, enquanto a de número 49 um pouco mais difícil, 0,04 *logitos*.

As medidas da faceta Tarefas segundo o modelo de crédito parcial não serão expostas por não diferiram significativamente das medidas desta faceta segundo o modelo de escala gradual.

Tabela 16 – Calibração das tarefas – Modelo: Escala gradual

					MQ		Média	
Item	Escore	Pontuação	Dificuldade	Erro padrão	*Infit*	*Outfit*	observada	justa
50	13805	3755	-0,04	0,02	0,97	0,97	3,68	3,76
49	14024	3849	0,04	0,02	1,04	1,03	3,64	3,74

Fonte: Dados da pesquisa

A Tabela 17 apresenta as medidas relacionadas à quarta faceta, Item segundo o modelo de escala gradual. Observa-se que os valores de *MQ–Infit* e *MQ–Outfit* estão todos entre 0,5 e 1,5, indicando que as medidas são produtivas. O item considerado o mais fácil foi o de número 1, com medida -0,74 *logitos*, e o mais difícil, o de número 2, com medida 0,29 *logitos*.

Tabela 17 – Calibração dos itens – Modelo: Escala gradual

					MQ		Média	
Item	Escore	Pontuação	Dificuldade	Erro padrão	*Infit*	*Outfit*	observada	justa
1	6165	1521	-0,74	0,04	1,09	1,07	4,05	4,15
3	5535	1520	0,04	0,03	1,27	1,26	3,64	3,74
4	5420	1521	0,18	0,03	0,99	0,98	3,56	3,66
5	5382	1521	0,23	0,03	0,83	0,84	3,54	3,64
2	5327	1521	0,29	0,03	0,84	0,87	3,50	3,60

Fonte: Dados da pesquisa

A Tabela 18 exibe as medidas relacionadas à faceta Item segundo o modelo de crédito parcial. A ordem dos itens quanto à dificuldade foi preservada, com o item de número 1, considerado o mais fácil e o de número 2 o mais difícil.

Tabela 18 – Calibração dos itens – Modelo: Crédito parcial

					MQ		Média	
Item	Escore	Pontuação	Dificuldade	Erro padrão	*Infit*	*Outfit*	observada	justa
1	6165	1521	-0,89	0,04	1,06	1,04	4,05	4,17
3	5535	1520	0,17	0,03	1,13	1,21	3,64	3,86
4	5420	1521	0,21	0,03	0,96	0,96	3,56	3,68
5	5382	1521	0,22	0,04	0,91	0,91	3,54	3,61
2	5327	1521	0,28	0,04	0,97	0,97	3,50	3,51

Fonte: Dados da pesquisa

5.5 INTERPRETAÇÃO DA QUALIDADE DA ESCALA DE CLASSIFICAÇÃO

Nesta seção são feitas interpretações sobre a qualidade da escala e o modo como os avaliadores atribuíram as pontuações conforme as diretrizes sugeridas por Linacre (2002a) para examinar a qualidade das escalas de avaliação no contexto das medidas de Rasch. O Quadro 19 da Seção 3.6.7 expõe um resumo dessas diretrizes.

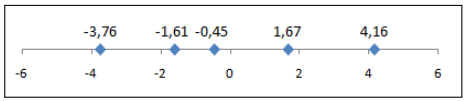
A Tabela 19 exibe um resumo da maneira como os avaliadores, em média, utilizaram a escala de classificação no teste conforme estabelece a formulação de escala gradual do modelo multifacetas de Rasch (equação (11)). A Figura 14 exibe a representação gráfica da locação das categorias (limiares).

Tabela 19 – Estrutura da escala – Modelo: Escala gradual

		MÉDIA		CONTAGEM		CATEGORIA	
Categoria	Nome	Observ.	Esper.	Categoria	*Outfit*	Locação (d_k)	$\lvert d_k - d_{k-1}\rvert$
1	Inadequado	-1,78	-1,81	107 (1%)	1,0		
2	Mínimo	-1,14	-1,21	1010 (13%)	1,1	-3,76	
3	Razoável	-0,60	-0,54	2096 (28%)	0,9	-1,61	2,15
4	Bom	0,28	0,28	2841 (37%)	1,0	-0,45	1,16
5	Muito bom	1,62	1,57	1250 (16%)	0,9	1,67	2,12
6	Excelente	3,89	3,96	300 (4%)	1,1	4,16	2,49

Fonte: Dados da pesquisa

Figura 14 – Localização das categorias – Modelo: Escala gradual



Fonte: Autora

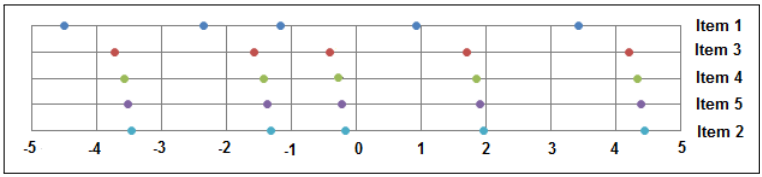
A Tabela 19 exibe a medida de dificuldade de cada categoria (k) que, por sua vez, são as mesmas para todos os itens (modelo de escala gradual). Os valores da dificuldade de cada item (b_i) foram dados na Tabela 17 e para a obtensão da dificuldade de cada categoria de cada item deve-se somar a dificuldade do item com a locação da categoria, isto é, $b_{ik} = b_i + d_k$. Estes dados são dados na Tabela 20 e uma ilustração da alocação dessas medidas na escala é feita na Figura 15.

Tabela 20 – Medidas da dificuldade das categorias – Modelo: Escala gradual

Item	Dificuldade (b_i)	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}
1	-0,74	-4,50	-2,35	-1,19	0,93	3,42
3	0,04	-3,72	-1,57	-0,41	1,71	4,2
4	0,18	-3,58	-1,43	-0,27	1,85	4,34
5	0,23	-3,53	-1,38	-0,22	1,90	4,39
2	0,29	-3,47	-1,32	-0,16	1,96	4,45

Fonte: Dados da pesquisa

Figura 15 – Dificuldade das categorias dos itens – Modelo: Escala gradual



Fonte: Autora

Linacre (2002) considera que a orientação direcional das categorias (diretriz 1), em sequência, da escala de classificação deve estar alinhada com a variável latente. Em outras palavras, espera-se que valores elevados nas observações correspondam a altas posições na variável latente. Pode-se observar que nessa aplicação existe uma estreita correspondência entre os valores das medidas observadas e esperadas (Figura 16). Esse fato fornece informações sobre a direcionalidade da escala de avaliação indicando que a progressão da dificuldade está implícita nas categorias ordenadas. Além disso, as categorias devem ter uma progressão monotônica na escala de classificação, que é observada pelo aumento contínuo das médias das medidas de cada categoria (diretriz 2).

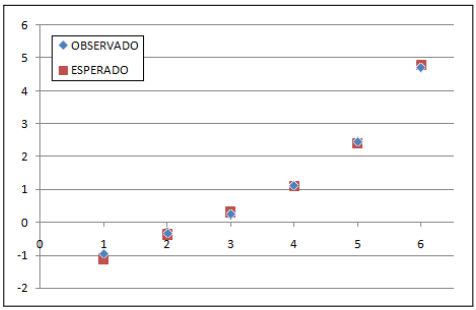
Em termos da utilização de cada categoria (diretriz 3), deve-se observar a distribuição das observações nas categorias da escala de classificação. Linacre (2002a) sugere que as categorias com menos de 10 observações limitam a precisão e a estabilidade dessas estimativas. Analisando os dados da Tabela 19, observa-se que as frequências de utilização das categorias e as porcentagens indicam que, de modo geral, há uma razoável distribuição na utilização das categorias da escala. Além disso, os valores das estatísticas *MQ–Infit* para todas as categorias da escala estão próximos do valor esperado (1,00) (diretriz 5), indicando que os dados observados possuem um bom

ajuste com os dados esperados pelo modelo.

A localização dos coeficientes da categoria devem ter uma progressão ao longo da variável latente (diretriz 6). Para medidas invariantes, a capacidade dos examinandos sobre a escala de avaliação depende de uma sequência monótona da localização dos coeficientes de categoria. Além disso, a descrição precisa do desempenho dos examinandos em relação à variável latente depende da localização das categorias na escala de avaliação (diretriz 7). Quando a localização das categorias é distinta, cada uma delas descreve uma gama única de indivíduos sobre a variável latente (Linacre, 2002a). Observa-se na Tabela 19 o aumento contínuo dos coeficientes das categorias sobre a variável latente. A menor diferença ocorre entre as categorias 3 e 4 (1,16 *logitos*) e a maior diferença ocorre entre as categorias 5 e 6 (2,49 *logitos*). Linacre (2002a) estabelece uma diferença mínima entre categorias adjacentes, em valor absoluto, de 1,40 *logitos* para que a localização das categorias possam ser consideradas distintas. Neste estudo as categorias 3 e 4 são as únicas consideradas muito próximas, não diferenciando bem os examinandos em relação ao traço latente (ver Figura 15).

O alinhamento entre as medidas observadas e esperadas para as categorias conforme o modelo de escala gradual é apresentado graficamente na Figura 16. Esse gráfico ilustra o aumento contínuo das medidas médias dos limiares das categorias sobre a escala de habilidades, indicando a monotonicidade exigida pela diretriz 2. Pode-se também observar a existência de uma estreita correspondência entre os valores observados e esperados em todos os itens. Esse é um dos requisitos para satisfazer a diretriz 1, sobre a direcionalidade das categorias, confirmando o poder discriminatório dos itens entre os indivíduos em relação à variável latente.

Figura 16 – Valores observados e esperados das categorias – Modelo: Escala gradual



Fonte: Autora

Um resumo das estruturas das escalas de avaliação para os cinco itens das tarefas de números 49 e 50, conforme foram utilizadas pelos avaliadores segundo o modelo multifacetas de Rasch, escala de crédito parcial (eq. (15)), são expostos nas Tabelas 21 e 22, respectivamente.

Tabela 21 – Estrutura da escala: Tarefa 49 – Modelo: Crédito parcial

Categoria	Nome	MÉDIA Observ.	MÉDIA Esper.	CONTAGEM Categoria	*Outfit*	d_{ik}	b_{ik}	$\lvert b_{ik} - b_{i(k-1)} \rvert$
TAREFA 49 – Item 1								
1	Inadequado	-0,56	-1,10	1 (0%)	1,0			
2	Mínimo	-0,27	-0,57	36 (5%)	1,1	-3,46	-4,35	
3	Razoável	-0,03	0,06	145 (19%)	0,9	-0,69	-1,58	2,77
4	Bom	0,74	0,79	312 (41%)	1,0	0,60	-0,29	1,29
5	Muito bom	1,70	1,90	224 (29%)	1,2	2,58	1,69	1,98
6	Excelente	4,47	4,33	52 (7%)	1,0	5,42	4,53	2,84
TAREFA 49 – Item 2								
1	Inadequado	-1,82	-1,95	18 (2%)	1,1			
2	Mínimo	-1,24	-1,35	94 (12%)	1,2	-3,59	-3,31	
3	Razoável	-0,75	-0,66	15 (41%)	0,9	-2,51	-2,23	1,08
4	Bom	0,27	0,19	23 (29%)	0,9	-0,20	0,08	2,31
5	Muito bom	1,62	1,55	90 (12%)	1,0	1,42	1,70	1,62
6	Excelente	3,63	3,78	30 (4%)	1,1	3,47	3,75	2,05
TAREFA 49 – Item 3								
1	Inadequado	-1,90	-1,76	19 (2%)	1,0			
2	Mínimo	-1,10	-1,20	181 (24%)	1,2	-3,92	-3,75	
3	Razoável	-0,50	-0,61	107 (14%)	1,2	-0,57	-0,40	3,35
4	Bom	0,02	0,12	314 (41%)	1,3	-1,53	-1,36	0,96
5	Muito bom	1,73	1,38	117 (15%)	0,9	1,48	1,65	3,01
6	Excelente	2,95	3,76	31 (4%)	1,7	3,69	3,86	2,21
TAREFA 49 – Item 4								
1	Inadequado	-1,58	-1,85	18 (2%)	1,2			
2	Mínimo	-1,35	-1,27	133 (17%)	0,9	-3,79	-3,58	
3	Razoável	-0,62	-0,64	217 (28%)	1,0	-1,68	-1,47	2,11
4	Bom	0,16	0,13	253 (33%)	0,8	-0,66	-0,45	1,02
5	Muito bom	1,47	1,43	122 (16%)	0,9	1,20	1,41	1,86
6	Excelente	3,93	3,81	27 (4%)	0,9	3,88	4,09	2,68
TAREFA 49 – Item 5								
1	Inadequado	-2,11	-1,94	10 (1%)	0,9			
2	Mínimo	-1,36	-1,34	105 (14%)	1,0	-4,24	-4,02	
3	Razoável	-0,72	-0,67	270 (35%)	0,8	-2,20	-1,98	2,04
4	Bom	0,25	0,17	259 (34%)	0,8	-0,48	-0,26	1,72
5	Muito bom	1,82	1,65	107 (14%)	0,9	1,45	1,67	1,93
6	Excelente	3,79	4,06	19 (2%)	1,1	4,37	4,59	2,92

Fonte: Dados da pesquisa

Tabela 22 – Estrutura da escala: Tarefa 50 – Modelo: Crédito parcial

		MÉDIA		CONTAGEM		CATEGORIA		
Categoria	Nome	Observ.	Esper.	Categoria	*Outfit*	d_{ik}	b_{ik}	$\lvert b_{ik} - b_{i(k-1)} \rvert$
TAREFA 50 – Item 1								
1	Inadequado	-1,02	-0,97	1 (0%)	1,1			
2	Mínimo	-0,26	-0,44	68 (9%)	1,4	-4,11	-5,00	
3	Razoável	0,15	0,18	159 (21%)	1,0	-0,16	-1,05	3,95
4	Bom	0,98	0,89	293 (39%)	1,1	0,73	-0,16	0,89
5	Muito bom	1,97	1,93	189 (25%)	1,0	2,62	1,73	1,89
6	Excelente	4,62	4,36	41 (5%)	0,7	5,37	4,48	2,75
TAREFA 50 – Item 2								
1	Inadequado	-2,03	-2,00	10 (1%)	1,0			
2	Mínimo	-1,34	-1,38	53 (7%)	1,0	-3,64	-3,36	
3	Razoável	-0,73	-0,65	349 (46%)	0,9	-3,18	-2,90	0,46
4	Bom	0,35	0,24	228 (30%)	0,9	-0,08	0,20	3,10
5	Muito bom	1,65	1,72	87 (12%)	1,0	1,57	1,85	1,65
6	Excelente	4,17	4,04	24 (3%)	0,9	3,94	4,22	2,37
TAREFA 50 – Item 3								
1	Inadequado	-1,39	-1,80	14 (2%)	2,2			
2	Mínimo	-1,15	-1,27	126 (17%)	1,4	-3,89	-3,72	
3	Razoável	-0,77	-0,68	63 (8%)	0,8	-0,44	-0,27	3,45
4	Bom	0,01	0,08	417 (56%)	1,2	-2,37	-2,20	1,93
5	Muito bom	1,52	1,49	103 (14%)	0,9	1,92	2,09	4,29
6	Excelente	3,88	4,02	28 (4%)	1,3	3,93	4,10	2,01
TAREFA 50 – Item 4								
1	Inadequado	-1,95	-1,84	11 (1%)	0,9			
2	Mínimo	-1,32	-1,25	107 (14%)	0,9	-4,02	-3,81	
3	Razoável	-0,60	-0,59	215 (29%)	0,9	-1,82	-1,61	2,20
4	Bom	0,19	0,20	288 (38%)	1,0	-0,70	-0,49	1,12
5	Muito bom	1,56	1,56	101 (13%)	1,0	1,65	1,86	2,35
6	Excelente	3,84	3,97	29 (4%)	1,3	3,83	4,04	2,18
TAREFA 50 – Item 5								
1	Inadequado	-2,63	-1,88	5 (1%)	0,9			
2	Mínimo	-1,34	-1,27	107 (14%)	1,0	-4,84	-4,62	
3	Razoável	-0,60	-0,58	256 (34%)	0,8	-2,00	-1,78	2,84
4	Bom	0,20	0,24	254 (34%)	0,8	-0,38	-,16	1,62
5	Muito bom	1,82	1,74	110 (15%)	0,9	1,52	1,74	1,90
6	Excelente	4,12	4,20	19 (3%)	1,1	4,59	4,81	3,07

Fonte: Dados da pesquisa

De modo similar às análises feitas para o modelo de escala gradual, as análises da qualidade das escalas para esse modelo também são feitas com base nas diretrizes sugeridas por Linacre (2002a) e resumidas no Quadro 19 da Seção 3.6.7. O modelo de crédito parcial permite que cada item possua a sua própria estrutura de escala de classificação. Nessas tabelas, a sétima

coluna exibe os limiares das categorias de cada item (d_{ik}) e a oitava os parâmetros da dificuldade das categorias $b_{ik} = b_i + d_{ik}$. Estes valores indicam o ponto na escala de habilidades no qual a probabilidade de um examinando de obter a classificação k é igual a probabilidade dele de obter a classificação $k-1$ para cada item. A nona coluna traz a distância entre parâmetros consecutivos ($|b_{ik} - b_{i(k-1)}|$).

Linacre (2002a) considera que a orientação direcional das categorias, em sequência, da escala de classificação deve estar alinhada com a variável latente (diretriz 1). As Tabelas 21 e 22 mostram que o crescimento de forma monótona das médias observadas para todos os cinco itens das duas tarefas se dá juntamente com o aumento da habilidade do examinando em relação ao construto (categorias). As duas últimas colunas das Tabelas 21 e 22 também indicam que a média da localização de cada categoria da escala de classificação aumenta de uma categoria para a outra, com exceção do item de número 3, com um decréscimo entre as categorias 2 e 3, fato comum às duas tarefas.

A distribuição das observações nas categorias demonstra que elas foram utilizadas de modo razoável em todos os itens das duas tarefas, com algumas exceções (diretriz 3). Obtiveram menos de 10 observações a categoria 1 do item 1 da tarefa de número 49 e as categorias 1 dos itens 1 e 6 da tarefa de número 50. Esse número mínimo de observações é recomendado por Linacre (2002a) para uma estimativa precisa da localização da categoria.

Os valores da estatística *MQ–Infit* para todas as categorias da escala, em todos os cinco itens da tarefa de número 49, estão próximos do valor esperado (1,00), com exceção da categoria 6 do item 3, com valor para *MQ–Infit* de 1,7. Para a tarefa de número 50, os valores dessa estatística para quase todas as categorias do item 3 não estão próximos do valor esperado (1,0), indicando que pode haver problemas com esses dados, mesmo assim, esse resultado não afeta demasiadamente as estimativas conforme estabelece Linacre (2002a).

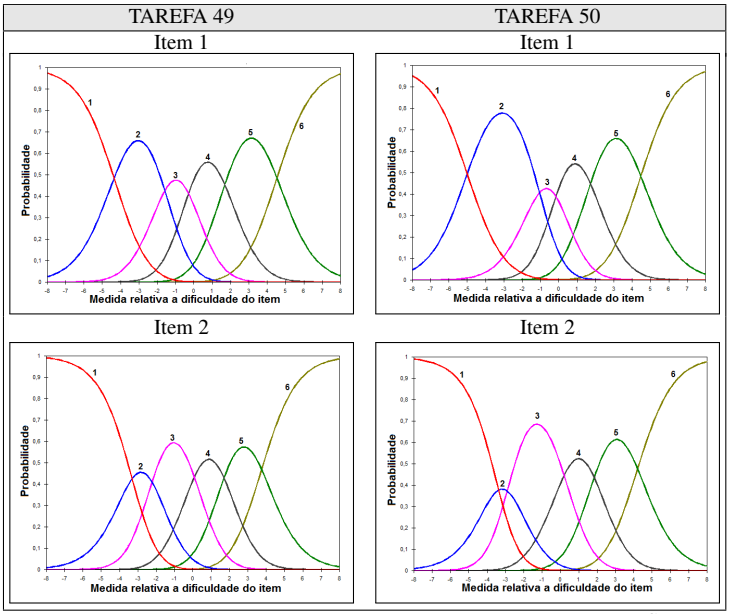
As ordens das categorias refletem a ordem pretendida para todos os itens, com exceção do de número 3 para as duas tarefas (diretriz 6). As localizações das categorias são, de modo geral, distintas umas das outras e seus coeficientes crescem de modo monótono sobre a escala de habilidades (Diretriz 7). Linacre (2002a) considera que os limiares das categorias devem estar a uma distância maior do que 1,40 *logito*s para diferenciar os examinandos em relação ao construto avaliado.

As localizações dos limiares das categorias não atenderam esse requisito para as categorias 2 e 3 do item 2 e para as categorias 3 e 4 dos itens 1, 3 e 4 da tarefa 49, assim como entre as categorias 3 e 4 do item 1, para

as categorias 2 e 3 do item 2 e para as categorias 3 e 4 do item 4 da tarefa de número 50. As outras categorias adjacentes, de todos os cinco itens, são suficientemente afastadas.

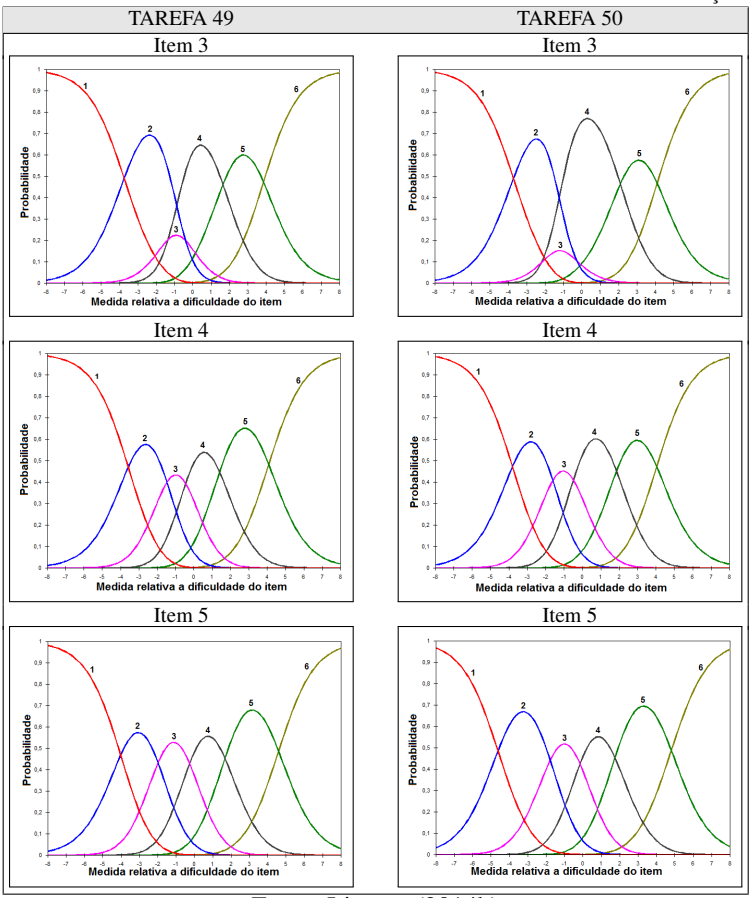
Os gráficos das curvas características dos itens das duas tarefas são expostos na Figura 17. Novamente, pode-se constatar o problema detectado nos itens 3 das duas tarefas; a categoria 3 fica "coberta" pelas categorias 2 e 4, evidenciando que a categoria 3 não está discriminando os candidatos. As distâncias entre as categorias não são muito evidentes nestas ilustrações porque a escala na qual os gráficos então sendo mostrados é muito pequena, mesmo assim, percebe-se que, para a tarefa de número 50, a distância entre as categorias 2 e 3 do item de número 2 é muito pequena. Essa foi a menor distância entre as categorias obtidas nessa aplicação.

Figura 17 – Curvas características dos itens – Tarefas 49 e 50



Continua

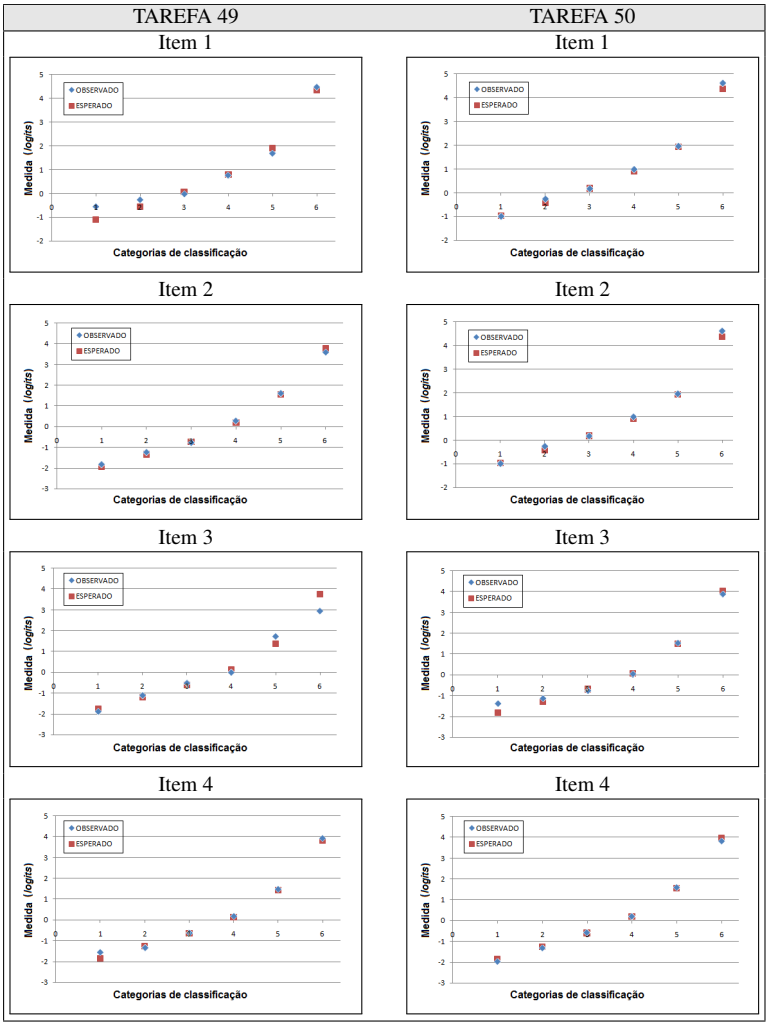
Continuação



Fonte: Linacre (2014b)

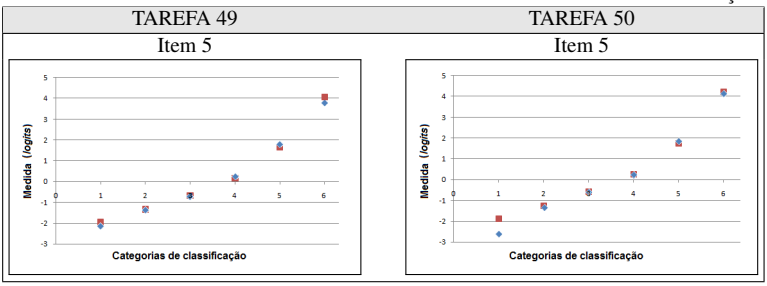
As médias dos valores observados e esperados para as categorias conforme o modelo de escala de crédito parcial para as duas tarefas, 49 e 50, são apresentadas graficamente na Figura 18. Cada um desses gráficos ilustra o aumento contínuo das medidas médias das categorias sobre a escala de habilidades, confirmando o poder discriminatório dos indivíduos em relação à variável latente.

Figura 18 – Valores observados e esperados – Modelo: Crédito parcial



continua

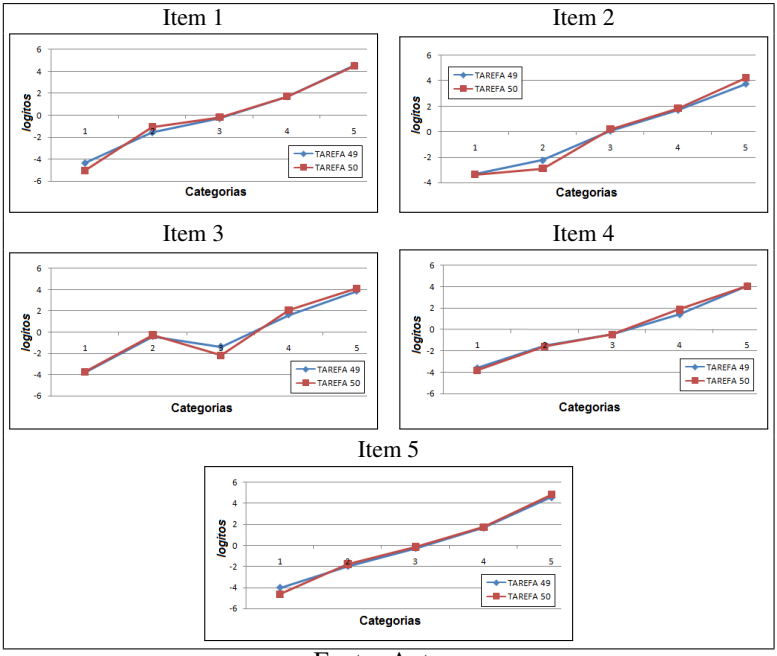
continuação



Fonte: Autora

A Figura 19 traz uma ilustração gráfica das locações das categorias dos itens das duas tarefas, 49 e 50, com o intuito de auxiliar a interpretação da qualidade das escalas conforme as diretrizes propostas por Linacre (2002a).

Figura 19 – Localização das categorias – Modelo: Crédito parcial



Fonte: Autora

Esta visualização gráfica proporciona também a comparação entre as duas tarefas que se mostram com comportamentos muito parecidos em todos os itens e em relação a todas as categorias. O item de número 3 mostra o mesmo comportamento em ambas as tarefas. Há um decréscimo na localização do coeficiente da categoria 3 em relação à categoria 2, demonstrando que deve haver problemas na definição dos critérios desse item.

A lista de respostas não esperadas fornecida pelo programa *Facets*, exibida na Tabela 7 para o modelo de escala gradual, também auxilia na detecção de problemas com as tarefas, os itens e as escalas de avaliação utilizadas pelos avaliadores. Nesse experimento, foi imposto que as pontuações com o residual padronizado maiores do que 3 deveriam constar na lista de respostas não esperadas. Para o modelo de crédito parcial, foram detectadas 40 respostas não esperadas, entre elas cerca de 35% são referentes à tarefa 49 e 65% à tarefa 50. Para ambas as tarefas, o item mais contemplado com respostas não esperadas foi o de número 3, com 42,9% para a tarefa 49 e 65,4% para a tarefa 50, confirmando mais uma vez que esse item deve ser revisto.

5.6 CONCLUSÃO SOBRE O PADRÃO DE QUALIDADE DA AVALIAÇÃO

As implicações deste estudo são importantes para o desenvolvimento e interpretação da qualidade de avaliações que necessitam do julgamento de avaliadores quanto ao desempenho dos examinandos. Essas avaliações exigem cuidados especiais para a elaboração uma vez que demandam a formulação dos itens, dos critérios de classificação e das escalas, como também o treinamento dos avaliadores para a correta utilização desses critérios e escalas, do modo como eles foram concebidos. Os resultados deste estudo sugerem que uma variedade de métodos, índices quantitativos e gráficos baseados nos modelos multifacetas de Rasch fornecem múltiplas fontes de evidência para o monitoramento da qualidade das avaliações e auxiliam de maneira incisiva na interpretação da pontuação atribuída pelo avaliador.

Jaeger *et al.* (1996) propuseram algumas questões para assegurar que a avaliação satisfaça padrões profissionais de medição (Seção 3.6). Com a utilização dos modelos multifacetas de Rasch, assim como com os índices estatísticos, os gráficos e outros recursos disponibilizados pelo programa *Facets*, foi possível obter respostas a essas perguntas.

Na avaliação implementada para a aplicação da sistemática proposta neste trabalho, os índices média quadrática *infit* e *outfit* indicam que os da-

dos, de modo geral, se ajustam satisfatoriamente ao modelo de Rasch, sendo que 81,14% dos examinandos possuem essas medidas dentro do intervalo de valores produtivos conforme sugerido por Wright e Linacre (1994). Além disso, entre as 7.604 observações, apenas 23 respostas, cerca de 0,3%, foram consideradas não esperadas pelo modelo multifacetas de escala gradual, por possuírem residual padronizado, em valor absoluto, maior do que 3,0, e pelo modelo de crédito parcial, foram detectadas 40 respostas não esperadas, cerca de 0,5% do total. Esses índices indicam um bom ajuste dos dados ao modelo, o que apoia a constatação de uma classificação confiável dos participantes da avaliação, respondendo, assim, à primeira questão de Jaeger *et al.* (1996).

A aplicação dessa sistemática propõe que a avaliação seja elaborada, desde o início, visando à validade. Os procedimentos adotados em todas as etapas da elaboração das tarefas, dos critérios para a avaliação e das escalas utilizadas pelos avaliadores devem ser baseados nas teorias sobre o construto a ser avaliado e, desse modo, garantir a validade do teste. Além disso, as análises empíricas, obtidas por meio dos recursos fornecidos pelos modelos multifacetas de Rasch descritas anteriormente, como os estudos sobre os critérios e categorias da escala de classificação e a qualidade da pontuação, demonstram que a segunda questão proposta por Jaeger *et al.* (1996) é respondida.

Jaeger *et al.* (1996) propõem que os avaliações devem ser imparciais quanto a sexo, raça, etnia, etc. Estudos para diferenciar os diversos grupos de participantes da avaliação não foram previstos na aplicação prática deste trabalho, embora os modelos multifacetas de Rasch estejam sendo utilizados em estudos desse tipo. Para tanto, devem-se considerar separadamente as variáveis a serem estudadas dentro dos grupos e verificar a existência de alguma influência entre a dificuldade relativa dos itens ou a pontuação proveniente dos avaliadores e os elementos observados nos grupos, por exemplo.

A última questão da lista de Jaeger *et al.* (1996) refere-se às normas institucionais para a classificação dos examinandos. Quando um sistema avaliativo é desenvolvido segundo a teoria de medição de Rasch, podem-se manter os padrões de classificação de uma avaliação para a outra. Desse modo, é possível comparar o nível de habilidade dos participantes de edições distintas dos exames e assegurar que as normas de certificação para cada categoria em particular foram respeitadas.

6 CONSIDERAÇÕES FINAIS

Os sistemas de avaliações em larga escala devem ter como objetivo o desenvolvimento de instrumentos de avaliação que possibilitem inferências válidas, confiáveis e justas em relação à medida obtida da habilidade dos participantes. Nas avaliações em larga escala, o diagnóstico dos problemas é importante para que eles sejam sanados antes da próxima edição do evento. Nesse contexto, são duas as principais demandas: os psicometristas devem desenvolver teorias e modelos que possam ser utilizados para compreender, conceituar e eficientemente resolver os eventuais problemas práticos; os profissionais e pesquisadores da área de avaliação devem utilizar a teoria de medição disponível para fornecer às pessoas resultados da avaliação que sejam, tanto quanto possível, válidos e justos.

Estudos utilizando novas abordagens e técnicas para determinar a qualidade da avaliação são necessárias, nesse sentido, a pesquisa foi norteada pelo seguinte problema de pesquisa: *"Qual é a contribuição que a utilização do modelo multifacetas de Rasch pode proporcionar para a análise de avaliações com itens de respostas construídas?"*

Desta forma, a tese teve como objetivo geral estabelecer como o modelo multifacetas de Rasch pode contribuir para a determinação da qualidade das avaliações com itens de respostas construídas. A abordagem utilizada pelo modelo multifacetas de Rasch proporciona análises sobre a qualidade das medidas relacionadas aos examinandos, aos avaliadores, às tarefas, aos itens e às escalas de classificação utilizadas para a pontuação das tarefas. Além disso, o modelo multifacetas de Rasch permite análises no nível individual de cada elemento participante das avaliações com itens de respostas construídas, mostrando-se eficaz para a detecção de erros. Desse modo, é possível que os erros sejam corrigidos resultando na melhoria dos processos avaliativos.

6.1 CONCLUSÃO

A pesquisa bibliográfica elaborada na etapa teórica deste trabalho permitiu a determinação dos aspectos importantes e também das etapas mais problemáticas nas avaliações em larga escala com itens de respostas construídas e, a partir daí, tornou-se possível o estabelecimento dos procedimentos essenciais em cada uma das etapas demandadas para concepção, elaboração, aplicação, pontuação, análises, entre outros, dessas avaliações. A partir des-

sas pesquisas, foi possível elaborar uma sistemática para ser utilizada pelas empresas provedoras de avaliações em larga escala, integrando os processos necessários para a construção de avaliações com itens abertos, com o intuito de guiar e auxiliar as pessoas em todas as etapas demandadas para a construção dessas avaliações, de modo que elas possam alcançar padrões profissionais de qualidade.

A aplicação prática foi realizada por meio de análises feitas com a utilização do modelo multifacetas de Rasch, da pontuação das respostas a duas tarefas de escrita que fizeram parte do concurso público para provimento de vagas da Polícia Militar do Estado do Paraná aplicado em fevereiro de 2010 pela Coordenadoria de Processos Seletivos da Universidade Estadual de Londrina (COPS/UEL). Foram consideradas quatro facetas nas análises; a habilidade dos examinandos, a dificuldade dos itens, a dificuldade das tarefas e a severidade dos avaliadores. A estrutura da escala de classificação utilizada também foi analisada.

As análises dos dados de modo geral, tomados em conjunto, indicaram um ajuste satisfatório dos dados ao modelo, e o mapa das variáveis (Figuras 10 e 11), mostrou-se um recurso muito informativo para auxiliar na interpretação dos dados, retratando todas as facetas da análise em um único quadro de referência. Esse recurso é de grande valia e facilita comparações dentro e entre as várias facetas.

As medidas dos desempenhos dos examinandos estimadas pelo modelo multifacetas de Rasch de escala gradual foram aproximadas às estimadas pelo modelo de crédito parcial. Estas variaram entre -2,63 e 6,46 *logitos*, com a maior concentração de indivíduos entre -2,0 e 2,0 *logitos*. A média das medidas da habilidade dos participantes foi de aproximadamente 0,22 e o desvio padrão é de SD = 1,52, com precisão de 0,05. No entanto, as medidas da habilidade dos examinandos obtidas pelo modelo de escala gradual de duas facetas, habilidade dos examinandos e dificuldade dos itens, que resulta no modelo original de Andrich, foram significativamente diferentes das medidas obtidas pelos modelos multifacetas de Rasch com quatro facetas. Segundo o modelo de duas facetas, a medida da habilidade dos examinandos variou entre -2,23 *logitos* e 5,71 *logitos*. O modelo de duas facetas não considera os efeitos causados pelos avaliadores, e por este motivo, a classificação dos examinandos é significativamente diferente da classificação obtida quando estes efeitos são considerados.

O examinando de menor habilidade obteve medida estimada pelo modelo de quatro facetas cerca de 0,40 *logitos* mais baixa do que a obtida pelo modelo de duas facetas, enquanto o de maior habilidade, a medida estimada

pelo modelo de quatro facetas é cerca de 0,75 *logitos* mais alta do que a estimada com o modelo de duas facetas.

Os examinandos que possuem habilidades intermediárias sofreram modificações maiores em suas classificações do que aqueles com habilidades extremamente altas ou extremamente baixas. Em termos gerais, a classificação dos examinandos segundo o modelo de escala gradual com quatro facetas foi modificada em relação à classificação obtida com o modelo de duas facetas em mais de 5 posições para 235 examinandos, o que equivale a 67,14% dos 350 indivíduos avaliados. Desses, 46,28% tiveram suas colocações modificadas em mais de 10 posições e 24% em mais de 20 posições.

A calibração das tarefas e dos itens pelos modelos multifacetas de Rasch de escala gradual e de crédito parcial obtiveram índices semelhantes, diferindo muito pouco. As medidas da dificuldade dos itens variaram entre -0,74 *logitos* e 0,29 *logitos*, o item considerado o mais fácil foi o de número 1 e o mais difícil, o de número 2. A variação entre essas medidas na escala de habilidades é muito pequena, estão todas localizadas em torno da origem, revelando que estes itens não são eficientes para discriminar entre examinandos com habilidades fora dessa faixa.

As tarefas foram consideradas equivalentes em relação à dificuldade, pois a diferença de suas medidas é muito pequena. A de número 50 foi considerada um pouco mais fácil, -0,04 *logitos*, enquanto a de número 49 um pouco mais difícil, 0,04 *logitos*.

Mesmo assim, não foi possível afirmar que as duas tarefas são comparáveis quanto a seus graus de dificuldade. Baseando-se nos índices quantitativos calculados, como a taxa de separação de 1,55 e o índice estrato de 2,40, pôde-se estabelecer que a diferença entre as dificuldades das duas tarefas não é muito grande, uma vez que estes valores são relativamente pequenos por serem dados em unidades de erro de medição. A mesma conclusão pôde ser constatada com base no valor do qui-quadrado de 6,8 com 1 grau de liberdade e p=0,01, por ser pequeno comparado com o valor deste índice para as outras facetas. A confiabilidade do índice de separação (0,74), apesar de ser a menor entre as facetas, não é suficientemente próxima de zero para indicar que os elementos da faceta se comportam de modo semelhante. Desse modo, esta questão necessita de mais investigação para uma resposta decisiva.

O modelo multifacetas de Rasch para a escala de crédito parcial permitiu um estudo aprofundado da estrutura de escala de classificação, como ela foi utilizada pelos avaliadores. Para cada item de cada uma das duas tarefas, foi possível identificar algumas diferenças e também as semelhanças na estrutura das escalas de classificação utilizadas. A fim de determinar a qua-

lidade da estrutura das escalas de avaliação, índices quantitativos fornecidos pelo programa *Facets* (LINACRE, 2014b), juntamente com ilustrações gráficas, foram examinados para cada um dos itens das duas tarefas. As escalas de classificação são, de modo geral, eficientes para a classificação dos candidatos de acordo com as diretrizes propostas por Linacre (2002a) (Quadro 19) para as escalas de classificação dos modelos de Rasch, embora algumas exceções tenham sido detectadas.

Todos os itens das duas tarefas atenderam inteiramente à diretriz 1 (direcionalidade), referente ao alinhamento das categorias de classificação com a variável latente e a diretriz 2 (monotonicidade), que estabelece que a habilidade dos examinandos deve aumentar juntamente com as categorias de classificação. A diretriz 3, que estabelece a utilização das categorias, não foi satisfeita para a o item 1, categoria 1 da tarefa 49 e para o item 1, categoria 1 e o item 5, categoria 1 da tarefa 50. A distribuição das observações (diretriz 4) se dá aproximadamente na forma normal para a maioria dos itens das duas tarefas, com exceção dos itens 2 e 3 da tarefa 49 e do item 3 da tarefa 50. A diretriz 5, que confere o ajuste da escala de classificação com o modelo, é verificada para todos os itens com os valores das médias quadráticas *infit* próximos de 1. Quanto à ordem das categorias (diretriz 6), não foi confirmada a ordem crescente para o item 3 das duas tarefas, com os limiares das categorias em uma sequência não monótona ao longo da escala. A diretriz 7, que trata da distância entre as categorias, não foi obedecida em vários itens das duas tarefas: item 1, categorias 3 e 4; item 2, categorias 2 e 3; item 3, categorias 3 e 4; item 4, categorias 3 e 4, tarefa 49; e item 1, categorias 3 e 4; item 2, categorias 2 e 3; item 4, categorias 3 e 4, tarefa 50. Em todos esses casos, os limiares entre essas categorias estavam a menos de 1,4 *logitos* de distância.

Quanto à comparabilidade das escalas utilizadas para a pontuação dos itens das duas tarefas, percebem-se comportamentos semelhantes. As semelhanças mais evidentes podem ser constatadas com o auxílio dos gráficos das curvas características dos itens (Figura 17) e na representação gráfica da localização das categorias (Figura 19).

O exame detalhado do funcionamento da escala de classificação para cada item é essencial para avaliar a qualidade da pontuação atribuída a esses itens. Nas avaliações em larga escala, estudos como este podem indicar a necessidade de reformulações em alguns critérios de classificação ou na estrutura da escala. As avaliações em larga escala que utilizam testes com respostas construídas como parte de seus exames podem ser beneficiados com análises desse tipo, uma vez que elas indicam exatamente os pontos mais frágeis e que necessitam de algum tipo de intervenção ou modificação.

Em relação à confiabilidade da pontuação, os índices de ajuste (*infit* e *outfit*) da faceta avaliadores indicam que as medidas de 86% dos avaliadores estão no intervalo de medidas produtivas, entre 0,5 e 1,5, confirmando que os dados se ajustam ao modelo multifacetas de Rasch. As medidas da severidade dos avaliadores variaram entre -1,54 *logitos* para o mais complacente até e 0,94 *logitos* para o mais severo, estas medidas foram estimadas com precisão de 0,5 *logitos*.

De modo geral, no nível de grupo, os avaliadores não mostraram tendências em pontuações sistemáticas, não apresentando os efeitos de severidade/complacência, tendência central, tendência de halo ou de aleatoriedade. Entretanto, a confiabilidade do índice de separação dos avaliadores perto de 1 indicou variações indesejadas entre os níveis de severidade dos avaliadores, sugerindo que esses avaliadores não podem ser intercambiáveis.

No nível individual, foi possível identificar alguns avaliadores do grupo portadores de tendências a pontuações sistemáticas causando os efeitos de severidade/complacência, central, halo e aleatoriedade. Entretanto, a constatação mais importante nessas análises é da possibilidade de identificar, no nível individual, os avaliadores com problemas em suas pontuações, fontes de vieses geradores de inconsistências nas pontuações.

A possibilidade de identificação, no nível individual, dos avaliadores portadores de comportamentos tendenciosos, que pontuam erradamente ou com graus diferentes de severidade, consiste em uma contribuição importante do modelo multifacetas de Rasch às avaliações mediadas por avaliadores. Os problemas causados por avaliadores são geradores de inconsistências graves nas pontuações dos testes. Outro estudo no nível individual de grande valia é o da estrutura da escala de avaliação utilizada. Os modelos multifacetas de Rasch utilizados nesse estudo proporcionam análises claras do modo como cada avaliador atribui as pontuações e do modo como cada um utiliza os critérios de pontuação e as escalas de classificação.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

Os dados provenientes de avaliações em larga escala reais, pontuadas por avaliadores profissionais, como o ENEM e os concursos vestibulares de instituições importantes, devem ser disponibilizados para propiciar análises de seus testes. Os problemas mais comuns presentes nas avaliações mediadas por avaliadores são os causados pelos julgamentos imprecisos e tendenciosos dos avaliadores. Essas avaliações são carentes de análises que auxiliem no

diagnóstico desses problemas e, consequentemente, nas suas soluções. Neste sentido, as análises de tais avaliações são importantes e podem contribuir com a melhoria da qualidade dos processos, uma vez que os resultados dessas podem gerar consequências sérias para as instituições e para as pessoas envolvidas.

Em avaliações do porte dos concursos vestibulares e do ENEM, não se pode deixar de evidenciar as vantagens da formação de um banco de avaliadores que pode ser proporcionada pela utilização do modelo MFR. Uma vez o banco criado, as informações e características sobre cada avaliador são conhecidas, desse modo, a organização do evento pode classificar os avaliadores conforme apresentem alguma tendência e, então, proporcionar treinamentos específicos visando à correção desses problemas. É claro que, como ocorre com os tradicionais bancos de itens, avaliadores podem ser incorporados ou desligados do grupo a cada edição do evento.

As provas de redação dos concursos vestibulares das principais universidades brasileiras, normalmente, são corrigidas com os avaliadores presencialmente. As análises fornecidas pelo modelo multifacetas podem ser feitas parcialmente enquanto ocorrem as sessões de pontuação, informando aos responsáveis se estão ocorrendo desajustes na pontuação e quais avaliadores necessitam de algum acompanhamento especial. Essa possibilidade de monitoramento da qualidade das pontuações, ocorrendo ao mesmo tempo em que elas são feitas, pode proporcionar um aumento significativo na precisão e qualidade das pontuações atribuídas às tarefas elaboradas pelos examinandos, mesmo porque será dada a oportunidade de correção dos erros graves que já ocorreram e também a prevenção da ocorrência desses erros outras vezes. O mesmo trabalho pode ser desenvolvido para as sessões de pontuação *on-line* como ocorre com a correção das redações do ENEM. Os dados podem ser testados sistematicamente para o monitoramento da qualidade.

Outras análises podem ser realizadas com a utilização do modelo multifacetas de Rasch e dos dados provenientes das avaliações mediadas por avaliadores em larga escala, uma delas é o estudo do comportamento diferencial dos avaliadores em relação a grupos específicos de examinandos, por exemplo, raça, gênero, nacionalidade, entre outros. Um comportamento diferencial comum entre os avaliadores é em relação à letra do examinando uma vez que as provas com itens abertos são escritas à mão. Os avaliadores despendem um grande esforço para que a letra da pessoa não influencie a pontuação, entretanto, este esforço nem sempre é suficiente. O estudo sobre o funcionamento diferencial do item consiste em outra análise importante que pode ser proporcionada pelo modelo multifacetas de Rasch. Com a utilização deste modelo,

pode-se identificar se existem itens no instrumento de avaliação que são mais ou menos favoráveis a um determinado grupo de examinandos. Essas análises sobre os efeitos diferenciais dos avaliadores e também dos itens possibilitam assegurar a imparcialidade em relação aos examinandos e, consequentemente, a justiça na avaliação.

Os modelos e métodos descritos neste trabalho fornecem uma base teórica consistente para análises das avaliações mediadas por avaliadores, principalmente no nível individual dos elementos participantes da avaliação. As vantagens na utilização do modelo MFR nessas avaliações são inúmeras e muitos estudos podem ser desenvolvidos com foco nos mais variados elementos. O modelo MFR proporciona análises por um conjunto diverso de índices quantitativos, ilustrações gráficas, tabelas, entre outros, auxiliando na determinação de evidências para o monitoramento da qualidade das avaliações.

6.3 LIMITAÇÕES DO TRABALHO

A proposta deste trabalho foi a de estabelecer como o modelo multifacetas de Rasch pode contribuir para a determinação da qualidade das avaliações com itens de respostas construídas. Estas avaliações são aquelas que necessitam da mediação de avaliadores para a pontuação das tarefas elaboradas pelos examinandos. Entretanto, os procedimentos e resultados estabelecidos neste trabalho, delimitam-se ao estudo de tarefas de escrita de textos, como as redações dos exames de seleção e concursos vestibulares.

Além disso, as respostas dos examinandos cujos dados foram utilizados neste trabalho, são resultantes de pontuações elaboradas por estudantes dos cursos de graduação e pós-graduação em Letras e por professores de língua portuguesa da rede de ensino. Mesmo que as sessões de pontuações tenham sido conduzidas de acordo com procedimentos recomendados na literatura, muitas vezes, as pontuações tiveram que ser refeitas por apresentarem discrepâncias sérias, provavelmente pela falta de experiência ou algumas vezes pela falta de seriedade dos avaliadores. O ideal seria a replicação deste estudo com dados provenientes de avaliações em larga escala reais.

REFERÊNCIAS

ABAURRE, M. B. M. Vestibular discursivo da UNICAMP: um espaço de interação entre a universidade e a escola. *Ensaio*: avaliação e políticas públicas em educação, v. 3, n. 9, out./dez. 1995.

American Educational Research Association (AERA); American Psychological Association (APA); National Council on Measurement in Education (NCME). *AERA: Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 1999.

ALDERSON, C. J.; BANERJEE, J. Language testing and assessment (Part 2). *Language Teaching*, 35, p. 79-113, 2002.

ALVES, T.; GOUVÊA, M. A., VIANA, A. B. N. The socioeconomic level of public school students and the conditions for the provision of education in the Brazilian municipalities. *Education Policy Analysis Archives*. v. 20, p. 1-29, 2012.

ANASTASI, A. *Testes psicológicos*. Trad. Dante Moreira Leite. 2 ed. São Paulo: EPU, 1977.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item*: Conceitos e Aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

APPLEBEE, A. Alternative Models of Writing Development. In *Writing*: Research/Theory/Practice. INDRISANO R.; SQUIRE, J. R. Newark, DE: International Reading Association, 2000.

AZEVEDO, C. L. N. *Métodos de Estimação na Teoria de Resposta ao Item*. (Dissertação de Mestrado). São Paulo: Universidade de São Paulo, instituto de matemática e estatística, 2003.

BACHA, N. Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, v. 29, n. 3. p. 371-383, 2001.

BAIRD, J. A. Alternative conceptions of comparability. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 166-206. London: QCA, 2007.

BARDOT, B.; TAN, M.; RANDI J.; SANTA-DONATO, G.; GRIGORENKO, E. L. Essential skills for creative writing: Integrating multiple domain-specific perspectives. *Thinking Skills and Creativity*, v.7, n. 3, p. 1-15, 2012.

BARKAOUI K. Effects of marking method and rater experience on ESL essay scores. *Assessment in Education:* Principles, Policy e Practice, v. 18, n. 3, p. 277-291, 2011.

_____. Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, v. 12, p. 86-107, 2007.

BECK, S. W; JEFFERY, J. V. Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing*. v.12. p. 60-79, 2007.

BECKER, A. Examining Rubrics Used to Measure Writing Performance in U.S. Intensive English Programs. *The Catesol Journal*, n. 22.1, 2010/2011.

BEHIZADEH, N.; ENGELHARD, G. Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*. v. 16, n. 3, p. 189-211, 2011.

BESSA, N. M. Fidedignidade de Notas Atribuídas a Redações: enfoque teórico e empírico. *Educação e Seleção*, n. 14, 1986.

BOCK, R. D.; BRENNAN, R. L.; MURAKI, E. The Information in Multiple Ratings. *Applied Psychological Measurement*, v. 26, n. 4, p. 364-375, 2002.

BONAMINO, A.; COSCARELLI, C.; FRANCO, C. Avaliação e letramento: Concepções de aluno letrado subjacentes ao SAEB e ao PISA. *Educação e sociedade*, v. 23, n. 81, p. 91-113, 2002.

BORSBOOM, D; MELLENBERG, G. J.; VAN HEERDEN, J. The Concept of Validity. *Psychological Review*, v. 111, n. 4, p. 1061-1071, 2004.

BRANDÃO, Z.; CANEDO, M. L.; XAVIER, A. Construção solidária do habitus escolar: Resultados de uma investigação nos setores público e privado. *Revista Brasileira de Educação*. v. 17, n. 49, 2012, p. 193-218.

BRASIL. Ministério da Educação (MEC). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Diretoria de Avaliação da Educação Básica (DAEB). Exame Nacional do Ensino Médio (ENEM). Disponível em: <http://www.enem.inep.gov.br/> Acesso: 2 set. 2012.

_____. *A redação no ENEM 2012 – Guia do participante*. Brasília, DF, 2012. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/downloads/2012/guia_participante_redacao_enem2012.pdf>. Acesso em: 12 set. 2012.

_____. *A redação no ENEM 2013 – Guia do participante*. Brasília, DF, 2013. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_participante_redacao_enem_2013.pdf>. Acesso em: 6 set. 2013.

_____. *Guia de elaboração e revisão de itens*. v. 1, Brasília, DF, 2010. Disponível em: <http://download.inep.gov.br/outras_acoes/bni/guia/guia_elaboracao_revisao_itens_2012.pdf>. Acesso em: 22 out. 2013.

BRASIL, Decreto N. 79.298, de 24 de Fevereiro de 1977. Disponível em <http:// www6.senado.gov.br/legislacao>. Acesso: 10 de out. 2012.

BRASIL. Ministério da Educação e Cultura. Secretaria de Educação Média e Tecnológica. *Parâmetros Curriculares Nacionais: ensino médio*. Brasília: Ministério da Educação e Cultura, 1999.

BRELAND, H.; LEE, Y. W.; NAJARIAN, M.; MURAKI, E. An Analysis of TOEFL CBT Writing Prompt Difficulty and Comparability for Different Gender Groups. *Educational Tsting Service* (ETS), Research Reports, Princeton, NJ, 2004.

BRENNAN, R. L. Using Generalizability Theory to Address Reliability Issues for PARCC Assessments: A White Paper. *Center for Advanced Studies in Measurement and Assessment* (CASMA), University of Iowa, 2011.

BRIDGEMAN, B.; MORGAN, R.; WANG, M. Choice Among Essay Topics: Impact on Performance and Validity. *Educational Testing Service*, v. 34, n. 3, p. 273-286, 1997.

BRIDGEMAN, B.; TRAPANI, C.; BIVENS-TATUM, J. Comparability of essay question variants. *Assessing Writing*, v. 16, n. 4, p. 237-255, 2011.

BROAD, B. Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English*, v. 35, n. 2, p. 213-260, 2000.

BROWN, G. T. L.; GLASSWELL, K.; HARLAND, D. Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, v. 9, n. 2, p. 105-121, 2004.

CASTRO, M. M. C. As Razões de uma ruptura: Elementos para uma história da prova de redação nos exames vestibulares isolados da UFRJ – 1987/88 – 2007/08. *Revista Contemporânea de Educação*, v. 3, n. 5, 2008.

CHALMERS, A. F. *O que é ciência afinal?* São Paulo: Brasiliense, 1995.

CHAPELLE, C. A. Validity in language assessment. *Annual Review of Applied Linguistics*. v. 19, p. 254-272, 1999.

COE, R. Common examinee methods. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 166-206. London: QCA, 2007.

_____. Understanding comparability of examination standards. *Research Papers in Education*, v. 25, n. 3, p. 271-284, 2010.

COE, R.; SEARLE, J.; BARMBY, P.; JONES, K.; HIGGINS S. *Relative difficulty of examinations in different subjects*. Report, CEM Centre, Durham University, 2008.

COHEN, J. Weighted kappa: nominal scale agreement or partial credit. *Psychological Bulletin*, v. 70, n. 4, 1968.

_____. Assessing written expression. In *Assessing language ability in the classroom*. Boston: Heinle & Heinle. 1994.

COHEN, A. S.; WOLLACK, J. A. Handbook on test development: Helpful tips for creating reliable and valid classroom tests. *Testing and Evaluation Services*, University of Wisconsin-Madison, 2004.

COLOMBINI, C. B.; McBRIDE, M. "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing writing*, v. 17, n. 4, p. 191-207, 2012.

COMVEST: Comissão Permanente para os Vestibulares. *Vestibular nacional UNICAMP 2011*: Manual do candidato. Pró-reitoria de graduação, 2010. Disponível em: <www.comvest.unicamp.br>. Acesso: 15 ago. 2012.

_____. *Vestibular nacional UNICAMP 2013*: Manual do candidato. Pró-reitoria de graduação, 2012. Disponível em: www.comvest.unicamp.br. Acesso: 25 fev. 2013.

_____. *Vestibular nacional UNICAMP 2012*: Primeira fase: Redação. Pró-reitoria de graduação, 2013. Disponível em: <www.comvest.unicamp.br>. Acesso: 25 fev. 2013.

COPS/UEL: Coordenadoria de Processos Seletivos. Universidade Estadual de Londrina. Disponível em: <www.cops.uel.br>. Acesso: 25 fev. 2014.

_____. *A UEL comenta suas provas: Vestibular 2012*. Diálogos Pedagógicos. Universidade Estadual de Londrina, 2012. Disponível em: <www.cops.uel.br/vestibular/2013/RevistaDialogosPedagogicos.pdf>. Acesso: 25 fev. 2013.

_____. *Manual do candidato: Vestibular 2013*. Universidade Estadual de Londrina, 2012. Disponível em: <http://www.cops.uel.br/vestibular/2013/manual_do_candidato.pdf>. Acesso: 25 fev. 2013.

CORTEZÃO, L. Formas de ensinar, formas de avaliar: breve análise de práticas correntes de avaliação. In. *Reorganização curricular do ensino básico*: avaliação das aprendizagens: das concepções às novas práticas. Universidade do Porto, 2002. Disponível em: <http://hdl.handle.net/10216/26195>. Acesso: 22 ago. 2012.

CRONBACH, L. J.; LINN, R. L.; BRENNAN, R. L.; HAERTEL, E. Generalizability Analysis for Educational Assessments, *Evaluation comment*, Summer, 1995.

CROMBACH, L J.; MEEHL, P. Construct validity in psychological tests. *Psychological Bulletin*, v. 52, n. 4, p. 281-302, 1955.

DARLING-HAMMOND, L.; SNYDER, J. Authentic assessment of teaching in context. *Teaching and Teacher Education*, n. 16, p. 523-545, 2000.

DEANE, P. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*. v. 18, p. 7-24, 2013

DE AYALA, R. J. *The theory and practice of item response theory*. New York: The Guilford Press, 2009.

DE CASTRO, M. H. G. A Consolidação da Política de Avaliação da Educação Básica no Brasil. *Meta:* Avaliação. v. 1, n. 3, p. 271-296, 2009.

DE SOUZA, A. R.; GOUVEIA, A. B. Os trabalhadores docentes da educação básica no Brasil em uma leitura possível das políticas educacionais. *Education Policy Analysis Archives*. v. 19, p. 1-22, 2011.

BEEL, J.; GIPP, B.; LANGER, S.; GENZMEHR M. Docear: *An academic literature suite for searching, organizing and creating academic literature*. In Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL 11), Ottawa, Ontario, Canada, p. 465-466, 2011. Disponível em: <http://www.docear.org/>. Acesso: 20 dez. 2011

DOWNING, S. M. Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine*, v. 77, n. 10, p. 103-104, 2002.

DOWNING, S. M.; HALADYNA,T. M. Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, n. 10, v. 1, p. 61-82, 1997.

EAST, M. Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, v. 14, p. 88-115, 2009.

ECKES, T. Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages*: Learning, teaching, assessment. 2009.

_____. Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessment. Frankfurt: *Peter Lang*. 2011.

ELLIOTT, G. A guide to comparability terminology and methods. *Research Matters*: A Cambridge Assessment Publication, Special Issue 2, 9-19, 2011.

_____. A guide to comparability terminology and methods. *Assessment Research & Development*, Cambridge assessment, 2013. Disponível em: <http://www.cambridgeassessment.org.uk/Images/130424-a-guide-to-comparability-terminology-and-methods.pdf>. Acesso em 17 out. 2013.

EMBRETSON, S. E.; REISE, S. P. emphItem response theory for psychologist. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000.

ENGELHARD, G. The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, v. 5, n. 3, p. 171-191, 1991.

_____. *Invariant measurement:* using Rasch models in the social, behavioral, and health sciences. New York: Routledge Academic, 2013.

ENGELHARD, G.; MYFORD, C. M. Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model. College Board Research Report, n. 1, New York: *College Entrance Examination Board*, 2003.

ENGELHARD, G.; MYFORD, C. M.; CLINE, F. Investigating assessor effects in National Board for Professional Teaching Standards assessments for early childhood/generalist and middle childhood/generalist certification. *Research Report-Educational Testing Service Priceton RR*, n. 13, 2000.

ENGELHARD, G.; WIND, S. A. Rating Quality Studies Using Rasch Measurement Theory. *Educational Testing Service* (ETS), Research Reports, Princeton, NJ, 2013.

ESFANDIARI, R.; MYFORD C. M. Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*. v. 18, p. 111-131, 2013.

ETS International Principles for Fairness Review of Assessments. A Manual for Developing Locally Appropriate Fairness Review Guidelines in Various Countries. Princeton, NJ: Educational Testing Service, 2009.

FAZAL, A.; HUSSAIN, F. K. H.; DILLON, T. S. An innovative approach for automatically grading spelling in essays using rubric-based scoring. *Journal of Computer and System Sciences*, 2013. Disponível em: <http://dx.doi.org/10.1016/j.jcss.2013.01.021>. Acesso: 22 jan. 1014.

FISCHER, G. H.; MOLENAAR, I. W. *Rasch models*: Foundations, recent developments and applications. New York: Springer-Verlag.

FONTANIVE, N.; KLEIN, R.; MARINO, L.; ABREU, M.; BIER, S. E. A alfabetização de crianças de 1° e 2° ano do Ensino Fundamental de 9 anos : uma contribuição para a definição de uma Matriz de Competências e Habilidades de Leitura , Escrita e Matemática. *Ensaio:* avaliação de políticas públicas na Educação, v. 18, n. 68, p. 527-548, 2010.

FREDERIKSEN, J. R.; COLLINS, A. A Systems Approach to Educational Testing. *Educational Researcher*. v. 18, n. 9, p. 27-32, 1989.

FUVEST: Fundação Universitária para o Vesstibular. *FUVEST 2014: Manual do candidato*. 2013. Disponível em: <www.fuvest.br.> Acesso: 22 out. 2013.

GEARHART, M.; HERMAN, J. L.; NOVAK, J. R.; WOLF, S. A. Toward the Instructional Utility of Large-ScpIe Writing Assessment: Validation of a New Narrative Rubric. *Assessing Writing*. v. 2, n. 2, p. 207-242, 1995.

GIL, A. C. *Métodos e técnicas de pesquisa social*. São Paulo: Atlas, 6 ed. 2008.

GIMENEZ, T. Concepções de linguagem e ensino na preparação de alunos para o vestibular. *Trabalhos em Linguística Aplicada*. n. 34, p. 21-37, 1999.

GOMES, C. M. A.; BORGES, O. O ENEM é uma avaliação educacional construtivista? Um estudo de validade de construto. *Estudos em Avaliação Educacional*. v. 20, n. 42, p. 73-88, jan./abr 2009.

GRAND, J. A.; GOLUBOVICH, J; RYAN, A. M.; SCHMITT, N. The detection and influence of problematic item content in ability tests: An examination of sensitivity review practices for personnel selection test development. *Organizational Behavior and Human Decision Processes*, n. 121, p. 158-173, 2013.

GREATOREX, J.; BAIRD, J.-A.; BELL, J. F. What makes rating reliable? Experiments with UK examinations. *Assessment in Education*, v. 11, n. 3, p. 331-347, 2004.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hil, 1954.

GUSTAFSON, J. E. Testing and obtaining fit of data to the Rasch model. *British Journal of mathematical and Statistical Psychology*, v.33, p. 220, 1980.

GYAGENDA, I. S.; ENGELHARD, G. Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, n. 10, p. 225-246, 2009.

HABERMAN, S. J. Maximum likelihood estimates in exponential response models. *The annals of statistics*, n. 5, p. 815-841, 1977.

HAERTEL, E. H.; LINN, R. L. Comparability. In G. PHILLIPS (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, p. 1-18, 1996.

HALADYNA, T. M.; DOWNING, S. M. Construct Irrelevant Variance in High-Stakes Testing. *Educational Measurement*: Issues and Practice, p. 17-27, Spring 2004.

HAMP-LYONS, L. Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing*. v. 16, n. 1, p. 3-5, 2011.

_____. Writing assessment: Expanding outwards and coming together. *Assessing Writing*, v. 13, n. 1, p. 1-3, 2004.

_____. Writing teachers as assessors of writing. In B. Kroll (Ed.). *Exploring the dynamics of second language writing*. Cambridge, England: Cambridge University Press, p. 162-189, 2003.

_____. The scope of writing assessment. *Assessing Writing*, v. 8, p. 5-16, 2002.

HAMP-LYONS, L.; MATHIAS, S. P. Examining Expert Judgments of Task Dificulty on Essay Tests. *Journal of second language writing*, v. 3, n. 1, p. 49-68, 1994.

HARSCH, C.; MARTIN, G. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, v. 17, n. 4, p. 228-250, 2012.

HARSCH, C.; RUPP, A. A. Designing and Scaling Level-Specific Writing Tasks in Alignment With the CEFR: A Test-Centered Approach. *Language Assessment Quarterly*, v. 8, n. 1, p. 1-33, 2011.

HOFFMAN, J. M. L. A controvérsia da redação no vestibular: questão de pertinência da prova ou de fidedignidade da medida? São Paulo: Fundação Carlos Chagas. *Educação e Seleção*, n. 17, 1988.

_____. How accurate are ESL students' holistic writing scores on large-scale Assessments? A generalizability theory approach. *Assessing Writing*, v. 13, n. 3, p. 201-218, 2008.

HUANG, J. Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, v. 17, n. 3, p. 123-139, 2012.

HUNG, S. P.; CHEN, P.H.; CHEN, H. C. Improving Creativity Performance Assessment: A Rater Effect Examination with Many Facet Rasch Model. *Creativity Research Journal*, v. 24, n. 4, p. 345-357, 2012.

HUOT, B. The literature of direct writing assessment: Major concern and prevailing trends. *Review of Educational Research*, v. 60, p. 237-264, 1990.

_____. Toward a New Theory of Writing Assessment. *National Council of Teachers of English*, v. 47, n. 4, p. 549-566, 1996.

JAEGER, R. M.; MULLIS, I. V. S.; BOURQUE, M. L.; SHAKRANI, S. Setting Performance Standards for Performance Assessments: Some Fundamental Issues, Current Practice, and Technical Dilemmas. In: G. PHILLIPS (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, p. 1-18, 1996.

JEFFERY, J. V. Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing*, v. 14, n. 1, p. 3-24, 2009.

JENNINGS, M.; FOX, J.; GRAVES, B.; SHOHAMY, E. The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing*, v. 16, p. 426-456, 1999.

JOHNSON, J. S.; LIM, G. S. The influence of rater language background on writing performance assessment. *Language Testing*, v. 26, n. 4, p. 485-505, 2009.

JOHNSTON, B. Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education*, v. 29, n. 3, p. 395-412, 2004.

JOHNSTONE, C. J.; THOMPSON, S. J.; BOTTSFORD-MILLER, N. A.; THURLOW, M. L. Universal design and multimethod approaches to item review. *Educational Measurement*: Issues and Practice, n. 27, p. 25-36, 2008.

JONES, N.; SHAW, S. D. SHAW. Task difficulty in the assessment of writing: Comparing performance across three levels of CELS. Cambridge ESOL Examinations: *Quarterly*, v. 11, 2003.

JONSSON, A.; SVINGBY, G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, v. 2, n. 2, p. 130-144, 2007.

JUZWIK, M. M.; CURCIC, S.; WOLBERS, K.; MOXLEY, K. D.; DIMLING, L. M.; SHANKLAND, R. K. Writing Into the 21st Century: An Overview of Research on Writing, 1999 to 2004. *Written Communication*, v. 23, p. 451-476, 2006.

KANE, M. T. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*. v. 50, n. 1, p. 1-73, 2013.

KANE, M. T.; COOKS, T.; COHEN, A. Validating measures of performance. *Educational Measurement*: Issues and Practice, v. 18, n. 2, p. 5-17, 1999.

KELLEY, T. L. *Interpretation of educational measurements*. New York: Macmillan, 1927.

KLEIN, R.; FONTANIVE, N. Uma nova maneira de avaliar as competências escritoras na redação do ENEM. *Ensaio*: avaliação e políticas públicas em educação, Rio de Janeiro, v. 17, n. 65, p. 585-598, 2009a.

_____. Alguns indicadores educacionais de qualidade no Brasil de hoje. *São Paulo em Perspectiva*, v. 23, n. 1, Pages 19-28, 2009b.

KLEIN, R.; FONTANIVE, N.; RESTANI, A. L.; TELLES, M. C. O desempenho dos alunos da Fundaçãoo Bradesco: uma comparação com os resultados do SAEB. *Estudos em Avaliação Educacional*. v. 19, n. 41, p. 499-515, 2008.

KNOCH, U. Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, v. 26, n. 20, p.275-304, 2011a.

_____. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, v. 16, n. 2, p. 81-96, 2011b.

KNOCH,U.; ELDER, C. Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System*. Elsevier, v. 38, n. 1, p. 63-74, 2010.

KNOCH,U.; READ, J.; RANDOW, J. V. Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, v. 12. p. 26-43, 2007.

KOBRIN, J. L.; DENG, H.; SHAW, E. J. The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing*. Elsevier, v.16, p. 154-169, 2011.

KROLL, B.; REID, J. Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing*, v. 3, n. 3, p. 231-255, 1994.

LAKATOS, E. M.; MARCONI, M. de A. *Fundamentos de metodologia científica*. 6. ed. 5. reimp. São Paulo: Atlas, 2007.

LEE, H. K.; ANDERSON, C. Validity and topic generality of a writing performance test. *Language Testing*, v. 24, n. 3, p. 307-330, 2008.

LEE, Y. W.; GENTILE, C.; KANTOR, R. Toward Automated Multi-trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics*. Oxford: Oxford University Press, v. 31, n. 3, p. 391-417, 2009.

LI, H. The resolution of some paradoxes related to reliability and validity. *Journal of Educational and Behavioral Statistics*, Thousand Oaks, CA: Sage, v. 28, n. 2, p. 89-95, 2003.

LIM, G. S. The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, v. 28, n. 4, p. 543-560, 2011.

LIMA, A. C.; PEQUENO, M. I. C.; MELO, M. N. R. Avaliação da alfabetização no Ceará: principais resultados da primeira edição do Spaece-Alfa. *Estudos em Avaliação Educacional*. v. 19, n. 41, p. 465-482, 2008.

LINACRE, J. M. *Many-facet Rasch measurement*, 2nd ed. Chicago: MESA Press, 1994.

_____. Detecting Multidimensionality: Which Residual Data-type Works Best? *Journal of Outcome Measurement*, v. 2, n. 3, p. 266-283, 1998.

_____. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, v. 3, n. 1, p. 85-106, 2002a.

_____. Judging debacle in pairs figure skating. *Rasch Measurement Transactions*, v. 15, p. 839-840, 2002b.

_____. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, n. 16, p. 878, 2002c.

_____. A user's guide to FACETS [computer program manual 3.71.4]. Chicago: *MESA Press*. 2014a.

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

LINACRE, J. M.; ENGLEHARD, G.; TATUM, D. S.; MYFORD, C. M. Measurement with judges: many-faceted conjoint measurement. *International Journal of Educational Research*, v. 21, n. 4, p. 569-577, 1994.

LINACRE, J. M.; WRIGHT, B. D. The "Length" of a Logit. *Rasch Measurement Transactions*, v. 3, n. 2, p. 54-55, 1989.

_____. Construction of measures from Many-Facet Data. *Journal of Applied Measurement*, v. 3, n. 4, p. 484-509, 2002.

LINN R. L. Linking results of distinct assessments. *Applied Psychological Measurement*. Thousand Oaks, CA: Sage, v. 6, p. 83-102, 1993.

LIPMAN, M. *O pensar na educação*. Petrópolis: Vozes, 1995.

LORD, F. M. A.; NOVICK, N. R. *Statistical Theories of mental test scores*. Massachusetts: Addison Weslley. 1968.

LOYD-JONES, R. Primary trait scoring. In Charles Raymond Cooper; Lee Odell (Eds.), *Evaluating writing*: Describing, measuring, judging. Urbana, IL: National Council of Teachers of English. p. 33-66, 1977.

LUNZ, M. E.; WRIGHT, B. Latent trait models for performance examinations. *Applications of latent trait and latent class models in the social sciences*, 1997.

MARTINS, R. A. Princípios da pesquisa científica. In CAUCHICK, P. A. (coord.) *Metodologia de pesquisa em engenharia de produção e gestão de operações*. Rio de Janeiro: Elsevier, 2010.

McMANUS, I. C., ELDER, A. T, DACRE, J. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC Medical Education*, v. 13:103, 2013. Disponível em: <http://www.biomedcentral.com/1472-6920/13/103>. Acesso em abr. 2014.

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

McNAMARA, T.; KNOCH, U. The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, v. 29, n. 4, p. 555-576, 2012.

MENDELEY DESKTOP. Versão: 1.7.1. 2008-2012. Mendeley LTD. Disponível em: <http://www.mendeley.com/>

MESSICK, S. Validity of performance assessments. In: G. PHILLIPS (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, p. 1-18, 1996.

_____. Validity. In: LINN, R. (Ed.). *Educational Measurement*. 3nd ed. New York: Macmillan, p. 13-103, 1989.

MISLEVY, R. J. Linking Educational Assessments: Concepts, Issues, Methods, and Prospects. Princeton, NJ: *Educational Testing Service*, 1992.

_____. Can There Be Reliability without "Reliability?" *Journal of Educational and Behavioral Statistics*. Thousand Oaks, CA: Sage, v. 29, n. 2, jan, p. 241-244, 2004.

MISLEVY, R. J.; HAERTEL, G. D. Implications of evidence-centered design for educational testing. *Educational Measurement*: Issues and Practice, v. 25, p. 6-20, 2006.

MITCHELL, A. W.; McCONNELL, J. R. A historical review of Contemporary Educational Psychology from 1995 to 2010. *Contemporary Educational Psychology*, Elsevier Inc., v. 37, p. 136-147, 2012.

MORAES, Z. H. O vestibular em discussão. *Estudos em Avaliação Educacional*, v. 15, p. 199-226, 1997.

MOSS, P. A. Can there be validity without reliability? *Educational Researcher*. Thousand Oaks, CA: Sage, v. 23, n. 2, p. 5-12, 1994.

MOREIRA JUNIOR, F. J. *Sistemática para implantação de testes adaptativos informatizados baseados na teoria da Resposta ao Item*. (Tese de doutorado). Universidade Federal de Santa Catarina. Programa de Pós-Graduação em engenharia de Produção. 2011.

MORETO, M. *Modelo de teses e dissertações do PPGEEL UFSC*. 2009. Disponível em: <http://code.google.com/p/pgeeltex/>. Acesso: 22 abr. 2014.

MOSKAL, B. M. Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation*, v. 7, n. 3, 2000.

MOSKAL, B. M.; LEYDENS, J. A. Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*. v. 7, n. 10, p. 71-81, 2000.

MYFORD, C. M. Investigating Design Features of Descriptive Graphic Rating Scales. *Applied Measurement in Education*, v. 15, n. 2, p. 187-215, 2002.

MYFORD, C. M.; WOLFE, E. W. Monitoring sources of variability within the Test of Spoken English assessment system. (Research Project 65). Princeton, NJ: *Educational Testing Service*, 2000.

_____. Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. journal of applied measurement, v. 5, n. 2, p. 189-227, 2004.

NETTO, A. R. O Vestibular ao longo do tempo: implicações e implicâncias. *Seminário*: Vestibular Hoje. Brasilia: MEC/SESU/CAPES, dez., 1985.

NEWTON, P. E. Comparability monitoring: progress report. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 166-206. London: QCA, 2007.

_____. Exploring tacit assumptions about comparability. Paper presented at the 34^{th} annual conference of the International Association for Educational Assessment, September 7-12, Cambridge, 2008.

NYSTRAND, M.; COHEN, A. S.; DOWLING, N. M. Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, v. 1, p. 53-70, 1993.

NOBRE, J. C. S. *Modelo computacional para valoração e avaliação de redações baseado em lógica Fuzzi*. (Tese de doutorado). Instituto Tecnológico de Aeronáutica (ITA). Engenharia eletrônica e computação. 2011.

NORTH, B. Linking language assessments: an example in a low stakes context. *System*. v. 28, n. 4, 2000.

PAGANO, N.; BERNHARDT, S. A.; REYNOLDS, D.; WILLIAMS, M.; McCURRIE, M. K. An Inter-Institutional Model for College Writing Assessmen. *College Composition and Communication*. v. 2, p. 285-320. 2008.

PARRA-LÓPES, E.; OREJA-RODRÍGUES, J. R. Evaluation of the competiveness of tourist zones of an island destination: An application of a Many-Facet Rasch Model (MFRM). *Journal of Destination Marketing & Management*, 2014. Disponível em: <http://dx.doi.org/10.1016/j.jdmm.2013.12.007i.>. Acesso: abr. 2014.

PASQUALI, L. Validade dos Testes Psicológicos: Será Possível Reencontrar o Caminho? *Psicologia*: Teoria e Pesquisa. Brasilia: Instituto de Psicologia, Universidade de Brasília, v. 23, n. especial, p. 099-107, 2007.

_____. Testes referentes a construtos: teoria e modelos de construção. In: PASQUALI, L. (org) *Instrumentação psicológica*: fundamentos e práticas. Porto Alegre: Artmed, 2010.

PENNY, J. A. Reading high stakes writing samples: My life as a reader. *Assessing Writing*, v. 8, 192-215, 2003.

PENNY, J.; JOHNSON, R.; GORDON, B. The effect of rating augmentation on inter-rater reliability an empirical study of a holistic rubric. *Assessing writing*, v. 7, 2000.

PHILLIPS, G. (Ed.). *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics, p. 1-18, 1996.

PINHO FILHO, A. G. O vestibular da Universidade de São Paulo: modelo adotado em 1995. *Estudos em Avaliação Educacional*, n. 11, p. 53-92, 1996.

POLLITT, A.; AHMED, A.; Crisp, V. The demands of examination syllabuses and question papers. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 166-206. London: QCA, 2007.

POMPLUN, M.; WRIGHT, D.; OLEKA, N.; SUDLOW, M. An Analysis of English Composition Test Essay Prompts for Differential Difficulty. College Board Report, New York: *College Entrance Examination Board*, 1992.

POPHAM, W. J. "What's wrong and what's right with rubrics". *Educational Leadership*, v. 55, n. 2, p. 72-75, 1997.

PRODANOV, C. C.; FREITAS, E. C. *Metodologia do trabalho Científico*[recurso eletrônico]: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013. Disponível em:<http://www.hugoribeiro.com.br/biblioteca-digital/FEEVALE-Metodologia_Trabalho_Cientifico.pdf>. Acesso em: 31 out. 2013.

QUEVEDO-CAMARGO, G. Efeito retroativo da avaliação na aprendizagem de línguas. In: *Centro de Estudos Linguísticos e Literários do Paraná CELLIP*, 2011, Londrina. Anais do Seminário do Centro de Estudos Linguísticos e Literários do Paraná, 2011. p. 1-16.

_____. Efeito retroativo da avaliação na aprendizagem de línguas estrangeiras: que fenômeno é esse? In: MULIK, K. B.; RETORTA, M. S. (Org.) *Avaliação no ensino-aprendizado de línguas estrangeiras*: diálogos, pesquisas e reflexões. Campinas, SP: Pontes Editores, 2014.

RASCH, G. *Probabilistic models for some intelligence and attainment tests*, Copenhagen: Danish Institute for Educational Research, 1960.

RAMINENI, C. Validating automated essay scoring for online writing placement. *Assessing Writing*. Elsevier, v. 18, n. 1, p. 40-61, 2013.

REDDY, M. Design and development of rubrics to improve assessment outcomes: A pilot study in a Master's level business program in India. *Quality Assurance in Education*, v. 19, n. 1, p. 84-104, 2011.

REZAEI, A. R.; LOVORN, M. Reliability and validity or rubrics for assessment through writing. *Assessing Writing*, v. 15, p.18-39, 2010.

ROCCO, M. T. F. O vestibular e a prova de Redação: mais concordâncias, menos controvérsias. *Estudos em Avaliação Educacional*, n. 11, 1995.

RUTH, L.; MURPHY, S. Designing Topics for Writing Assessment: Problems of Meaning. *College Composition and Communication*, v. 35, n. 4, p. 410-422, 1984.

SAN MARTIN, E.; GONZÁLEZ, J.; TUERLINCKX, F. Identified parameters, parameters of interest and their relationships. *Measurement*: Interdisciplinary Research and Perspective, v. 7, p. 97-105, 2009.

SAN MARTIN, E. ROLIN, J. M. Identification of parametric Rasch-type models. *Journal of Statistical Planning and Inference*, 2013.

NOVENTA, S.; TEFANUTTI, L.; VIDOTTO, G. An analysis of item response theory and Rasch models based on most probable distribution method. *Psychometrika*, v. 79, n. 3, p. 377-402, 2014.

SAXTON, E.; BELAGER, S.; BECKER, W. The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing*, n. 17, p. 251-270, 2012.

SCARAMUCCI, M. V. R. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte . *Trabalhos em Linguística Aplicada*. Campinas, SP, v. 2, n. 43, p. 203-226, 2004.

_____. Validade e consequências sociais das avaliações em contexto de ensino de línguas. *LINGVARVM ARENA*, Porto: Universidade do Porto, v. 2, p. 103-120, 2011.

SERVISS, T. A history of New York state literacy test assessment: Historicizing calls to localism in writing assessment. *Assessing Writing*, Elsevier Inc., v. 17, p. 208-227, 2012.

SOSSAI, J. A.; SOSSAI, A.; CARVALHO, D. A. Provas objetivas e dissertativas nos vestibulares: a experiência da U.F. do Espírito Santo, Est. Aval. Educ., n.12, p. 103-117. 1995.

SLOMP, D. H. Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, v. 17, n. 2, p. 81-91, 2012.

SLOMP, D. H.; FUITE, J. Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing*. Elsevier, n. 9, p. 190-207, 2005.

SMITH, R. M. et al. Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, v. 2, p. 66-78, 1998.

STEMLER, S. E. An overview of content analysis. *Practical Assessment, Research and Evaluation*, v. 7, n. 17, 2001. Disponível em: <http://PAREonline.net/getvn.asp?v=7&n=17>. Acesso em: 22 maio 2013.

_____. A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, v. 9, 2004.

SUDWEEKS, R. R.; REEVE, S.; BRADSHAW, W. S. A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*. Elsevier, v. 9, p. 239-261, 2005.

TATTERSALL, K. A brief history of policies, practices and issues relating to comparability. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, P. Tymms, 43-96. London: QCA, 2007.

TENNANT A., PALLANT J. F. Unidimensionality Matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, v. 20, n. 1, p. 1048-51. 2006.

TEZZA, R.; BORNIA, A. C.; ANDRADE, F. A. Measuring web usability using item response theory: Principles, features and opportunities. *Interacting with Computers*, v. 23, p. 167-175, 2011.

TRAUB, R. E. Classical Test Theory in Historical Perspective. *Educational Measurement Issues and Practice*, Winter 1997.

U.S. Department of Education, National Center for Education Statistics. The NPEC Sourcebook on Assessment, Volume 1: *Definitions and Assessment Methods for Critical Thinking, Problem Solving, and Writing, NCES 2000-172*, prepared by T. Dary Erwin for the Council of the National Postsecondary Education Cooperative Student Outcomes Pilot Working Group: Cognitive and Intellectual Development. Washington, DC: U.S. Government Printing Office, 2000.

_____. *The Nation's Report Card*: Writing 2002, Trial Urban District Assessment, NCES 2003-530, by A. D. Lutkus, M. C. Daane, A. W. Weiner, and Y. Jin. Washington, DC: 2003.

_____. *Online Assessment in Mathematics and Writing*: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series (NCES 2005-457). by SANDENE, B.; HORKAY, N.; BENNETT, R.; ALLEN, N.; BRASWELL, J.; KAPLAN, B.; ORANJE, A. Washington, DC 2005.

_____. *Writing framework for the 2011 national assessment of educational progress*. Washington: National Assessment Governing Board, U.S. Department of education. 2010.

U.S. Department of Labor Employment and Training Administration. *Testing and assessment*: an employer's guide to good practices. Washington, DC: Author. 2000.

WEIGLE, S. C. Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing*, v. 6, n. 2, p. 145-178, 1999.

_____. *Assessing Writing*. New York: Elsevier, Cambridge University Press. 2002.

_____. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*. Elsevier, v. 18, p. 85-99, 2013.

WHITE E. M. Holisticism. *College Composition and Communication*, n. 35, p. 400-409, 1984.

WIGGINS, G. The Constant Danger of Sacrificing Validity to reaeliability: Making Writing Assessment Serve Writers. *Assessing Writing*. v. 1, n. 1, p. 129-139, 1994.

WILLIAMSON, D.; XI, X.; BREYER, F. J. A framework for evaluation and use of automated scoring. *Educational Measurement*: Issues and Practice, v. 31, n. 1, p. 2-13, 2012.

WISEMAN, C. S. Rater effects: Ego engagement in rater decision-making. *Assessing Writing*. n. 17, p. 150-173, 2012.

WORDEN, D. L. Finding process in product: Prewriting and revision in timed essay responses. *Assessing Writing*, v. 14, n. 3, p. 157-177, 2009.

WRIGHT, B. D.; LINACRE, J. M. Observations are always ordinal; measurements, however, must be interval. Chicago: *Mesa psychometric laboratory*, MESA Research Memorandum, n. 44, 1987.

_____. Reasonable mean-square fit values. *Rasch Measurement Transactions*, v. 8, n.3, p. 370, 1994.

WRIGHT, B. D.; PANCHAPAKESAN, N. A. A procedure for sample-free item analysis. *Educational and Psychologic al Measurement*, N. 29, pp. 2-48, 1969.

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

VAN MOERE, A. Validity evidence in a university group oral test. *Language Testing*, v. 23, p. 411-440, 2006.

VIANNA. H. M. Redação e medida da expressão escrita: algumas contribuiçóes da pesquisa educacional. *Cadernos de Pesquisa*, v. 16, p. 41-47. 1976a.

_____. Flutuações de julgamentos em provas de redação. *Cadernos de Pesquisa*, n. 19, p. 5-9. 1976b.

_____. Aplicação de critérios de correção em provas de redação. *Cadernos de Pesquisa*. São Paulo: Fundação Carlos Chagas, n. 26, p. 29-34, 1978.

_____. *Testes em educação*. 4 ed. São Paulo: Ibrasa, 1982.

_____. Os novos modelos de vestibular: preocupações metodológicas. *Estudos em avaliação educacional*, n. 11, p. 47-52, 1995.

_____. Avaliações nacionais em larga escala: análises e propostas. *Estudos em Avaliação Educacional*. São Paulo: Fundação Carlos Chagas, n. 27, 2003.

VIANELLO, M.; ROBUSTO, E. The many-facet Rasch model in the analysis of the go/no-go association task, *Behavior Research Methods*, v.42, n. 4, p. 944-956, 2010.

VICENTINI, M. P. *Exame nacional do Ensino Médio*: A relevância de pesquisas empíricas sobre validade e efeitos retroativos. 2011. 75. Dissertação de mestrado. Universidade Estadual de Campinas - Instituto de estudos da linguagem.

VIDAL, E. M.; FARIAS, I. M. S. Avaliação da Aprendizagem e Política Educacional: desafios para uma nova agenda. *Estudos em Avaliação Educacional*, São Paulo: Fundação Carlos Chagas, v. 19, n. 40, 2008.

VILLAS, M. V.; VAN ADUARD MACEDO-SOARES, T. D. L.; RUSSO, G. M. Bibliographical research method for business administration studies: a model based on scientific journal ranking. *Brazilian Administration Review*, v. 5, n. 2, p. 139-159, 2008.

YANCEY, K. B. Looking Back as We Look Forward: Historicizing Writing Assessment. *College Composition and Communication*. Urbana, IL: National Council of Teachers of English, v. 50, n. 3, p. 483-503, 1999.

YANG, H. C. Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes*. v.31. n. 3. p. 174-187, 2012.

YAO, L. Multidimensional Linking for Domain Scores and Overall Scores for Nonequivalent Groups. *Applied Psychological Measurement*, v. 35, n. 1, p. 48-66, 2011.

ZAINAL, A. Validation of an ESL writing test in a Malaysian secondary school context. *Assessing Writing*, v. 17, n. 1, p. 1-17, 2012.

APÊNDICE A – CRITÉRIOS DE AVALIAÇÃO UTILIZADOS PARA A PONTUAÇÃO DAS TAREFAS

- **Competência 1. Demonstrar domínio da norma padrão da língua escrita.**

  O candidato deve demonstrar conhecimento das regras gramaticais de:

  a) concordância nominal e verbal;
  b) regência nominal e verbal;
  c) pontuação;
  d) flexão de nomes e verbos;
  e) colocação de pronomes átonos;
  f) grafia das palavras;
  g) acentuação gráfica;
  h) emprego de letras maiúsculas e minúsculas;
  i) divisão silábica na mudança de linha (translineação).

- **Competência 2. Desenvolver o tema dentro dos limites estruturais de um texto dissertativo.**

  Um texto dissertativo deve ser escrito com a finalidade de expressar uma ideia ou expor uma opinião sobre um determinado assunto, com argumentos lógicos e buscando convencer o leitor. Deve ser organizado da seguinte forma:

  Primeiro parágrafo: Consiste na introdução, deve apresentar a ideia principal da dissertação. Pode conter fatos históricos, exemplos, dados estatísticos, pensamento filosófico, comparações diversas. Pode também conter perguntas, desde que sejam respondidas durante o texto.

  Parágrafos intermediários: Consiste na argumentação e desenvolvimento do tema. O autor deve explicar a ideia principal e tentar convencer o leitor sobre o seu ponto de vista por meio de argumentos, explicações ou dados. O texto não deve ser escrito na primeira pessoa.

  Último parágrafo: É a conclusão do texto, que pode ser feita por meio de um resumo, de um questionamento ou mesmo de uma proposta para solucionar o problema apresentado no texto.

- **Competência 3. Atender os requisitos relacionados ao propósito e à leitura.**

  Devem ser estabelecidos pontos de contato com o material fornecido para a tarefa. O escritor deve mostrar a relevância desses pontos para o seu projeto de escrita e não simplesmente copiar partes dos textos. Devem ser apresentadas informações, fatos e opiniões relacionados ao tema proposto para a elaboração da argumentação e desenvolvimento do texto. O autor deve explicar a ideia principal e tentar convencer o leitor sobre o seu ponto de vista.

- **Competência 4. Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.**

  Os textos produzidos pelo candidato devem propiciar uma leitura fluida e envolvente, mostrando uma articulação entre as partes do texto apoiada na utilização adequada e diversificada de recursos coesivos para a sua organização.

- **Competência 5. Aplicar conceitos das várias áreas de conhecimento e vocabulário rico e variado.**

  O candidato deve aplicar conceitos das várias áreas de conhecimento, com vocabulário rico e variado para desenvolver o tema dentro dos limites estruturais do texto dissertativo. O texto deve refletir o conhecimento de mundo do autor e a coerência da argumentação.

APÊNDICE B – CRITÉRIOS DE AVALIAÇÃO E NÍVEIS DE DESEMPENHO UTILIZADOS PARA A PONTUAÇÃO DAS TAREFAS

Os Quadros 22, 23, 24, 25 e 26 apresentam os seis níveis de desempenho que foram utilizados para avaliar cada uma das competências nos textos escritos pelos participantes da avaliação. Esses quadros consistem nos critérios de avaliação para o evento de avaliação.

Quadro 22 – Competência 1: Demonstrar domínio da norma padrão da língua escrita

6 pontos	Não cometeu erros em relação às regras gramaticais. Demonstra excelente domínio da norma padrão da língua escrita.
5 pontos	Cometeu poucos erros gramaticais sem apresentar reincidência, demonstra bom domínio da norma padrão da língua escrita.
4 pontos	Cometeu alguns erros gramaticais. Demonstra domínio mediano da norma padrão da língua escrita.
3 pontos	Cometeu muitos erros gramaticais. Demonstra domínio insuficiente da norma padrão da língua escrita.
2 pontos	Cometeu muitos erros, de forma sistemática, diversificados e frequentes. Demonstra domínio precário da norma padrão da língua escrita.
1 ponto	Demonstra desconhecimento total da norma padrão da língua escrita.

Fonte: Autora

Quadro 23 – Competência 2: Compreender o propósito da tarefa e desenvolver o tema dentro dos limites estruturais de um texto dissertativo

6 pontos	Toma uma posição clara e com sucesso excepcional expressa um ponto de vista. Desenvolve o texto demonstrando excelente domínio da estrutura de texto dissertativo, com introdução, argumentação e conclusão.
5 pontos	Toma uma posição em defesa de um ponto de vista e de forma competente desenvolve o texto demonstrando bom domínio da estrutura de texto dissertativo, com introdução, argumentação e conclusão.
4 pontos	Toma uma posição em defesa de um ponto de vista e de forma adequada desenvolve o texto demonstrando domínio mediano da estrutura de texto dissertativo, com introdução, argumentação e conclusão.
3 pontos	Toma uma posição em defesa de um ponto de vista e desenvolve o texto demonstrando domínio insuficiente da estrutura de texto dissertativo, com introdução, argumentação e conclusão.
2 pontos	Não toma uma posição em defesa de um ponto de vista e desenvolve o texto demonstrando domínio precário da estrutura de texto dissertativo, com traços constantes de outros tipos textuais.
1 ponto	Não toma uma posição em defesa de um ponto de vista e desenvolve o texto não atendendo à estrutura de texto dissertativo.

Fonte: Autora

Quadro 24 – Competência 3: Atender os requisitos relacionados ao propósito e à leitura

6 pontos	Desenvolve o tema por meio de argumentação consistente e utiliza muito bem o material fornecido como suporte para o seu projeto de escrita, sem divagações. O autor convence plenamente o leitor sobre o seu ponto de vista.
5 pontos	Desenvolve o tema por meio de argumentação consistente e utiliza bem o material fornecido como suporte para o seu projeto de escrita, com divagações ocasionais. O autor convence o leitor sobre o seu ponto de vista.
4 pontos	Desenvolve o tema por meio de argumentação previsível e utiliza razoávelmente o material fornecido como suporte para o seu projeto de escrita, com algumas divagações. O autor convence medianamente o leitor sobre o seu ponto de vista.
3 pontos	Desenvolve o tema recorrendo à cópia de trechos dos textos auxiliares e utiliza de modo precário o material fornecido. O autor convence de modo frágil o leitor sobre o seu ponto de vista.
2 pontos	Desenvolve o texto tangenciando o tema e utiliza de modo precário o material fornecido. O autor não convence o leitor sobre o seu ponto de vista.
1 ponto	Fuga ao tema.

Fonte: Autora

Quadro 25 – Competência 4: Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação

6 pontos	Articula muito bem as partes do texto e utiliza recursos coesivos de forma adequada e diversificada.
5 pontos	Articula bem as partes do texto e utiliza recursos coesivos com poucas inadequações e de forma diversificada.
4 pontos	Articula medianamente as partes do texto e utiliza recursos coesivos com algumas inadequações e de forma pouco diversificada.
3 pontos	Articula insuficientemente as partes do texto e utiliza recursos coesivos com muitas inadequações e de forma limitada.
2 pontos	Articula precariamente as partes do texto.
1 ponto	Ausência de marcas de articulação, resultando em fragmentação das ideias.

Fonte: Autora

Quadro 26 – Competência 5: Aplicar conceitos das várias áreas de conhecimento e vocabulário rico e variado

6 pontos	Demonstra excelente conhecimento de mundo e extensa gama de vocabulário para fins comunicativos. Escolha de palavras apropriadas.
5 pontos	Demonstra bom conhecimento de mundo e competente gama de vocabulário para fins comunicativos. Evidência de escolha de palavras apropriadas.
4 pontos	Demonstra mediano conhecimento de mundo e adequada gama de vocabulário para fins comunicativos. Algumas evidências de escolha de palavras apropriadas.
3 pontos	Demonstra pobre conhecimento de mundo e mínima gama de vocabulário para fins comunicativos. Evidência mínima de escolha de palavras apropriadas.
2 pontos	Demonstra precário conhecimento de mundo e gama limitada de vocabulário para fins comunicativos. Possibilidade de escolha das palavras apropriadas.
1 ponto	Não demonstra conhecimento de mundo e mostra estreita faixa de vocabulário para fins comunicativos. Pouca evidência de escolha de palavras apropriadas.

Fonte: Autora

B.1 PONTUAÇÃO ANALÍTICA

Competência	6 pontos	5 pontos	4 pontos	3 pontos	2 pontos	1 ponto
1. Norma padrão da língua escrita	• Não cometeu erros gramaticais	• Cometeu poucos erros gramaticais	• Cometeu alguns erros gramaticais	• Cometeu muitos erros gramaticais	• Cometeu muitos erros de forma sistemática, diversificados e frequentes	• Demonstra desconhecimento total da norma padrão da língua escrita
	• Demonstra domínio excelente da norma padrão da língua escrita	• Demonstra bom domínio da norma padrão da língua escrita	• Demonstra domínio mediano da norma padrão da língua escrita	• Demonstra domínio insuficiente da norma padrão da língua escrita	• Demonstra domínio precário da norma padrão da língua escrita	
2. Compreensão da tarefa	• Toma uma posição clara e com sucesso excepcional expressa um ponto de vista	• Toma uma posição e de forma competente expressa um ponto de vista	• Toma uma posição e de forma adequada expressa um ponto de vista	• Toma uma posição em defesa de um ponto de vista	• Não toma uma posição em defesa de um ponto de vista	• Não toma uma posição em defesa de um ponto de vista
	• Desenvolve o texto demonstrando excelente domínio da estrutura de texto dissertativo, com introdução, argumentação e conclusão	• Desenvolve o texto demonstrando bom domínio da estrutura de texto dissertativo, com introdução, argumentação e conclusão	• Desenvolve o texto demonstrando domínio mediano da estrutura de texto dissertativo, com introdução, argumentação e conclusão	• Desenvolve o texto demonstrando domínio insuficiente da estrutura de texto dissertativo, com introdução, argumentação e conclusão	• Desenvolve o texto atendendo precariamente a estrutura de texto dissertativo, com traços constantes de outros tipos textuais	• Desenvolve o texto não atendendo a estrutura de texto dissertativo
3. Atender os requisitos relacionados ao propósito e à leitura	• Desenvolve o tema por meio de argumentação consistente	• Desenvolve o tema por meio de argumentação consistente	• Desenvolve o tema por meio de argumentação previsível	• Desenvolve o tema recorrendo à cópia de trechos dos textos auxiliares	• Desenvolve o tema tangenciando o tema	• Fuga ao tema
	• Utiliza muito bem o material fornecido, sem divagações	• Utiliza bem o material fornecido, com divagações ocasionais	• Utiliza de modo razoável o material fornecido, com divagações frequentes	• Utiliza de modo precário o material fornecido	• Utiliza de modo precário o material fornecido	
4. Mecanismos linguísticos	• Articula muito bem as partes do texto	• Articula bem as partes do texto	• Articula medianamente as partes do texto	• Articula insuficientemente as partes do texto	• Articula precariamente as partes do texto	• Ausência de marcas de articulação resultando em fragmentação das ideias
	• Utiliza recursos coesivos de forma adequada e diversificada	• Utiliza recursos coesivos com poucas inadequações e de forma diversificada	• Utiliza recursos coesivos com algumas inadequações e de forma pouco diversificada	• Utiliza recursos coesivos com muitas inadequações e de forma limitada	• Utiliza inadequadamente os recursos coesivos	
5. Conhecimento de mundo e vocabulário	• Demonstra excelente conhecimento de mundo	• Demonstra bom conhecimento de mundo	• Demonstra mediano conhecimento de mundo	• Demonstra pobre conhecimento de mundo	• Demonstra precário conhecimento de mundo	• Não demonstra possuir conhecimento de mundo
	• Utiliza extenso vocabulário para fins comunicativos	• Utiliza vocabulário competente para fins comunicativos	• Utiliza vocabulário adequado para fins comunicativos	• Utiliza vocabulário básico para fins comunicativos	• Utiliza vocabulário limitado para fins comunicativos	• Utiliza estreita faixa de vocabulário para fins comunicativos

Fonte: Autora

B.2 PONTUAÇÃO HOLÍSTICA

6 pontos	Texto excepcionalmente bem escrito, toma uma posição clara e com sucesso expressa um ponto de vista. As ideias são completamente desenvolvidas, com exemplos ricos e fornece pelo menos dois pontos de contato relacionados com o material de apoio fornecido. O ensaio é claramente e logicamente organizado e sem divagações. O escritor utiliza recursos coesivos de forma adequada e diversificada e demonstra possuir extensa gama de vocabulário para fins acadêmicos, com poucos problemas na escolha ou uso da palavra. Alguns erros gramaticais são perceptíveis porém raramente esses erros interferem com o sentido da frase. A variedade e complexidade das sentenças refletem um ótimo conhecimento das normas padrão da língua escrita.
5 pontos	Texto escrito competentemente com posição clara mas pode apresentar algumas divagações. O escritor fornece apoio substancial para o desenvolvimento das ideias, embora alguns exemplos não são totalmente relevantes ou apropriados para o tema. O trabalho é organizado de forma eficaz, o que demonstra o uso de dispositivos coesos de forma competente e diversificada mas pode apresentar alguns erros. A faixa de vocabulário para fins acadêmicos é competente e o escritor geralmente demonstra um controle preciso e apropriado de escolha de palavras e expressões idiomáticas para a escrita acadêmica. Possui alguns erros relacionados com as normas padrão da língua escrita, mas esses erros geralmente não interferem com o sentido.
4 pontos	Texto escrito adequadamente, a posição do escritor é clara, apesar de algumas divagações e contradições. O escritor fornece suporte adequadamente detalhado de dois ou mais pontos que se relacionam diretamente com o tema. O trabalho é geralmente organizado, demonstrando uso adequado de recursoss coesivos. O escritor utiliza alguma variedade de frases simples e, algumas vezes, frases complexas, embora nem sempre corretamente. O ensaio pode conter erros frequentes que, ocasionalmente, podem prejudicam o sentido. O vocabulário é adequado para fins acadêmicos, mas a utilização de algumas palavras são inapropriadas ou imprecisas.
3 pontos	O texto consegue minimamente expor uma posição relacionada com um padrão organizacional discernível (introdução, argumentação, conclusão), embora o foco no desenvolvimento da ideia central não é claro. O escritor utiliza exemplos, na maioria, irrelevantes para o desenvolvimento do tema. O escritor faz uso mínimo de recursos coesivos e demonstra uma faixa mínima de variedade de sentenças e vocabulário, com a utilização de palavras imprecisas ou inadequadas. Possui domínio mínimo das normas padrão da língua escrita, com erros frequentes e alguns deles prejudicam o sentido da sentença.
2 pontos	O texto possui sucesso limitado para representar um ensaio com alguma estrutura organizacional (introdução, argumentação, conclusão). O escritor fornece limitado desenvolvimento do tema com um ou mais pontos que direta ou indiretamente se relacionam com o material de apoio fornecido. A escrita mostra evidências limitadas de organização de idéias ou uso apropriado de dispositivos coesivos. A gama de vocabulário e a escolha de palavras apropriadas para a escrita acadêmica é limitada. O domínio da língua padrão é desigual, com erros frequentes e que resultam em significados obscuros. A escrita carece de variedade sentenças.
1 ponto	O texto é uma tentativa fracassada para representar um ensaio. O escritor não desenvolve plenamente o assunto, falta pontos de apoio relacionados ao tema. Muitas vezes não há padrão organizacional claro, com começo, meio e fim. O escritor não usa dispositivos coesivos. O escritor demonstra uma estreita faixa de vocabulário, com pouca evidência de escolha de palavras apropriadas para uso acadêmico. Os erros são frequentes, de todos os tipos, e geralmente possuem significados obscuros. Demonstra total desconhecimento da norma padrão da língua escrita.

Fonte: Autora

ANEXO A – TAREFAS PROPOSTAS PARA A AVALIAÇÃO

Para a elaboração da resposta às tarefas de avaliação, o participante foi alertado a observar rigorosamente as instruções a seguir:

INSTRUÇÕES

1. Focalize o tema proposto.
2. A resposta deve, necessariamente, referir-se ao texto de apoio ou dialogar com ele. Atenção, evite mera colagem ou reprodução.
3. Organize a resposta de modo que preencha entre 10 (mínimo) e 15 (máximo) linhas plenas, considerando-se letra de tamanho regular.
4. Use a prosa como forma de expressão.
5. Comece a desenvolver a resposta na linha 1.

Quadro 27 – Tarefa 1

TEMA 1

LEGADO ÀS FUTURAS GERAÇÕES

O mundo avança em vertiginosas transformações: ele se transforma a todo momento em nossos usos e costumes, na vida, no trabalho, nos governos, na família, nos modelos que nos são apresentados, em nossa capacidade de fazer descobertas, no progresso e na decência. Se há 100 anos a vida era mais previsível – o pai mandava e o resto da família obedecia, o professor e o médico tinham autoridade absoluta, os governantes eram nossos heróis e havia trilhas fixas a serem seguidas ou seríamos considerados desviados –, hoje ser diferente pode dar status.

Não adianta falar em ética, se vasculho bolsos e gavetas de meus filhos, se escuto atrás da porta ou na extensão do telefone – a não ser que a ameaça de drogas justifique essa atitude. Não adianta falar de justiça, se trato miseravelmente meus funcionários. Nem se deve pensar em respeito, se desrespeitamos quem nos rodeia, e isso vai dos empregados ao parceiro ou parceria, passando pelos filhos, é claro. Se sou tirana, egoísta, bruta; se sou tola, fútil, metida a gatinha gostosa; se vivo acima das minhas possibilidades e ensino isso aos meus filhos, o efeito sobre a moral deles e sua visão de vida vai ser um desastre. Nós somos aquele primeiro modelo que crianças recebem e assimilam, e isso passa pelo ar, pelos poros, pelas palavras, por silêncios e posturas.

(Adaptado de: LUFT, L. Legado aos nossos filhos. Veja. São Paulo, ed. 2082, p. 24, 15 out. 2009.)

Com base na reportagem, elabore um texto dissertativo cujo foco seja a contribuição que podemos deixar aos nossos jovens em relação a ética e cidadania.

Fonte: COPS/UEL

Quadro 28 – Tarefa 2

TEMA 2

TRABALHO INFANTIL, ONTEM E HOJE

A noção de que a infância é uma fase peculiar da vida, com necessidades, ritmo e tolerância diferentes, é uma descoberta recente. Até o século XIX, as crianças trabalhavam ao lado dos pais em indústrias e lavouras. Isso era comum mesmo em países mais desenvolvidos. A conscientização de que o lugar da criança é na escola é coisa que só chegou para os pobres dos Estados Unidos ou da Europa nos primeiros anos do século XX. No Brasil, o ensino obrigatório até os 14 anos só entrou na Constituição em 1937 e demorou décadas até alcançar as regiões mais pobres. Leis contra o trabalho infantil são ainda mais recentes. A prática passou a ser denunciada e combatida nos últimos anos depois de uma conclusão óbvia: os pais só irão recolocar os filhos na escola se isso não ameaçar o sustento da família.

(Disponível em: <www.midiaindependente.org/blue/2007/11/402823.shtml> Acesso: 3 dez. 2009.)

(KAISER, A. Ao sabor do café: fotografias de Arminio Kaiser. Organizadores Edson Vieira e Tati Costa. Londrina: Câmara Clara, 2008.)

Apesar das ligeiras quedas nos indicadores de trabalho infantil e abandono da escola, a situação de boa parte da juventude brasileira ainda é dramática. Considerando o texto e a foto, elabore um texto dissertativo apresentando o seu ponto de vista sobre o assunto.

Fonte: COPS/UEL

ANEXO B – ESTIMAÇÃO DOS PARÂMETROS PELO MÉTODO JMLE

Quadro 29 – Estimação preliminar da locação das pessoas e dos itens com o método JMLE

Passo	Equações	Descrição
*Cálculos preliminares para a locação dos itens (*logitos*)*		
I1	$p_i = \sum_{j=1}^{N} x_{ji}/N$	Média de respostas corretas ao item i ($p_i \equiv p-$valor para o item i)
I2	$\hat{b}_i = \ln[(1-p_i)/p_i]$	Estimativas preliminares dos itens
I3	$\bar{b} = \sum_{i=1}^{L} \hat{b}_i/L$	Média dos itens
I4	$\hat{b}_{i[0]} = \ln[(1-p_i)/p_i] - \bar{b}$	Estimativas centradas dos itens. Iteração=[0]
*Cálculos preliminares para a locação das pessoas (*logitos*)*		
P5	$p_j = \sum_{i=1}^{L} x_{ji}/L$	Média de respostas corretas para a pessoa j ($p_j \equiv p-$valor para a pessoa j)
P6	$\hat{\theta}_{j[0]} = \ln[p_j/(1-p_j)]$	Estimativas preliminares para as pessoas. Iteração=[0]

Adaptado de: Engelhard (2013)

Quadro 30 – Algoritmo de Newton Raphson para ajustar os parâmetros de dificuldade dos itens e da habilidade das pessoas com o método JMLE

Passo	Equações	Descrição
Iterações para ajustar as locações dos itens (I1 a I5)		
I1	$s_i = \sum_{j=1}^{N} x_{ji}$	$[s_i]$ é o número de respostas corretas (escore) para o item i e N é o número de pessoas
I2	$p_{ji[k]} = \frac{\exp(\hat{\theta}_j - \hat{b}_{i[k]})}{1+\exp(\hat{\theta}_j - \hat{b}_{i[k]})}$	p_{ji} esperado para a iteração=[k]
I3	$A_{i[k]} = \frac{\sum_{j=1}^{N-1} p_{ji[k]} - s_i}{-\sum_{j=1}^{N-1} p_{ji[k]}(1-p_{ji[k]})}$	Ajuste da dificuldade do item no passo [k]
I4	$\hat{b}_{i[k+1]} = \hat{b}_{i[k]} - A_{i[k]}$	Ajuste estimado da dificuldade do item para a iteração [k+1]
I5	$\lvert\hat{b}_{i[k+1]} - \hat{b}_{i[k]}\rvert < 0,01$	Repetir os passos I2, I3 e I4 até a diferença entre as dificuldades dos itens em valor absoluto ser pequena
I6	$\hat{b}_{i[k]} - \bar{b}_{[k]}$	Recentrar a dificuldade do item
Iterações para ajustar as habilidades das pessoas (P1 a P5)		
P1	$r_j = \sum_{i=1}^{L} x_{ji}$	$[r_j]$ é o escore bruto da pessoa j, L é o número de itens
P2	$p_{ji[k]} = \frac{\exp(\hat{\theta}_{j[k]} - \hat{b}_i)}{1+\exp(\hat{\theta}_{j[k]} - \hat{b}_i)}$	p_{ji} esperado para a iteração [k]
P3	$A_{j[k]} = \frac{r_j - \sum_{j=1}^{L-1} p_{ji[k]}}{-\sum_{j=1}^{L-1} p_{ji[k]}(1-p_{ji[k]})}$	Ajuste da habilidade da pessoa no passo [k]
P4	$\hat{\theta}_{j[k+1]} = \hat{\theta}_{j[k]} - A_{j[k]}$	Ajuste estimado da habilidade da pessoa para a iteração [k+1]
P5	$\lvert\hat{\theta}_{j[k+1]} - \hat{\theta}_{j[k]}\rvert < 0,001$	Repetir os passos P2, P3 e P4 até a diferença entre as dificuldades dos itens em valor absoluto ser pequena
Repetir os passos I1-I6 e P1-P5 até as estimativas convergirem		

Adaptado de: Engelhard (2013)