Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Matemática

Teoria da resposta ao item: o uso do modelo de Samejima como proposta de correção para itens discursivos

por

Bruno Marx de Aquino Braga

Brasília

2015

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de matemática

Teoria da resposta ao item: o uso do modelo de Samejima como proposta de correção para itens discursivos

por

Bruno Marx de Aquino Braga *

Dissertação apresentada ao Departamento de Matemáticatica da Universidade de Brasília, como parte dos requisitos do "Programa" de Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, para obtenção do grau de

MESTRE

Brasília, 15 de Julho de 2015

Comissão Examinadora:

Prof. Dr. Mauro Luiz Rabelo – MAT/UnB (Orientador)

Prof. Dr. Angel Rodolfo Baigorri – MAT/UnB

Prof. Dr. Antônio Luiz de Melo – FUP/UnB

*O autor foi bolsista da Capes durante a elaboração deste trabalho.

Dedico este trabalho aos meus pais e à minha irmã, porto seguro sempre, ao meu orientador Mauro Rabelo pelo estímulo constante e aqueles que, com paciência, estiveram ao meu lado nesse período, em especial ao amigo de primeira hora, Moacir Carvalho.

Agradecimentos

Agradeço a todas as pessoas que me apoiaram direta ou indiretamente neste período de estudos, esforço, dedicação, mas, principalmente, conquistas.

Sei que não estive sozinho e alguns, em especial, não faltaram nos momentos de dificuldades: Artidório, meu pai, Marli, minha mãe, Cristina, minha irmã, meus amigos Moacir, Lúcio, Anderson, Luiz Fernando e todos meus companheiros da turma PROFMAT 2013.

Não poderia me eximir de agradecer a todos os docentes da Universidade de Brasília que, direta ou indiretamente, apoiaram o programa PROFMAT na missão de qualificar de professores nos mais diversos rincões deste país, em especial aos professores Rui Seimetz, Lineu Neto, Ari Medino, Raquel Dorr, Carlos Alberto Santos, Aline Pinto, Lucas Ferreira, Adail Cavalheiro, Daniele Baratela e ao Prof. Dr. Mauro Rabelo, meu orientador neste projeto e incentivador.

Agradeço à *CAPES* pelo apoio financeiro a este trabalho.

"Eu não sou quem eu gostaria de ser; eu não sou quem eu poderia ser, ainda, eu não sou quem eu deveria ser. Mas graças a Deus eu não sou mais quem eu era"
Martin Luther King

Resumo

Para a avaliação educacional em larga escala é necessário um sistema de dados que forneça instrumentos de gestão da aprendizagem visando a implementação ou manutenção de políticas educacionais públicas ou privadas, visto que promove um contínuo monitoramento das estratégias adotadas, buscando detectar suas fragilidades e avanços.

Logo, um sistema de avaliação deve obter e organizar informações periódicas e comparáveis sobre os diferentes aspectos do sistema educacional.

Nesse sentido, para a avaliação educacional em larga escala, vários países utilizam-se da Teoria da Resposta ao item (TRI) que, em princípio, veio complementar algumas limitações da Teoria Clássica de Medidas.

No Brasil, a TRI foi usada pela primeira vez em 1995 na análise dos dados do Sistema Nacional de Ensino Básico (SAEB) e, entre outras avaliações em larga escala, é utilizada também no Exame Nacional do Ensino Médio (ENEM) nas provas de conhecimentos de Matemática, Ciências Humanas, Ciências da Natureza e Linguagens e Códigos, a correção da redação do ENEM é feita sob aspectos da Teoria Clássica de Medidas.

A proposta deste trabalho é de apresentar, para os modelos para itens não dicotômicos, o modelo de Samejima, o qual permite a criação de uma escala de correção da redação do ENEM a partir da TRI.

Palavras-Chave: avaliação em larga escala, TRI, Samejima, ENEM, Redação.

Abstract

In large-scale educational evaluation, there is a need for a data system whose objective is to provide instruments of learning management, aiming for the implementation or maintenance of educational policies, either public or private. This data system is necessary, since it promotes a continuous monitoring of the adopted strategies, aiming to detect its fragilities and improvements.

Therefore, an evaluation system must obtain and organize periodical and comparable information about the different aspects of the educational system.

In this sense, for large-scale educational evaluation, several countries make use of the Item Response Theory (IRT), which, in principle, has come to complement some limitations of the Classic Testing Theory.

In Brazil, IRT was used for the first time in 1995, to analyze data from the Sistema Nacional de Ensino Básico[1]– SAEB. Among other large-scale evaluations, it is also used in the Exame Nacional do Ensino Médio[2]–ENEM.

However, even though the IRT is used in the ENEM knowledge tests of Mathematics, Human Sciences, Nature Sciences and Languages and Codes, grading of the ENEM essay is made under aspects of the Classic Testing Theory.

The purpose of this work is to present, among non-dichotomous items models, the Samejima model, which permits the creation of a grading scale for the ENEM essay based on the IRT.

1: National Basic Education System

2: National High School Examination

Key-Words: Large-scale evaluluation, IRT, Samejima, ENEM, essay.

Sumário

Introdução

Um sistema de avaliação em larga escala deve obter e organizar informações periódicas e comparáveis sobre os diferentes aspectos do sistema educacional.

Nesse sentido, para a avaliação educacional em larga escala, vários países utilizam-se da Teoria da Resposta ao item (TRI) que, em princípio, veio complementar algumas limitações da Teoria Clássica de Medidas.

De acordo com Andrade, D. F. e Tavares & Valle (2000), a Teoria da Resposta ao Item (TRI) é uma metodologia que sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes.

Traços latentes são características do indivíduo que não podem ser observadas diretamente, isto é, não existe um aparelho capaz de medi-las diretamente, como um termômetro que mede diretamente a temperatura. Portanto, essas características são mensuradas por meio de variáveis secundárias que sejam relacionadas com o traço latente em estudo.

A TRI é uma poderosa ferramenta estatística que surgiu para suprir as necessidades decorrentes das limitações da Teoria Clássica da Medida (TCM) ou Teoria Clássica dos Testes (TCT), teoria que tradicionalmente era, e ainda é, utilizada nas avaliações.

No presente trabalho, serão estudados alguns modelos da TRI para itens dicotômicos e politômicos tendo por objetivos:

- Apresentar justificativas para o porquê de se avaliar.
- Traçar o histórico da avaliação em larga escala no Brasil.
- Revisar brevemente a literatura referente aos modelos de TRI hoje utilizados.
- Propor uma estratégia de correção da redação do ENEM pela TRI, via modelo de Samejima.

Capítulo

1

Avaliação em larga escala

1.1 Por que avaliar?

É notória a importância da avaliação no processo de ensino-aprendizagem, porém a prática docente nos mostra que, muitas vezes, atribui-se à avaliação o significado de uma atividade isolada do restante do processo, ainda que isso não apareça nos discursos. Isso pode ser percebido pelo uso de frases que expressam o entendimento do que seja avaliar como um momento pontual de verificação do rendimento das aprendizagens dos alunos: registrar os resultados; verificar o nível de conhecimento; verificar se o aluno atingiu ou

não os objetivos; ver se os alunos compreenderam ou não os conteúdos ensinados; medir o aproveitamento do aluno, entre outras.

Verifica-se, ainda, que a avaliação é vista como momento de diagnóstico, que irá oferecer subsídios para orientar novas possibilidades para as práticas docentes, como retomar o que não foi aprendido e perceber a dificuldade do aluno.

Quando atribuímos à avaliação somente o papel de representar um momento de balanço das aprendizagens, às vezes, até mesmo inconscientemente, constituímos um jeito de avaliar que é essencialmente somativo. Isso significa que não estamos considerando o processo de ensino-aprendizagem efetivamente como um processo contínuo e ininterrupto, mas apenas como um momento do processo, separado dos demais.

Assim, estamos investindo esforços exclusivamente em uma perspectiva em que o resultado da aprendizagem é constatado somente no final do período de formação, reduzindo consideravelmente as possibilidades de uma ação mais efetiva, que favoreça uma regulação contínua durante todo o processo. Centralizar a expressão do aproveitamento das aprendizagens na sua condição final estreita o entendimento de processo, reduz a busca por informações visando à melhoria da qualidade do ensino e limita a ideia de desempenho dos alunos, classificando-os em: aprovado; reprovado; em recuperação; está progredindo, está se esforçando etc.

Por outro lado, quando atribuímos à avaliação apenas o significado de diagnóstica, estamos considerando que o levantamento de informações ocorre em determinado momento do processo de ensino-aprendizagem. Pensar e agir somente desse modo também empobrece o processo de regulação. Se quisermos privilegiar a atuação e o desempenho do aluno, a regulação deve ser contínua (e não pontual), e essa continuidade sistemática acontece por meio das informações que levantamos.

Desse modo, para a aprendizagem ser dinâmica e significativa, é preciso considerar como parâmetros do diagnóstico as perguntas feitas pelos alunos, as atividades desenvolvidas em sala, as correções de tarefas, os trabalhos em equipe e outros trabalhos desenvolvidos dentro e fora da sala de aula.

Sem essa visão sistêmica, como é possível estabelecer uma regulação contínua com informações recolhidas somente em momentos pontuais? Nesse sentido, pode-se afirmar que avaliar é:

- Uma prática diagnóstica, embasada em múltiplas informações coletadas durante todo o processo, para favorecer uma regulação contínua e um julgamento mais adequado das produções dos alunos, promovendo, assim, uma mediação que esteja de acordo com as necessidades de cada um.

- Uma prática prognóstica, pois visa melhorar as condições das aprendizagens por

meio do amparo da avaliação.

- Uma prática somativa, em certo sentido, uma vez que avaliamos depois da ação de formação, em final de processo. Porém, o registro dos resultados das aprendizagens possibilita a análise e o julgamento para tomada de decisão em relação à continuidade ou não, ou à melhoria das práticas adotadas.

- Uma prática formativa, porque proporciona informações para promover o aperfeiçoamento da qualidade do projeto pedagógico desenvolvido. A análise e o julgamento, nesse caso, têm o objetivo de nortear as decisões de aprimoramento do processo de ensino-aprendizagem e dos diferentes elementos que o compõem. As informações sobre os avanços e as dificuldades nas aprendizagens interessam tanto aos alunos quanto aos professores.

- Um instrumento valioso, uma vez que possibilita colocar o "erro" em discussão para potencializar as aprendizagens . (OLIVEIRA 2012; SANDI & CHIQUITO 2009)

1.2 Avaliação em larga escala

A avaliação é apresentada também como uma política de Estado, haja vista que assim é instituída na constituição de 1988 após a promulgação da segunda LDB, Lei 9394/1996.

No artigo 9° desta Lei, a ênfase esteve no aprimoramento e sistematização das informações coletadas pelo Sistema de Avaliação da Educação Básica (SAEB), desde o início dos anos 1990, tendo em vista sua capacidade de orientar as autoridades governamentais na elaboração de políticas públicas.

> Art. 9º A União incumbir-se-á de: (Regulamento)
> V - coletar, analisar e disseminar informações sobre a educação;
> VI - assegurar processo nacional de avaliação do rendimento escolar no ensino fundamental, médio e superior, em colaboração com os sistemas de ensino, objetivando a definição de prioridades e a melhoria da qualidade do ensino;
> § 2° Para o cumprimento do disposto nos incisos V a IX, a União terá acesso a todos os dados e informações necessários de todos os estabelecimentos e órgãos educacionais. (LDB,1996)

Logo, o governo federal estabelece regras e diretrizes do sistema de ensino, por meio da definição de sua estrutura, organização, funcionamento e currículo. Essa regulação é também viabilizada pelos dados produzidos nas avaliações nacionais.

As atividades avaliativas, no entendimento de Kells (1999), inseridas na regulação estatal, são definidas como:

> [...] o processo informado e periódico através do qual um sistema, uma instituição, um programa ou um procedimento é, com o passar do tempo, colocado em sintonia com suas expectativas (intenções, padrões, normas), através de escolhas e ações julgadas necessárias pelo(s) regulador(es), tendo como base o resultado de uma avaliação formativa ou somativa.

Os dois papéis assumidos pelas avaliações nacionais se prestam a, no caso da avaliação formativa, oferecer informações úteis para a melhoria de um programa ou projeto no decorrer de suas diferentes fases. Na avaliação somativa, disponibilizar resultados aos responsáveis pelo processo decisório e demais interessados, como pais, professores, alunos, de maneira que possam julgar o mérito dos programas ou projetos desenvolvidos.

A avaliação educacional, então, possibilita conhecer os efeitos das ações realizadas, para acompanhar os resultados, subsidiar decisões a respeito da continuidade, encerramento ou ampliação dos trabalhos empreendidos.

Já a avaliação em larga escala, assim denominada devido à grande quantidade de pessoas envolvidas em cada uma das etapas, tem por especificidades a:

- Elaboração do projeto de avaliação.
- Construção de instrumentos padronizados (testes e questionários).
- Validação estatística dos instrumentos.
- Constituição e treinamento das equipes de trabalho.
- Execução e monitoramento simultâneos da avaliação em diferentes instituições pelo território nacional.
- Disseminação e processamento de resultados e dados.
- Repercussão dos resultados na sociedade.

Discutiremos neste trabalho a construção de instrumentos padronizados, notadamente com a adoção de modelos da Teoria de Resposta ao Item.

De acordo com Andrade, D. F. e Tavares & Valle (2000), a TRI é uma metodologia que sugere formas de representar a probabilidade de um indivíduo dar uma certa resposta a um item em função de seus traços latentes.

A TRI é uma poderosa ferramenta estatística que surgiu para suprir as necessidades decorrentes das limitações da Teoria Clássica da Medida (TCM) ou Teoria Clássica do Teste (TCT), que tradicionalmente era, e ainda é, utilizada nas avaliações (Andrade, D. F. e Tavares & Valle 2000).

A TRI categoriza as questões em fáceis e difíceis. Por isso consegue apenar candidatos que marcam respostas ao acaso e valorizar os que apresentam coerência no padrão de respostas. Dessa forma, aqueles que acertam mais questões difíceis do que fáceis têm a nota diminuída, uma vez que, na lógica da TRI, deveriam acertar as mais fáceis.

Assim, é possível que candidatos com mais itens acertados possam ter uma nota inferior a outros candidatos com menos itens acertados.

Entretanto, Andrade, D. F. e Tavares & Valle (2000) consideram que o maior avanço da TRI foi devido ao fato de permitir a comparação de indivíduos que fizeram exames diferentes, cujos itens encontram-se na mesma escala, e a criação de escalas interpretáveis, que são de grande importância para dar *feedback* para os participantes.

1.3 História da TRI no Brasil

Embora a TRI já tenha uma longa história (PASQUALI, 1996), as primeiras aplicações na área educacional no Brasil começaram em 1995, segundo Andrade, D. F. e Tavares & Valle (2000), através da pesquisa AVEJU, da Secretaria de Estado da Educação de São Paulo, e continuaram no Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP) e no (SAEB) do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Entretanto Gatti (1996) considera que a primeira aplicação da TRI foi realizada em 1993 pela Secretaria de Educação do Estado de São Paulo. No entanto, ela não foi aplicada com toda a sua potencialidade. Buscava-se uma metodologia mais sofisticada e precisa que permitisse a construção de escalas de habilidade a fim de acompanhar o progresso do conhecimento adquirido ao longo do tempo (Andrade, D. F. e Tavares & Valle 2000).

Nessas aplicações, a TRI tem mostrado a sua potencialidade no que diz respeito à avaliação educacional, por meio da construção de uma escala comparável, permitindo o acompanhamento do progresso do conhecimento adquirido pelo aluno ao longo do tempo, como tem sido feito nos países desenvolvidos. A partir de então, cada vez mais institutos de educação têm aderido a TRI para as suas avaliações educacionais, por exemplo, o

Sistema Mineiro de Avaliação da Educação Pública–SIMAVE (SOARES, GENOVEZ e GALVÃO, 2005) da Secretaria de Estado de Educação de Minas Gerais e no Projeto GERES (PERRY, 2009) e, mais recentemente, o ENEM (FERREIRA, F. F. G., 2009).

Capítulo

2

Revisão de literatura

2.1 Introdução à Teoria da Resposta ao Item na Avaliação

É comum verificar-se que, em processos avaliativos cuja finalidade é a seleção de candidatos ou verificação de aprendizagem são utilizados resultados obtidos em provas (instrumentos avaliativos de desempenho), expressos apenas por seus escores brutos ou padronizados. Isso significa dizer que, por exemplo, quanto maior a nota do respondente em uma prova, melhor sua classificação, fato que, por se tratar de instrumentos avaliativos de medição de desempenho, não reflete o todo da referida prova, tão pouco o grau/índice de conhecimento do respondente sobre os temas investigados.

Estatisticamente, esse procedimento se caracteriza em análises e interpretações, sempre associadas ao grau obtido pelo examinado e não a um item (questão) em particular. Em termos específicos, as análises e interpretações estão sempre associadas à prova como um todo; pressuposto característico da TCT, conforme descrito por Vianna: "contudo, uma característica dessa teoria consiste no fato de a análise psicométrica do construto ter por ênfase o instrumento de medição (prova) como um todo e não o item" (Vianna,1973; Gulliksen,1967). Disto decorre a impossibilidade de comparação entre respondentes que não foram submetidos às mesmas provas ou, pelo menos, ao que se denomina de formas paralelas de testes.

Assim procedendo, buscando-se medidas avaliativas em instrumentos de medição de desempenho, muitas questões em Educação, cujo teor avaliativo investiga aquisição qualitativa de conhecimentos, permanecem sem respostas.

Pleiteando um significativo avanço, em termos estatísticos, vislumbra-se a TRI como

uma metodologia de melhor validação de análises das respostas, haja vista a já citada questão da investigação qualitativa de conhecimentos, na qual se parte do pressuposto de que os itens constitutivos do instrumento avaliativo assumem a característica de "elementos centrais da análise", quebrando o paradigma observado em análises da TCT, cuja centralização faz referência à prova como um todo (VALLE, 1999).

Assumindo este pressuposto de validação aos itens como elementos centrais da análise, pode-se, por exemplo, comparar populações (grupos de respondentes) submetidas a provas diferentes, considerando instrumentos investigativos que venham a abranger os mesmos temas, ou seja, que ocorram comunalidades em suas características estruturantes em termos do conteúdo a ser avaliado.

Segundo Andrade (2000), outra comparação possível com a utilização da TRI é aquela entre respondentes de mesmo grupo, em provas totalmente diferentes. Em termos práticos, suponha comparar os níveis de conhecimento entre alunos de duas séries distintas (1° e 3° anos do ensino médio, por exemplo). Na TCT, essa comparação somente será possível caso seja aplicada a mesma prova para as duas turmas (dois grupos de respondentes). É também possível avaliar o desenvolvimento de determinada série de um ano para outro, ou, ainda, comparar o desempenho entre escolas públicas e privadas. Assim, várias questões de interesse prático na área da educação podem ser respondidas com a utilização da TRI, trazendo um pouco mais de informação.

Mesmo os processos avaliativos educacionais cuja finalidade principal é obter resultados classificatórios podem vir a ser, por meio da TRI, melhor compreendidos em termos qualitativos, do que, como o são tradicionalmente contextualizados, em termos quantitativos, até porque o reflexo de tais análises deve ser entendido como a estimação de parâmetros de investigação. E não pura e simplesmente como uma escala, como se pudéssemos fragmentar o respondente em partes iguais de zero a dez, por exemplo. A TRI constitui-se, portanto, em um instrumento poderoso de análise e interpretação, o qual, ao propor modelos para os traços latentes, realiza observações de variáveis secundárias, relacionadas a estes.

A TRI propõe modelos de variáveis latentes para representar a relação entre a probabilidade de um respondente apresentar determinada resposta a um item e seus traços latentes ou proficiências na área do conhecimento avaliada, para permitir, inclusive, a construção de escalas de proficiências calibradas, ou seja, permite analisar as interações entre os respondentes e os itens.

> "... talvez o aspecto mais importante da nova teoria é a promessa de fornecer medidas invariantes do desempenho cognitivo, que não dependem dos itens que compõem a prova ou das pessoas investigadas na amostra...". A calibração fornece, a cada item, parâmetros que caracterizam suas qualidades técnicas, independentes da população investigada. (...) Sendo invariantes, eles não dependem da amostra selecionada para fins de calibração. Sendo invariantes, podem ser aplicados a qualquer outra população, proporcionando resultados na mesma escala de proficiência" (Fletcher 1994).

A interpretação qualitativa sobre instrumentos de avaliação quantitativa tem se tornado cada vez mais necessária no contexto educacional, principalmente em termos de Brasil, onde este tipo de abordagem de validações qualitativas sobre instrumentos tradicionalmente quantitativos há pouco vem sendo implantada. Deste modo, a TRI avança sobremaneira na interpretação estatística, pois propicia novas condições de análises em direção à consolidação do conceito de qualidade em se tratando do processo de ensino e aprendizagem.

Pode-se entender por traço latente ou competências cognitivas as diferentes modalidades estruturais da inteligência que compreendem determinadas operações que o indivíduo utiliza para estabelecer relações com e entre os objetos físicos, conceitos, situações, fenômenos e pessoas. As habilidades instrumentais referem-se especificamente ao plano do saber fazer e decorrem, diretamente, do nível estrutural das competências já adquiridas e que se transformam em habilidades, isto é, a "capacidade de agir eficazmente em um determinado tipo de situação, apoiando-se em conhecimentos, mas sem se limitar a eles" (Perrenoud, 1999).

2.1.1 Considerações Gerais Sobre a Teoria da Resposta ao Item

Os instrumentos de avaliação de desempenho passam a ter, então, nos itens (questões, perguntas) a função de "elementos centrais" e, destes, como resultado agregados, a interpretação da prova / teste como um todo, assegurando uma validação qualitativa em tal perspectiva constitutivamente quantitativa.

Para tanto, a TRI tem utilizado duas funções matemáticas para caracterizar os parâmetros métricos dos itens componentes de um teste: a função logística e a função distribuição da normal padronizada (Muñiz & Hambleton, 1992) também conhecida como ogiva Gaussiana. Ambas variam de 0 a 1 e nessa escala situa-se a probabilidade de um examinado acertar a um item específico. De modo que os modelos usados pela TRI procuram se

adequar a essas funções. Cada item tem a sua Curva Característica de Informação (CCI) que segue um modelo baseado em uma daquelas funções. As CCIs descrevem os resultados para um item em termos das avaliações dos parâmetros dos itens e, evidentemente, através das suas formas, as quais serão analisadas mais especificamente na descrição dos Modelos Matemáticos da Teoria da Resposta ao Item.

As informações contidas nas CCIs a respeito dos parâmetros métricos dos itens dependem do modelo teórico escolhido. Rasch, em 1960, propõe o modelo denominado "Modelo Logístico de um Parâmetro", o qual será descrito detalhadamente. Este modelo contém o pressuposto de que a probabilidade de acerto de um item é influenciada apenas pelo grau de dificuldade do item. O parâmetro grau de dificuldade costuma ser representado por b.

Um segundo modelo, denominado Modelo Logístico de dois Parâmetros, foi formulado por A. Birnbaum em 1968. Neste modelo, a probabilidade de acerto de um item é influenciada pelo grau de dificuldade b e pelo grau de discriminação a.

O terceiro modelo desenvolvido, denominado Modelo Logístico de Três Parâmetros, foi construído a partir dos trabalhos de A. Birnbaum e assume que a probabilidade de acerto de um item é influenciada pela sua dificuldade, discriminação e probabilidade de acerto ao acaso. Consequentemente têm-se três parâmetros: a, b e c, sendo c a probabilidade de acerto ao acaso.

Os primeiros modelos estatísticos da TRI datam da década de 50 e eles foram primeiramente desenvolvidos na forma da função distribuição da normal, ou seja,

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt.$$

Depois, foram descritos para uma forma matemática mais fácil de ser tratada, sendo usada até hoje. Essa forma é a da função logística:

$$P_i(\theta) = \frac{e^{D^{(\theta - b_i)}}}{1 + e^{D^{(\theta - b_i)}}}$$

que é computacionalmente mais conveniente, pois é uma função explícita dos parâmetros do item e da proficiência e não envolve cálculos mais complexos.

Lord, em 1952, desenvolveu o modelo unidimensional de dois parâmetros, baseado na distribuição normal acumulada (ogiva normal):

$$P_i(\theta) = \frac{e^{Da_i{}^{(\theta - b_i)}}}{1 + e^{D_i{}^{(\theta - b_i)}}}$$

Este modelo foi aplicado com as limitações computacionais da época e, após algumas aplicações, sentiu-se a necessidade da incorporação de um parâmetro que tratasse do

problema do acerto casual. Então, no decorrer dos estudos, surgiu o modelo de três parâmetros

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i{}^{(\theta_j - b_i)}}}$$

No ano de 1977, Bock & Zimowski apresentaram os modelos logísticos de l, 2 e 3 parâmetros para duas ou mais populações de respondentes. A introdução desses modelos trouxe novas possibilidades para as comparações de rendimentos de duas ou mais populações submetidas a diferentes testes com itens comuns.

Uma das dificuldades da TRI é a estimação dos parâmetros envolvidos nos modelos, em particular quando se necessita estimar tanto os parâmetros dos itens quanto as proficiências (θ). No começo, a estimação era feita por meio do método da máxima verossimilhança conjunta que envolve um número muito grande de parâmetros a serem estimados simultaneamente e, consequentemente, grandes problemas computacionais.

Em 1970, Bock & Lieberman introduziram o método da máxima verossimilhança marginal para a estimação dos parâmetros em duas etapas. Na primeira etapa, estimam os parâmetros dos itens, assumindo-se uma certa distribuição para as proficiências. Na segunda etapa, assumindo os parâmetros dos itens conhecidos, estimam-se as proficiências. Apesar do avanço que esse método trouxe para o problema, ele requeria que todos os parâmetros dos itens fossem estimados simultaneamente. Em 1981, Bock & Aitkin propuseram uma modificação no método acima, utilizando o algoritmo EM de Dempster, Laird & Rubin (1977), de modo a permitir que os itens pudessem ter seus parâmetros estimados em separado, facilitando em muito o aspecto computacional do processo de estimação. Mais recentemente, métodos bayesianos foram propostos para, entre outros aspectos, resolver o problema de estimação dos parâmetros dos itens respondidos corretamente ou incorretamente por todos os respondentes e, também o problema da estimação das proficiências dos respondentes que acertaram ou erraram todos os itens da prova (Andrade 2000).

Atualmente, a TRI vem tornando-se a técnica predominante no campo de testes em vários países e, aqui no Brasil, a TRI foi usada pela primeira vez em 1995 na análise dos dados do SAEB. A introdução da TRI permitiu que os desempenhos de alunos de 5° e 9° anos do Ensino Fundamental e de 3^{a} série do Ensino Fundamental pudessem ser comparadas e colocados em uma escala única de conhecimento.

2.2 Modelos Matemáticos da Teoria da Resposta ao Item

A TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um j dar uma resposta certa a um item i de um instrumento de avaliação em função dos parâmetros dos itens e do conhecimento (ou proficiências) θ_j do respondente. Assume-se que, quanto maior a proficiência (habilidade), maior a probabilidade de acerto do item. Segundo Valle (1999), os modelos propostos dependem fundamentalmente de três fatores:

1. da natureza do item – dicotômicos ou não dicotômicos;
2. do número de populações envolvidas – apenas uma ou mais de uma;
3. do número de traços latentes que estão sendo medidos – apenas um ou mais de um.

2.2.1 Formulação do Modelo de Rasch

Trata-se de um modelo dicotômico pensado na sua forma mais simples. Prediz, por exemplo, a probabilidade condicional de um resultado binário (correto/incorreto, acerto/não acerto), dada a competência do respondente e a dificuldade da questão.

A codificação adotada pelo matemático suíço Rasch foi de "1" para resposta correta e "0" para a resposta incorreta. O modelo, então, expressa a probabilidade de se observar uma resposta correta, ou seja, de se observar "1" ao invés de "0", como uma função da diferença entre a competência/proficiência (θ) da pessoa e a dificuldade (b) da questão.

Assim, tem-se a diferença ($\theta - b$), que é a relação fundamental no modelo e que funciona como expoente na função logística. O modelo de Rasch é, portanto, uma expressão matemática para a relação entre a probabilidade de sucesso (P) e a diferença entre a habilidade/competência do examinado (θ) e a dificuldade de um item b.

Algebricamente, tem-se, com a função logística, a probabilidade:

$$P = \frac{\exp(\theta - b)}{1 + \exp(\theta - b)}$$

e, desenvolvendo-se, temos:

$$P + (1 + \exp(\theta - b)) = \exp(\theta - b)$$

$$P + P\exp(\theta - b) = \exp(\theta - b)$$

$$P = \exp(\theta - b) - P\exp(\theta - b)$$

$$P = \exp(\theta - b).(1 - P)$$

$$\left(\frac{P}{1-P}\right) = \exp(\theta - b)$$

$$\log\left(\frac{P}{1-P}\right) = \log(\exp(\theta - b)$$

aplicando a propriedade dos logaritmos:

$$\log\left(\frac{P}{1-P}\right) = \theta - b$$

Pode-se afirmar, agora, segundo Rasch, que, quando a habilidade/competência é igual à dificuldade ($\theta = b$), o resultado da operação de subtrair a dificuldade b da habilidade (θ) é zero ($\theta - b = 0$), e ao atuar como expoente do número neperiano e faz com que o resultado seja um, $e^0 = 1$ (todo número elevado a zero é igual à unidade).

Então, sendo a habilidade igual à dificuldade ($\theta - b = 0$), a probabilidade aplicada em:

$$P = \frac{\exp(\theta-b)}{1+\exp(\theta-b)}$$

$$P = \frac{\exp(\theta-\theta)}{1+\exp(\theta-\theta)}$$

$$P = \frac{1}{(1+1)}, \text{ou}$$

$$P = \frac{1}{2} = 0,50\text{x}100 = 50\%$$

Logo, a probabilidade de acertar um item quando $\theta = b$ é de 50%.

A unidade de medida usada por Rasch para calibrar itens (estimar dificuldade) e medir a habilidade dos respondentes (estimar competência) passou a ser chamada de *logit* por causa da transformação logarítmica (*log odds*) da probabilidade de uma resposta correta, sempre variando com valores positivos e negativos em torno do zero arbitrário da escala.

Wright & Stone (1979, p. 17) afirmam que a competência de uma pessoa em *logit* é o logaritmo natural da sua chance *(is the natural log odds)* de acertar questões do

tipo escolhido para definir o ponto "zero" da escala. Por outro lado, a dificuldade de um item em *logit* é o logaritmo natural da sua chance de induzir o não-acerto em pessoas de competência "zero" *(is the natural log odds for eliciting failure from persons with "zero" ability)*, (Ziviani, 2002).

2.2.2 Propriedades Específicas Utilizadas no Modelo de Rasch

Uma suposição importante adotada nos modelos da Teoria de Resposta ao Item é a de que a resposta a uma questão não deve influenciar a resposta de outras (suposição da independência local) ou, ainda, "(...) mantidas constantes as habilidades/proficiências, as respostas dos examinados a quaisquer dois itens são estatisticamente independentes" (PASQUALI, 1996).

Correlações entre os itens seriam explicáveis somente pelo que se quer estimar nos respondentes, a competência/proficiência, ou aptidão, ou capacidade, ou traço latente frequentemente denotada na literatura da TRI por meio da letra grega θ (correspondendo, em inglês, a *latent trait, ability, proficiency*), (ZIVIANI, 2002).

A possibilidade da rápida distinção entre respostas intuitivamente prováveis ou verossímeis e respostas improváveis ou inverossímeis facilita o entendimento da essência do modelo Rasch.

Intuitivamente, vê-se que a cada resposta de cada examinado pode-se fazer corresponder uma probabilidade de acerto. Pois é isso que o modelo Rasch faz, atribui uma probabilidade de acerto à resposta para determinada questão dependendo apenas de dois parâmetros a serem estimados, a proficiência θ_j do respondente j e a dificuldade do item i, b_i.

A partir dessa ordenação de respondentes (pela suposta competência/proficiência) e de itens (pela suposta dificuldade), Rasch (1960) desenvolveu um modelo matemático com a função logística para a construção de medidas baseadas na relação probabilística entre a competência da existência de apenas uma aptidão responsável pela realização de um conjunto de tarefas ou itens.

Sobre este ponto, Muñiz (1997) observa que a unidimensionalidade perfeita "aparece como uma idealização matemática difícil, senão impossível, de se alcançar com dados psicológicos reais" e que a avaliação psicológica (traço latente do respondente) terá que se acostumar a conviver com uma unidimensionalidade imperfeita.

O Modelo de Rasch, como um modelo da TRI, centra-se na estimação conjunta da dificuldade (b_i) dos n itens i, $i = 1, 2, ..., n$ e das proficiências dos j examinados (θ_j) em uma mesma escala.

Estes parâmetros, θ_j e b_i, não se definem pela pontuação observada ou pelo número

de examinados que acertam o item como na TCT, mas, sim, pela avaliação do chamado traço latente (proficiência) e pela dificuldade do item de uma forma mais geral de um domínio, da qual um teste particular seria um indicador.

Esta medição conjunta, na mesma escala, das proficiências dos examinados (θ) e da dificuldade do item b, promove uma sensível vantagem da TRI sobre a TCT porque permite analisar as interações entre os respondentes e os itens, apresentando um diagnóstico referido à variável, identificando o tipo de situação em que um respondente (avaliando) teria alta ou baixa probabilidade de responder acertadamente.

Também traz outras vantagens como a independência de seus resultados em relação às condições com que foram obtidos (tipos de amostras ou itens). As diferenças iguais de desempenho entre os respondentes e de graus de dificuldade entre os itens têm o mesmo significado independentemente do ponto da escala em que se encontram (propriedades de intervalo) e os erros de medida, encontrados em qualquer processo de medição, são quantificados mais precisamente, permitindo observar-se para qual faixa de proficiência um teste é mais preciso.

Este cometário será melhor detalhado adiante na análise da figura 2.1 para o Modelo Logístico de um Parâmetro (ML1). O Modelo de Rasch considera que a probabilidade de resposta correta ao item i (P_i) depende somente da diferença entre o nível de proficiência (conhecimento) do respondente ou examinado (θ_j) e a dificuldade do item (b_i), sendo a formulação conhecida como: Modelo Logístico de Um Parâmetro.

2.2.3 Modelo Logístico de um Parâmetro (ML1).

Este modelo é definido pela expressão:

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

em que:

- $P_i(\theta)$: chamada de função resposta do item que é a probabilidade de um respondente escolhido ao acaso e com proficiência θ acertar o item;
- θ :nível de proficiência (conhecimento) do respondente;
- b_i: é o parâmetro que representa a dificuldade (ou de posição) do item i, medido na mesma escala da habilidade/proficiência;
- e: base dos logaritmos neperianos;

- D: é um fator de escala usado para aproximar a função logística da ogiva Gaussiana com valor 1,7.

A expressão

$$P_i(\theta) = \frac{e^{D^{(\theta - b_i)}}}{1 + e^{D^{(\theta - b_i)}}}$$

é desenvolvida para a sua forma mais simples de apresentação, dividindo-se o numerador e o denominador por $e^{D^{(\theta - b_i)}}$, obtém-se:

$$P_i(\theta) = \frac{\frac{e^{D^{(\theta - b_i)}}}{e^{D^{(\theta - b_i)}}}}{\frac{1 + e^{D^{(\theta - b_i)}}}{e^{D^{(\theta - b_i)}}}} = \frac{1}{\frac{1}{e^{D^{(\theta - b_i)}}} + 1}$$

o que fica:

$$P_i(\theta) = \frac{1}{1 + e^{-D^{(\theta - b_i)}}}$$

Agora, passando o denominador para o numerador, o expoente do numerador tornar-se-á negativo, obtendo-se:

$$P_i(\theta) = \left(1 + e^{-D^{(\theta - b_i)}}\right)^{-1}$$

que é a descrição do (ML1) na sua forma mais reduzida possível para os itens, $i = 1, 2, \ldots, n$.

Essa expressão é representada graficamente na figura adiante. Esse gráfico é denominado Curva Característica do Item (CCI). A CCI representa a probabilidade de uma resposta correta (ordenada) para cada nível do construto medido (abscissa).

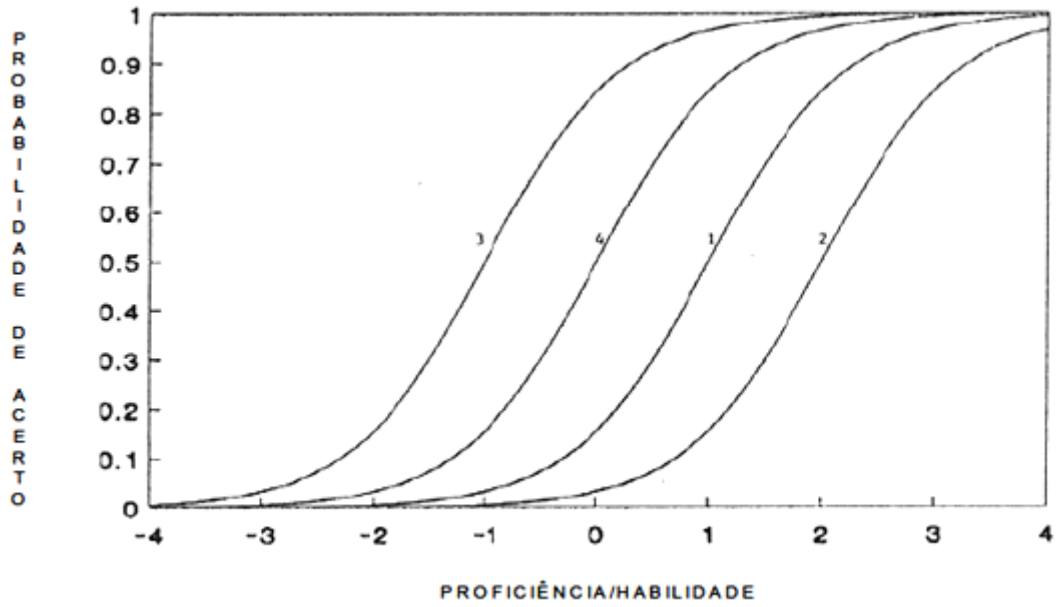

Figura 2.1: Curvas Características de 4 Itens Típicos do ML1.

Pode-se observar na figura 2.1 que, quando a probabilidade de resolver o item 2 (último a direita) é fixada em 0,5 (valor do eixo vertical da figura), tem-se em correspondência uma proficiência/habilidade $\theta = 2$ (valor do eixo horizontal da figura). Mas, se o item considerado é o 1 (penúltimo a direita), a proficiência/habilidade necessária ao acerto diminui para $\theta = 1$, quando se considera a mesma chance de 0,5. Assim, deslocando os itens para a esquerda a proficiência/habilidade necessária diminui até $\theta = -1$. Geralmente, este valor 0,5 é identificado como o grau de dificuldade limitativo b, de modo que, no gráfico, $b_3 < b_4 < b_1 < b_2$, sendo o item 3 o mais fácil; e o item 2, o mais difícil.

Esses valores da abscissa podem ser expressos em diferentes métricas (EMBRETSON; REISE, 2000). A mais utilizada é a *logit* $(\theta - b_i)$, em que a diferença de uma unidade significa que o quociente entre a probabilidade de acerto e erro é igual ao número de Neper, pois

$$\ln\left(\frac{P}{1-P}\right) = \theta - b$$

$$\ln\left(\frac{P}{1-P}\right) = 1$$

$$e^1 = \frac{P}{1-P}$$

Esta interpretação é a mesma em toda a escala, ou seja, ela possui propriedades de intervalo. A localização do ponto zero é arbitrária. Normalmente, no modelo de Rasch, usa-se situá-lo na média das dificuldades dos itens.

O objetivo, ao se aplicar um teste, é estimar tanto a proficiência dos examinados (θ) como a dificuldade dos itens b. O procedimento mais usual é determinar as estimativas que tornam mais prováveis as respostas observadas.

Após a estimação de θ e b, deve-se comprovar o ajuste do modelo aos dados empíricos, análise que pode identificar itens e/ou respondentes que não se ajustam ao modelo. A ausência deste ajuste pode significar, por exemplo, itens impróprios por não serem unidimensionais ou por estarem mal formulados e também pode significar falta de conhecimento dos respondentes. Cabe ao pesquisador eliminar dos resultados finais tanto os respondentes quanto os itens que apresentam desajuste e decidir por outro modelo de análise (HAMBLETON, 1991).

2.2.4 Modelo Logístico de dois Parâmetros (ML2).

Este foi o primeiro modelo para TRI, Modelo Matemático Unidimensional com dois parâmetros. Foi criado por Lord em 1952, baseado primeiramente na função distribuição Gaussiana padronizada. A expressão desse modelo inicial é:

$$P(U_i|\theta) = \int_{-\infty}^{a_i(\theta - b_j)} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} dz,$$

Birnbaum mudou o suporte do modelo para a função logística. E, assim, a equação utilizada para avaliar a probabilidade de um examinado j com proficiência θ_j responder corretamente ao i-ésimo item de um teste é dada por (Hambleton et al, 1991).

$$P\left(U_i|\theta_j\right) = \frac{e^{Da_i{}^{(\theta_j - b_i)}}}{1 + e^{Da^{(\theta_j - b_i)}}}$$

utilizando-se dos mesmos passos efetuados no ML1, para simplificação, obtém-se:

$$P\left(U_i|\theta_j\right) = \frac{1}{1+e^{-Da_i{}^{(\theta_j - b_i)}}}$$

$$P\left(U_i|\theta_j\right) = \left[1 + e^{-Da_i{}^{(\theta_j - b_i)}}\right]^{-1}$$

que é a sua forma mais reduzida, em que:

- $P\left(U_i|\theta_j\right)$: é a Função Característica do Item;
- a_i: é o parâmetro que representa o poder de discriminação do item i, com valor proporcional à inclinação da tangente à curva no ponto com abscissa b_i;

- b_i: é o grau de dificuldade do item;
- θ_j:é o nível de proficiência (conhecimento) do respondente j;
- e: base dos logaritmos neperianos;
- D: é um fator de escala usado para aproximar a função logística da ogiva Gaussiana com valor 1,7.

A figura 2.2 mostra as CCIs do ajuste do modelo logístico para quatro itens distintos.

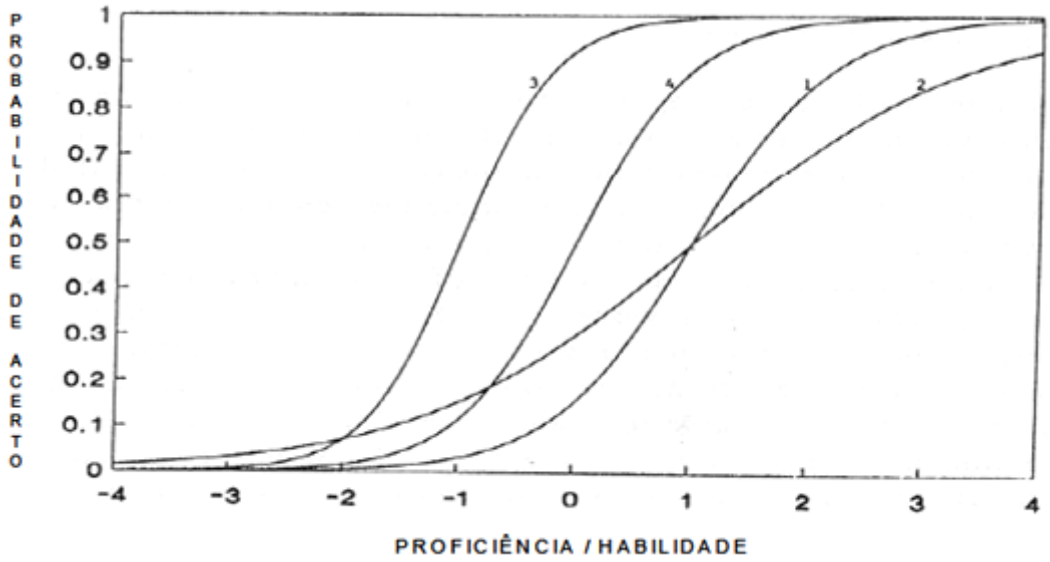

Figura 2.2: Curvas Características de 4 Itens Típicos do ML2 .

O modelo logístico de dois parâmetros (ML2) é obviamente o ML1 acrescido do parâmetro índice de discriminação, permitindo, então, a discriminação dos itens.

Para o item 1: $b_1 = 1$, para o item 2: $b_2 = 1$, para o item 3: $b_3 = -1$, para o item 4: $b_4 = 0$. As CCIs não são paralelas como elas eram anteriormente no (ML1). Assim cada CCI deste modelo tem uma inclinação diferente.

Tanto no modelo logístico de dois parâmetros como no modelo logístico de um parâmetro não é considerado que o examinando acerte o item por acaso. Esta possibilidade implica um novo parâmetro a ser incorporado no modelo.

Sabe-se que, em exames de múltipla escolha, é possível o examinando responder acertadamente um item sem ter conhecimento do assunto. Assim, pensou-se no modelo com mais um parâmetro e surgiu o ML3 (HAMBLETON et al., 1991).

2.2.5 Modelo Logístico de três Parâmetros (ML3).

Este modelo é denominado Modelo Logístico de Três Parâmetros, e é definido por:

$$P\left(U_{ij}=1/\theta_j\right)=c_i+\frac{1-c_i}{1+e^{-Da_i(\theta_j-b_i)}}$$

$$P\left(U_{ij}=1/\theta_j\right)=c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta_j-b_i)}}$$

$$P\left(U_{ij}=1/\theta_j\right)=c_i+(1-c_i)\left(1+e^{-Da_i(\theta_j-b_i)}\right)^{-1}$$

onde se chega à sua forma reduzida, com *i = 1, 2, 3, ...,I* e *j = 1, 2, 3, ..., n*, em que:

- $P\left(U_{ij}|1/\theta_j\right)$: é a probabilidade do j-ésimo indivíduo escolhido ao acaso com grau de proficiência θ_j responder corretamente ao i-ésimo item;
- U_{ij}: variável dicotômica que assume o valor 1 (um), quando o j-ésimo indivíduo responde corretamente ao item i, e assume 0 (zero) quando o j-ésimo respondente não responde acertadamente ao item i;
- a_i: é o parâmetro correspondente ao índice de discriminação;
- b_i: é o parâmetro correspondente ao grau de dificuldade do item;
- c_i: é o parâmetro que representa a probabilidade de acerto ao acaso;
- θ_j: representa o grau de proficiência (traço latente) do j-ésimo respondente;
- e: base dos logaritmos neperianos;
- D: é um fator de escala usado para aproximar a função logística da ogiva Gaussiana com valor 1,7.

A figura 2.3 mostra as CCI de seis ajustes do ML3 respectivamente a seis itens distintos.

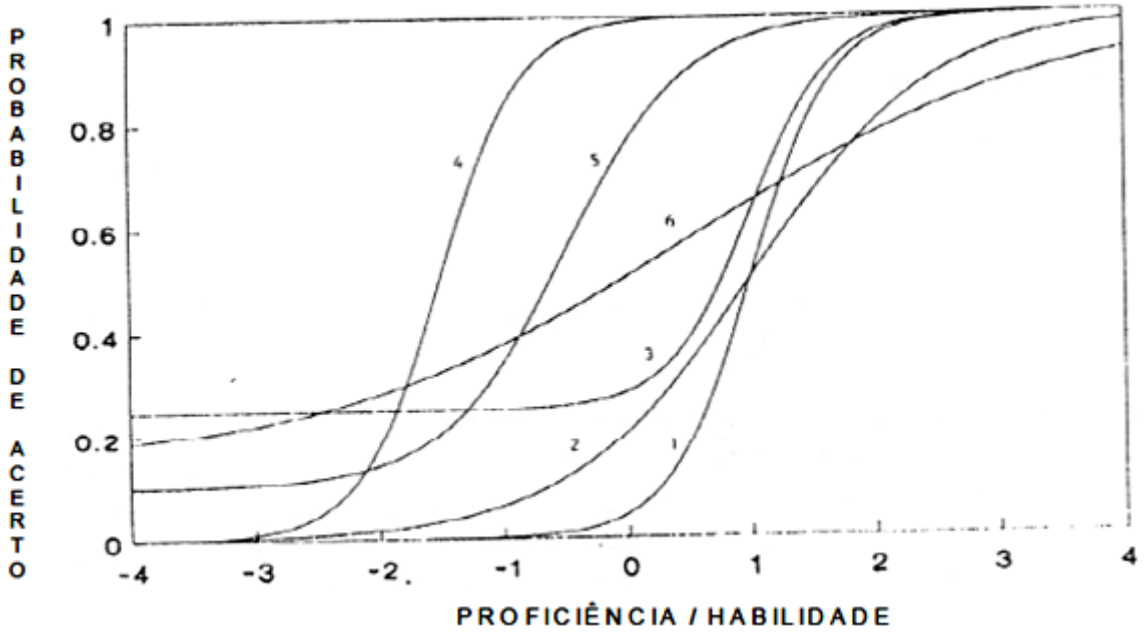

Figura 2.3: Curvas Características de 6 Itens Típicos do ML3.

Da figura 2.3 pode-se comparar os itens 1, 2 e 3 com 4, 5 e 6. De início, comparando o item 1 e o item 4, conclui-se que o grau de dificuldade do item 1 é muito superior ao do item 4, pois, para ter uma probabilidade de 50% de acertar o item 1, necessita-se de uma proficiência perto de 1, já para o item 4 basta ter uma proficiência perto de -2.

Logo, itens mais difíceis costumam situar-se mais à direita no eixo das proficiências. Observando, agora, o item 6, nota-se que ele não está tão inclinado em relação ao eixo das abscissas, como os outros, então isto indica que ele é o menos discriminativo dos itens. Assim, itens mais difíceis (itens 1, 2 e 3) estão localizados na extremidade mais alta da escala de habilidades (à direita da origem), enquanto os itens mais fáceis estão localizados na extremidade mais baixa da escala de habilidades (à esquerda da origem). Isto, como se observou, pode ser visto claramente no gráfico da CCI.

Comparando-se os itens 1 e 2 (ou itens 1, 3 e 4 "curvas mais íngremes" com itens 2, 5 e 6 "curvas mais suaves"), percebemos a influência do parâmetro de discriminação (a_i) na inclinação da CCI. A comparação dos itens 1 e 3 mostra a influência do parâmetro de acerto ao acaso c no eixo vertical desta figura, pois, com uma baixíssima proficiência, tem-se uma chance maior que 20% de se acertar o item.

2.2.6 Interpretação dos parâmetros aplicados nas Curvas Características dos Itens.

Nota-se que P(Uij = 1|θ_j) pode ser vista como a proporção de respostas corretas ao item i entre todos os respondentes da população com habilidade/proficiência θ fixada. A relação existente entre P(Uij = 1|θ) e os parâmetros do modelo é mostrada na figura 2.4, que é chamada de Curva Característica do Item – CCI.

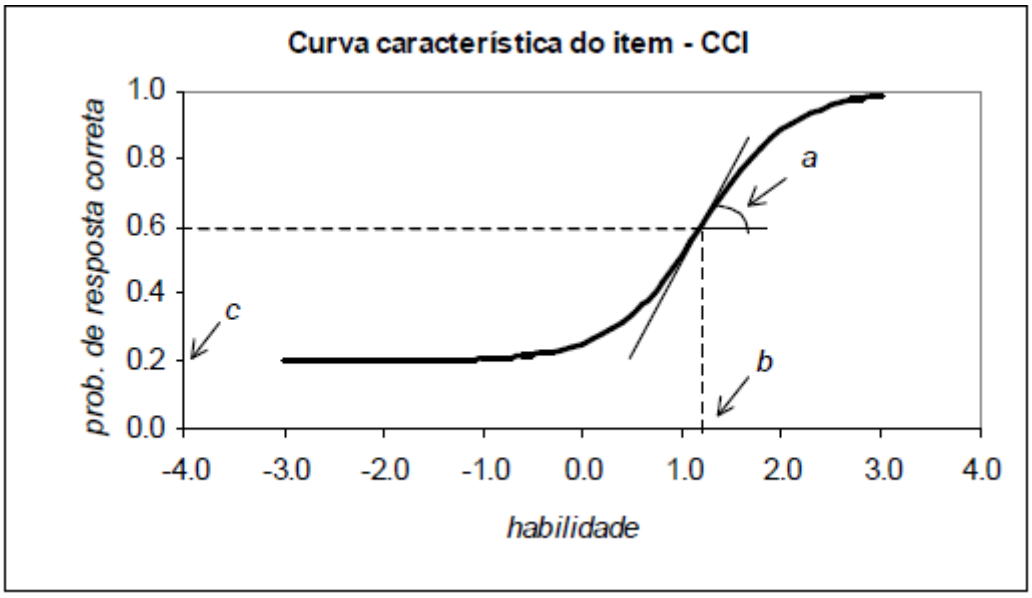

Figura 2.4: Modelo de Curva Característica com a visualização dos Itens.

A seguir, faremos um estudo mais pormenorizado sobre a influência de cada parâmetro na CCI na área educacional.

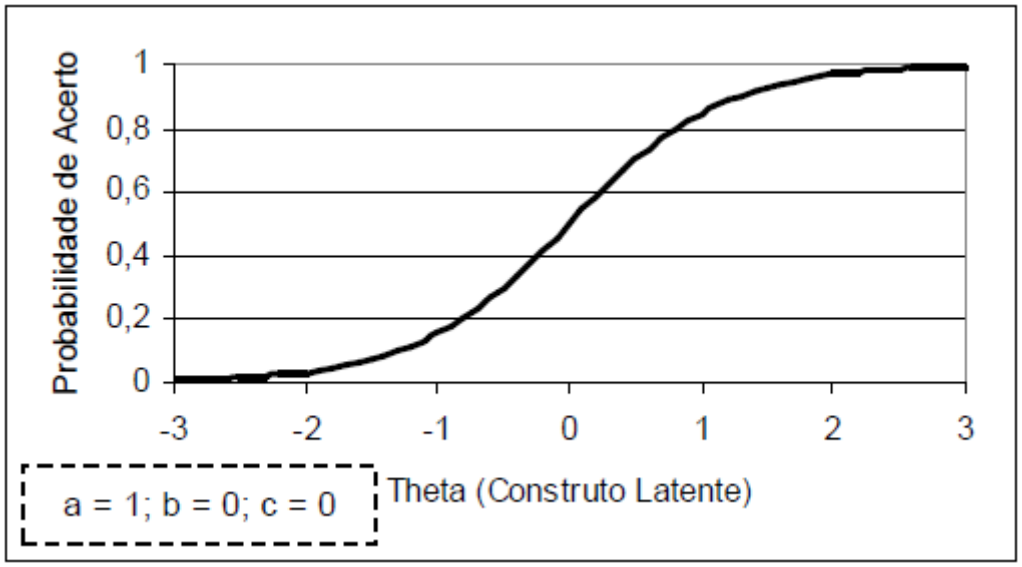

Figura 2.5: efeitos da variação dos parâmetros 1

Na figura 2.5, o parâmetro c é igual a 0. Isso significa que a probabilidade de acerto ao acaso é praticamente nula, que também é interpretada como a probabilidade de acerto por um indivíduo de baixa habilidade.

O parâmetro b, a dificuldade do item, é medido na mesma escala do construto latente. Ele é interpretado como a habilidade necessária para o indivíduo responder corretamente com probabilidade igual a $\frac{1+c_i}{2}$ que, neste caso, é igual a 0,5 ($c_i = 0$). Ou seja, a habilidade necessária para que o indivíduo tenha probabilidade igual a 0,5 de acertar o item é igual a zero (valor do parâmetro b).

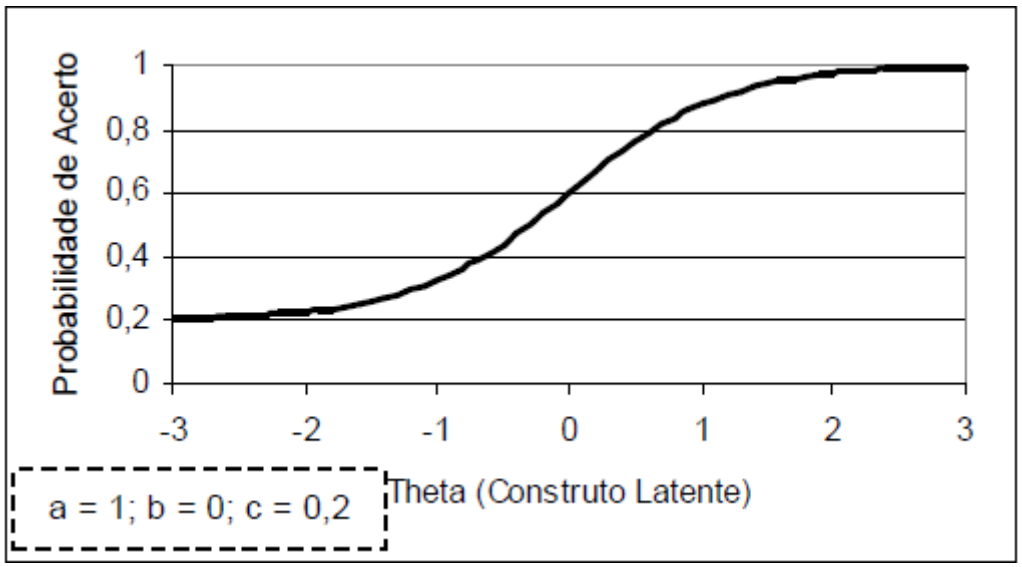

Figura 2.6: efeitos da variação dos parâmetros 2

O gráfico 2.6 ilustra a modificação no parâmetro c. Observa-se que a probabilidade de acerto do item por indivíduos de baixa habilidade aumentou, deslocando toda a curva verticalmente para cima, mas sem introduzir alteração na forma da curva.

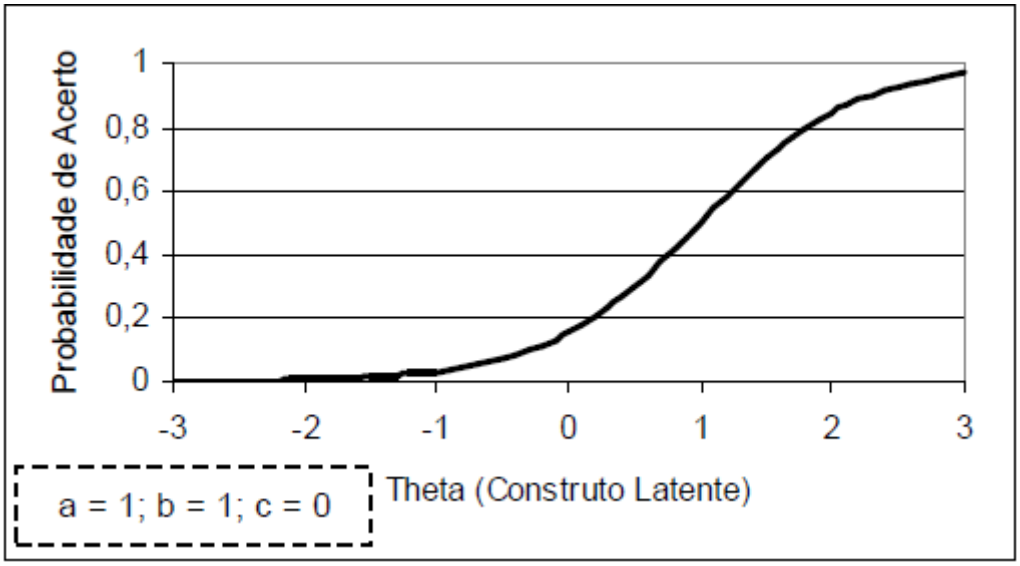

Figura 2.7: efeitos da variação dos parâmetros 3

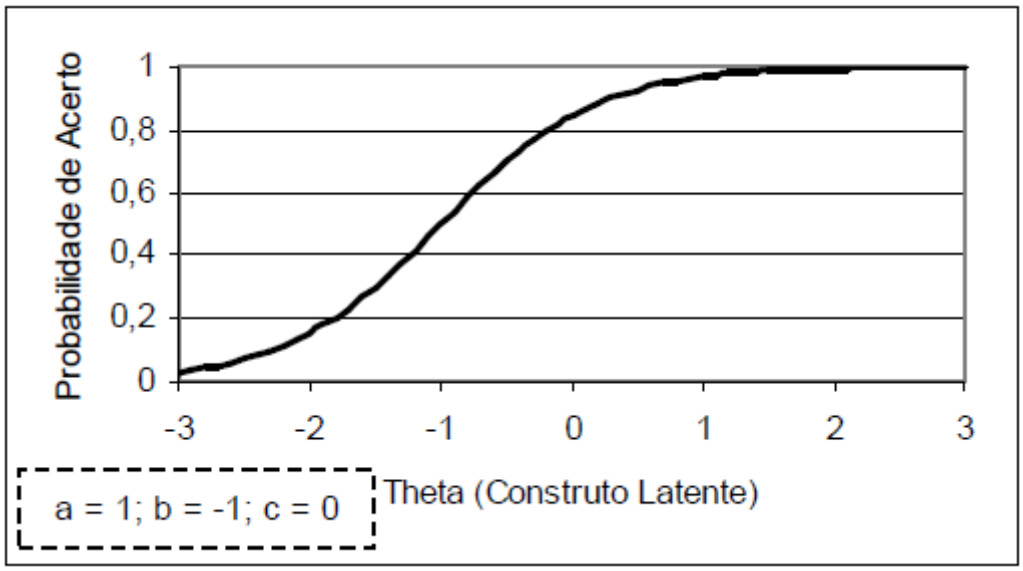

Figura 2.8: efeitos da variação dos parâmetros 4

As figuras 2.7 e 2.8 simulam as alterações no parâmetro b. Quanto maior o valor de b, maior é a habilidade necessária para que a probabilidade de resposta correta ao item seja igual a $\frac{1+c_i}{2}$.

A modificação no parâmetro b provoca um deslocamento horizontal na curva, sem que haja uma mudança em sua forma.

Quanto maior é o valor de b, mais a curva se desloca para a direita.

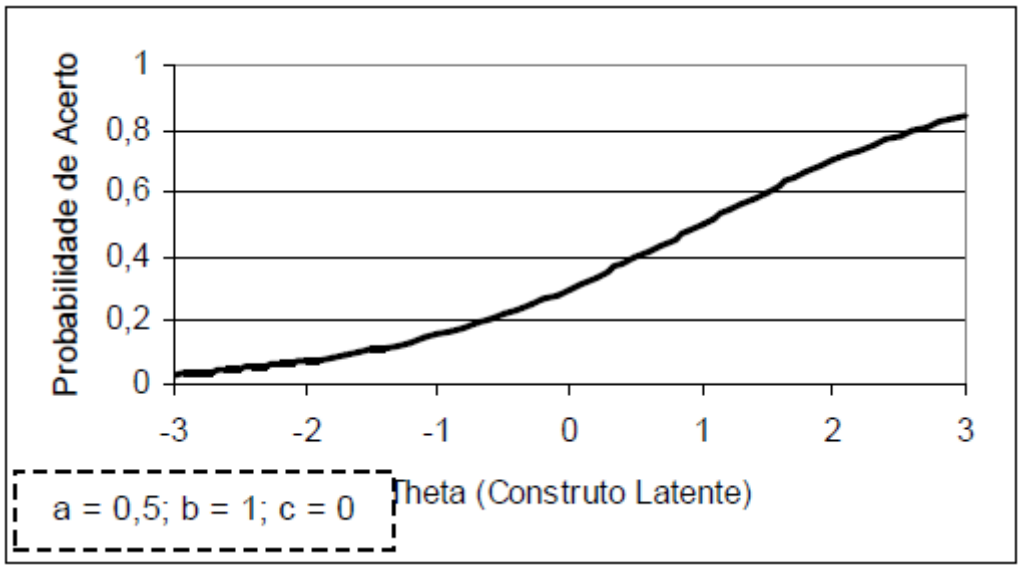

Figura 2.9: efeitos da variação dos parâmetros 5

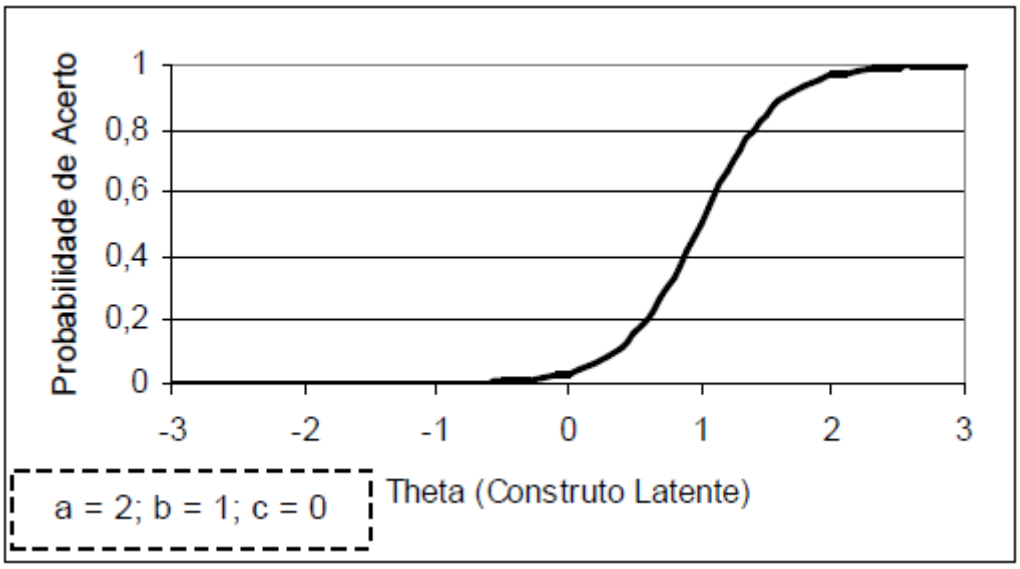

Figura 2.10: efeitos da variação dos parâmetros 6

As figuras 2.9 e 2.10 simulam as alterações no parâmetro a. Quanto menor é o valor de a ($a>0$), menor é o poder de discriminação do item.

No extremo, quando $a=0$, a curva gerada é uma reta, ou seja, qualquer que seja a habilidade, a probabilidade de acerto no item é exatamente a mesma.

Baixos valores de a indicam que o item tem pouco poder de discriminação, ou seja, indivíduos com habilidades muito diferentes têm aproximadamente a mesma probabilidade de acertar o item.

Apesar de matematicamente ser possível, não se esperam valores negativos de a, já que indicariam a probabilidade de o acerto diminuir com o aumento da habilidade.

Na TCT, não se pensa diferente, pois alta discriminação é entendida como uma característica desejável no item, sendo um indicador da qualidade deste. Uma aplicação importante que pode ser feita com o índice de discriminação é a seleção do melhor, como exemplo da discriminação de um item na TCT a qual os pesquisadores demonstram de duas maneiras:

- Comparando a proporção(θ_{GS}) de examinandos que responderam ao item corretamente na parte superior do grupo (GS: escores mais altos) com a proporção (θ_{GI}) de examinandos que responderam corretamente no grupo inferior (GI: escores mais baixos). Esta comparação é feita mediante um teste estatístico de proporções e, se a diferença entre as proporções for estatisticamente significativa, então o item é discriminativo. O grupo superior é considerado como os 27% dos examinandos que tiveram escores mais altos no teste e o grupo inferior é composto pelos 27% dos examinados que tiveram escores mais baixos.

  Nesse procedimento, tem-se, então, a hipótese nula $H_0 : \theta_{GS} = \theta_{GI}$, que é testada em confronto com a hipótese alternativa $H_1 : \theta_{GS} > \theta_{GI}$. A estatística do teste de hipótese tem distribuição normal padrão, assintótica, e é dada pela expressão

$$z = \frac{\left(\hat{\theta}_{GS} - \hat{\theta}_{GI}\right) - (\theta_{GS} - \theta_{GI})}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \sim N(0,1)$$

  A diferença entre as proporções amostrais dos grupos e a diferença hipotética que figura no numerador e o erro padrão da estatística amostral no denominador. O parâmetro verdadeiro θ é estimado pela expressão

$$z = \frac{n_{GS}\hat{\theta}_{GS} + n_{GI}\hat{\theta}_{GI}}{n_{GS} + n_{GI}},$$

  com $n = n_{GS} + n_{GI}$ sendo n_{GS} o número de indivíduos nos grupo superior e n_{GI} o número no grupo inferior.

- A maneira mais utilizada é a que calcula um coeficiente de correlação entre as variáveis: escore dos examinandos no teste (discreta: 0, 1, 2, ..., n (acertos)) e

resposta dos examinandos ao item (dicotômica: 1 ou 0). O coeficiente de correlação que deve ser aplicado é o bisserial, cuja expressão é:

$$\hat{\rho}_b = \frac{\overline{X}_p - \overline{X}_t}{S_t} \cdot \frac{\hat{\theta}}{y}.$$

em que $\overline{X}_p$ é o escore médio entre todos os examinandos que acertaram o item i, $\overline{X}_t$ é o escore médio global, S_t é o desvio-padrão do teste, $\hat{\theta}$ e a proporção dos examinandos que acertaram o item i e $y = f(z)$ a ordenada na curva da Normal Padrão correspondente à área de $\hat{\theta}$.

Geralmente, aplica-se este coeficiente, pois se obtêm uma variável discreta e outra dicotômica.

O parâmetro b é medido na mesma escala do traço latente θ. Na área educacional, ele está associado à dificuldade de um avaliando responder corretamente a uma questão. À medida que b cresce, aumenta o grau de dificuldade do item.

Como já foi citado, a vantagem significativa da TRI sobre a TCT é a de que o parâmetro de dificuldade b e o traço latente/proficiência (θ) estão na mesma escala.

No contexto da educação, o parâmetro c está associado à "probabilidade de um respondente (examinando) com baixa proficiência no tema do item acertá-lo". Pode ser denominado acerto casual. O parâmetro c não depende da escala, é uma probabilidade e, portanto, assume valores entre 0 e 1.

Tipicamente c assume valores menores do que o valor que resultaria se os examinados respondessem aleatoriamente ao item. Como Lord (1967) notou, esse fenômeno provavelmente pode ser atribuído à esperteza dos confeccionadores dos itens de uma prova em desenvolver escolhas bastante atrativas para os respondentes, mas incorretas.

Já o parâmetro θ representa o traço latente/proficiência, uma variável que não pode ser medida diretamente. Teoricamente, o parâmetro da proficiência pode assumir valores de $-\infty$ a $+\infty$. É preciso, portanto, estabelecer uma origem e uma unidade de medida para definição da escala. Estes valores são determinados de forma a representarem, respectivamente, a média (μ) e o desvio-padrão (σ) das proficiências inseridas no estudo. Uma escala bastante utilizada na TRI é aquela com $\mu = 0$ e $\sigma = 1$, e, nesse caso, os valores significativos do parâmetro b variam (tipicamente) entre -2 e $+2$.

Exemplo: seja uma escala com $\mu = 0$ e $\sigma = 1$, representada por (0, 1), uma

proficiência $\theta = 1,20$. Esta proficiência está 1,20 desvio-padrão acima da proficiência média θ.

Caso fosse utilizada a escala (200, 40) e o valor da proficiência $\theta = 248$, a interpretação seria a mesma, ou seja, estaria 1,20 desvio-padrão acima da proficiência média. O grande desafio está na criação de uma interpretação prática para a escala de proficiência θ. É evidente que a interpretação fica muito fácil na escala com $\mu = 0$ e $\sigma = 1$. Isto pode ser visto a partir da transformação de escala, voltando então à forma elaborada por Rasch e melhorada por Lord, em que:

$$a\,(\theta - b) = (a/40)\,[(40 \cdot \theta + 200) - (40 \cdot b + 200)] = a\,(\theta - b)$$

sendo que $a\,(\theta - b)$ é a parte do modelo probabilístico proposto envolvida na transformação. Assim, segue que, para:

- proficiência: $\theta^* = 40 \cdot \theta + 200$
- dificuldade: $b^* = 40 \cdot b + 200$
- discriminação: $a^* = \frac{a}{40}$

– a probabilidade $P\,(U_i = 1|\theta) = P\,(U_i = 1|\theta^*)$.

Exemplificando, se os valores dos parâmetros a e b, na escala (0,1) são, respectivamente, $a = 0,80$ e $b = -0,20$, então seus correspondentes na escala (200; 40) são, respectivamente, $a = 0,02 = 0,80/40$ e $b = 192 = 40(-0,20) + 200$.

Além disso, dado um respondente com proficiência $\theta = 1$, medida na escala (0 , 1) tem sua proficiência representada por $\theta^* = 40.1 + 200 = 240$ na escala (200; 40) e dado c = 0,20 , que é o acerto ao acaso para aplicar no ML3, tem-se:

$$P(U_1 = 1|\theta = 1) = 0,20 + (1 - 0,20)\frac{1}{1+e^{-1,7x0,80x(1-(-0,20))}} = 0,8692$$

$$P(U_1 = 1|\theta^* = 240) = 0,20 + (1 - 0,20)\frac{1}{1+e^{-1,7x0,02x(240-192)}} = 0,8692$$

ou seja, a probabilidade de determinado avaliado responder corretamente a um item é sempre a mesma, independentemente da escala utilizada para medir a sua proficiência θ, ou seja, θ é invariante à escala de medida. Segundo Andrade (2000), não faz o menor sentido analisar itens a partir dos valores dos parâmetros de discriminação a e de dificuldade b sem conhecer a escala de proficiência θ na qual eles foram determinados.

2.3 Função de Informação do Teste

Uma vez aplicado um conjunto de itens (teste) e estimado o nível de proficiência θ de um respondente, a TRI permite calcular o erro padrão (EP) de estimação do nível de proficiência deste respondente no qual o teste foi aplicado. Essa é uma diferença fundamental da TRI com a TCT, que assume que o erro é o mesmo para todos os examinandos.

A informação fornecida pelo teste é a soma das funções de informação dos itens em θ. Para analisar as CCIs, uma medida bastante utilizada em conjunto é a função de Informação do Item, que é dada por:

$$I_i(\theta) = \frac{\left[P_i^{'(\theta)}\right]^2}{P_i(\theta)Q_i(\theta)}.$$

em que $i = 1, 2, ..., n$.

Sendo:

- $I_i(\theta)$: é a "informação" fornecida pelo item i no nível de proficiência θ;
- $P_i(\theta)$: é a função de resposta ao item;
- $P_i(\theta) = P(X_{ij} = 1|\theta)$;
- $Q_i(\theta) = 1 - P_i(\theta)$;
- $P_i'(\theta)$ é a derivada $P_i(\theta)$.

No caso do modelo logístico de 3 parâmetros, a equação de informação do item é:

$$I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2$$

ou simplificando para parâmetros dos itens com valores já conhecidos, obtém-se a seguinte função de informação do item:

$$I_i(\theta) = \frac{2,89a_i^2(1 - c_i)}{\left[c_i + e^{1,7a_i(\theta - b_i)}\right]\left[1 + e^{-1,7a_i(\theta - b_i)}\right]^2}$$

Ela fornece a contribuição do item na estimação da proficiência, ao longo de toda a sua escala.

Os gráficos da figura 2.11 ilustram o efeito da variação dos parâmetros na forma da Curva de Informação do Item, ou seja, na contribuição de um item para a medida da escala do construto latente.

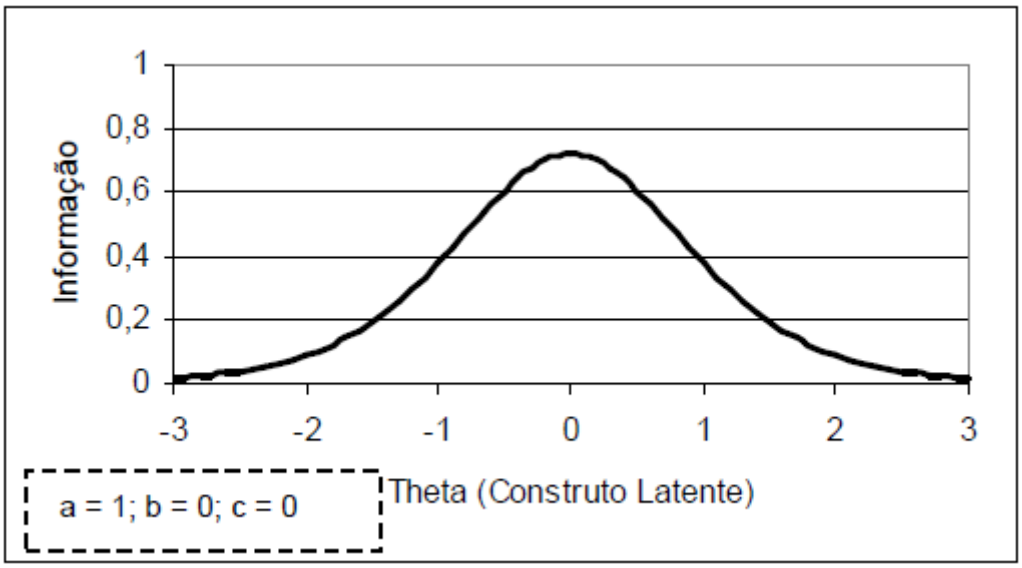

Figura 2.11: Efeito da variação dos parâmetros na forma da curva 1

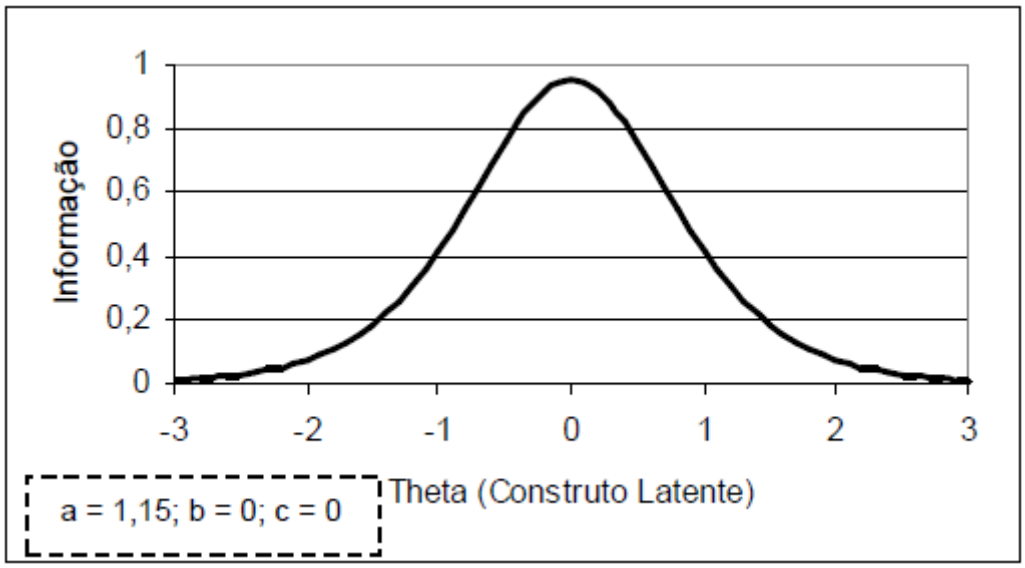

Figura 2.12: Efeito da variação dos parâmetros na forma da curva 2

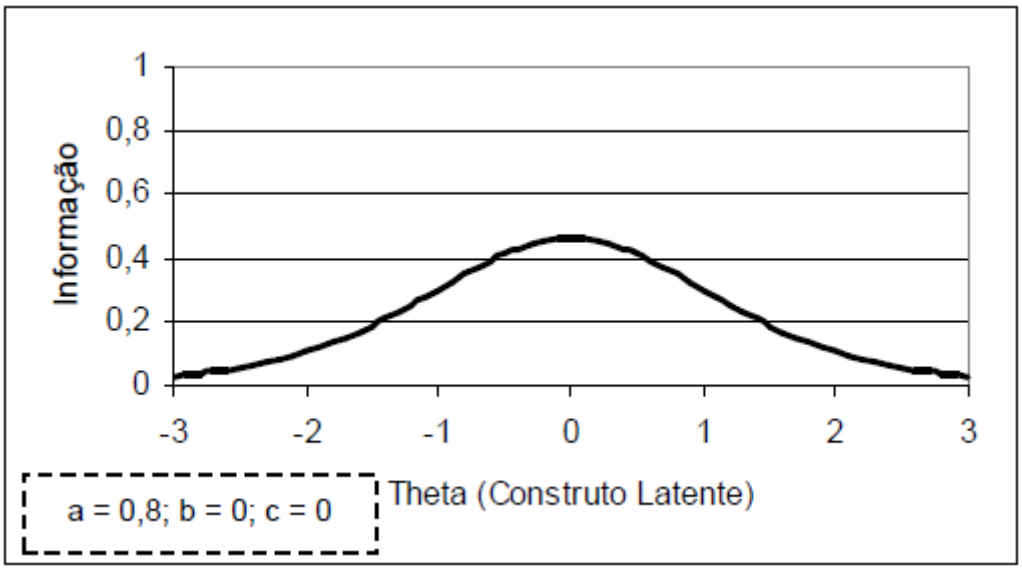

Figura 2.13: Efeito da variação dos parâmetros na forma da curva 3

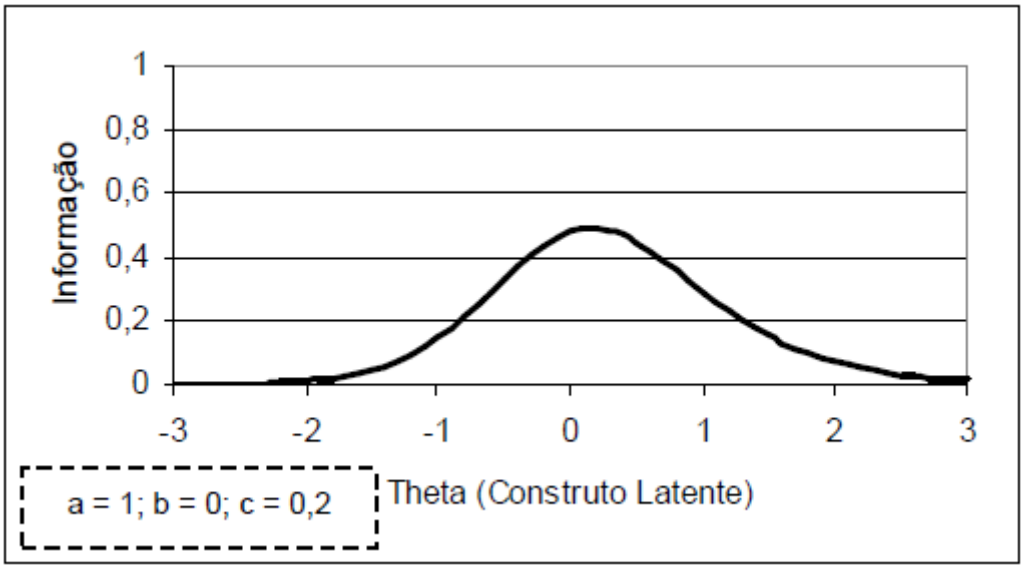

Figura 2.14: Efeito da variação dos parâmetros na forma da curva 4

As curvas das figuras 2.11, 2.12, 2.13 e 2.14 possibilitam a visualização do efeito dos parâmetros do modelo sobre a quantidade de informação do item.

A informação é maior quando b_i se aproxima de θ. Nos quatro gráficos, observa-se que a informação é maior para valores de θ próximos de zero, ou seja, para valores próximos deb_i.

Tem-se também que quanto maior o valor dea_i, ou seja, quanto maior a discriminação do item, maior é a informação.

Isso pode ser visualizado por meio da comparação dos gráficos das figuras 2.11, 2.12 e 2.13.

Por fim, quanto maior é c_i, menor é a informação, ou seja, quanto maior é a probabilidade de acerto ao acaso, menor é a quantidade de informação com a qual o item contribui para a medida da escala do construto latente.

A função de informação do teste é dada por:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$

sendo ela inversamente proporcional ao erro padrão quadrático de estimação de θ, ou seja:

$$EP = \frac{1}{\sqrt{I(\theta)}}$$

$$\sqrt{I(\theta)} = \frac{1}{EP}$$

$$I(\theta) = \frac{1}{(EP)^2}$$

Quanto maior for a função de informação $I(\theta)$, menor será o erro padrão de estimação (EP), e, portanto, maior será a precisão da estimação de habilidade. O erro padrão de estimativa dá a precisão com que é estimado θ.

Quanto maior o erro, menor a precisão, e seu tamanho depende de alguns fatores, tais como:

- 1º–Do número de itens aplicados, em geral, pois, ao se aumentar a quantidade de itens, diminui-se o EP,
- 2º–Da capacidade discriminatória dos itens, pois, ao se aumentar o parâmetro de discriminação a, diminui-se o EP,
- 3º–Da diferença entre b e θ, pois quanto mais próximo está b de θ, menor será o EP.

2.4 Unidimensionalidade e Independência Local

Unidimensionalidade:

Supõe-se que deve haver apenas uma habilidade/proficiência (θ) responsável pela elaboração de todos os itens que compõem um teste (prova). No entanto, parece claro que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa. Contudo, para satisfazer

o postulado da unidimensionalidade, é suficiente admitir que haja uma proficiência (θ) dominante (um fator dominante) responsável pelo conjunto de itens. Este fator é o que se supõe estar sendo medido pelo teste (ANDRADE, 2000).

Independência Local:

A independência local entre os itens de um teste, a qual será demonstrada a seguir, pressupõe que a resposta de um examinado a determinado item não depende das demais respostas dadas aos outros itens. A independência local é decorrente da unidimensionalidade porque, simplesmente, significa que a resposta a um item só depende de seus parâmetros (a_i , b_i , c_i) e de θ e não está influenciada pela ordem de apresentação dos itens, ou pelas resposta que já tenham sido dadas.

Sendo U_i a resposta do examinado para um item i (i = 1,2,..., n), então $P(U_i|\theta)$ é a probabilidade de resposta de um examinado com proficiência θ; $P(U_i = 1|\theta)$ denota a probabilidade de uma resposta correta, e $P(U_i = 0|\theta)$ denota a probabilidade de uma resposta incorreta. Sob a condição de independência local, tem-se que a probabilidade de um examinando com proficiência θ acertar os n itens é:

$$P(U_1, U_2, ..., U_n|\theta) = P(U_1|\theta)\, P(U_2|\theta) ... P(U_n|\theta)$$
$$P(U_1, U_2, ..., U_n|\theta) = \prod_{i=1}^{n} P(U_i|\theta)$$

A propriedade de independência local estabelece que, para determinado examinado, a probabilidade de um padrão de resposta em um conjunto de itens é igual ao produto de probabilidades associado às respostas dos examinados em respostas aos itens individuais.

Por exemplo, se o padrão de resposta para um dado examinado em três itens for $(1, 1, 0)$, isto é, $U_1 = 1$, $U_2 = 1$, e $U_3 = 0$, então:

$$P(U_1 = 1, U_2 = 1, U_3 = 0|\theta) = P(U_1 = 1|\theta)\, P(U_2 = 1|\theta)\, P(U_3 = 0|\theta) = P_1 P_2 Q_3$$

onde

$$P(U_1 = 1|\theta)\, P(U_2 = 1|\theta)\, P(U_3 = 0|\theta) = P_1 P_2 Q_3$$
$$P_i = P(U_i = 1|\theta)$$
$$\mathrm{Q}_i = 1 - P_i$$

(HAMBLETON, 1991).

Capítulo

3

Proposta para correção da redação

3.1 Modelo atual da correção da redação

De acordo com o documento "Guia do participante: a redação no ENEM de 2013", publicado pelo INEP, a prova de redação exigirá a produção de um texto em prosa, do tipo dissertativo-argumentativo, sobre um tema de ordem social, científica, cultural ou política. Os aspectos a serem avaliados relacionam-se às competências que devem ter sido desenvolvidas durante os anos de escolaridade.

Na redação, o candidato defenderá uma tese, uma opinião a respeito do tema proposto, apoiada em argumentos consistentes estruturados de forma coerente e coesa, de modo a formar uma unidade textual. O texto será redigido de acordo com a modalidade escrita formal da Língua Portuguesa. Por fim, o candidato elaborará uma proposta de intervenção social para o problema apresentado no desenvolvimento do texto que respeite os direitos humanos.

O texto produzido pelo candidato será avaliado por, pelo menos, dois professores de forma independente, sem que um conheça a nota atribuída pelo outro de acordo com seguintes critérios:

- Competência 1: Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.
- Competência 2: Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
- Competência 3: Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.

- Competência 4: Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
- Competência 5: Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Os critérios constituem a matriz de referência da redação publicada no Diário Oficial da União em 9 de maio de 2013 (anexo IV, página 83) e indicam ao candidato as bases a seguir para realizar uma boa redação, conforme a transcrição a seguir.

ANEXO IV

MATRIZ DE REFERÊNCIA PARA REDAÇÃO

Baseada nas cinco competências da Matriz de Referência para Redação, a proposta da Redação do Enem é elaborada de forma a possibilitar que os participantes, a partir de uma situação-problema e de subsídios oferecidos, realizem uma reflexão escrita sobre um tema de ordem política, social ou cultural, produzindo um texto dissertativo-argumentativo em prosa.

COMPETÊNCIAS EXPRESSAS NA MATRIZ DE REFERÊNCIA PARA REDAÇÃO DO ENEM E NÍVEIS DE CONHECIMENTOS ASSOCIADOS

I - Demonstrar domínio da modalidade escrita formal da língua portuguesa.

Nível 0: Demonstra desconhecimento da modalidade escrita formal da língua portuguesa.

Nível 1: Demonstra domínio precário da modalidade escrita formal da língua portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.

Nível 2: Demonstra domínio insuficiente da modalidade escrita formal da língua portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.

Nível 3: Demonstra domínio mediano da modalidade escrita formal da língua portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.

Nível 4: Demonstra bom domínio da modalidade escrita formal da língua portuguesa e de escolha de registro, com poucos desvios gramaticais e de convenções da escrita.

Nível 5: Demonstra excelente domínio da modalidade escrita formal da língua portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizem reincidência.

II - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.

Nível 0: Fuga ao tema/não atendimento à estrutura dissertativo-argumentativa.

Nível 1: Apresenta o assunto, tangenciando o tema ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais.

Nível 2: Desenvolve o tema recorrendo à cópia de trechos dos textos motivadores ou apresenta domínio insuficiente do texto dissertativo-argumentativo, não atendendo à estrutura com proposição, argumentação e conclusão.

Nível 3: Desenvolve o tema por meio de argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.

Nível 4: Desenvolve o tema por meio de argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.

Nível 5: Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sociocultural produtivo e apresenta excelente domínio do texto dissertativo-argumentativo.

III - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.

Nível 0: Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista.

Nível 1: Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista.

Nível 2: Apresenta informações, fatos e opiniões relacionados ao tema, mas desorganizados ou contraditórios e limitados aos argumentos dos textos motivadores, em defesa de um ponto de vista.

Nível 3: Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista.

Nível 4: Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista.

Nível 5: Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista.

IV - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.

Nível 0: Não articula as informações.

Nível 1: Articula as partes do texto de forma precária.

Nível 2: Articula as partes do texto, de forma insuficiente, com muitas inadequações e apresenta repertório limitado de recursos coesivos.

Nível 3: Articula as partes do texto, de forma mediana, com inadequações, e apresenta repertório pouco diversificado de recursos coesivos.

Nível 4: Articula as partes do texto com poucas inadequações e apresenta repertório

diversificado de recursos coesivos.

Nível 5: Articula bem as partes do texto e apresenta repertório diversificado de recursos coesivos.

V - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Nível 0: Não apresenta proposta de intervenção ou apresenta proposta não relacionada ao tema ou ao assunto.

Nível 1: Apresenta proposta de intervenção vaga, precária ou relacionada apenas ao assunto.

Nível 2: Elabora, de forma insuficiente, proposta de intervenção relacionada ao tema, ou não articulada com a discussão desenvolvida no texto.

Nível 3: Elabora, de forma mediana, proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.

Nível 4: Elabora bem proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.

Nível 5: Elabora muito bem proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.

A redação será corrigida por dois avaliadores e, cada avaliador atribuirá uma nota entre 0 (zero) e 200 (duzentos) pontos para cada uma das cinco competências. A soma desses pontos comporá a nota total de cada avaliador, que pode chegar a 1000 (mil) pontos. A nota final do participante será a média aritmética das notas totais atribuídas pelos dois avaliadores.

No caso de discrepância nas notas atribuídas, a redação será corrigida por um terceiro avaliador independente, e a nota final será a média aritmética das duas notas totais que mais se aproximarem.

Considera-se discrepância a divergência de notas atribuídas pelos avaliadores quando elas diferirem, no total, por mais de 100 (cem) pontos ou a diferença for superior a 80 (oitenta) pontos em qualquer uma das competências.

As tabelas seguintes apresentam os seis níveis de desempenho utilizados para avaliar cada competência bem como a pontuação correspondente.

Tabela 3.1: competência 1

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Demonstra excelente domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizem reincidência.
160	4	Demonstra bom domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com poucos desvios gramaticais e de convenções da escrita.
120	3	Demonstra domínio mediano da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.
80	2	Demonstra domínio insuficiente da modalidade escrita formal da Língua Portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.
40	1	Demonstra domínio precário da modalidade escrita formal da Língua Portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.
0	0	Demonstra desconhecimento da modalidade escrita formal da Língua Portuguesa.

Tabela 3.2: competência 2

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sociocultural produtivo, e apresenta excelente domínio do texto dissertativo-argumentativo.
160	4	Desenvolve o tema por meio de argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
120	3	Desenvolve o tema por meio de argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
80	2	Desenvolve o tema recorrendo à cópia de trechos dos textos motivadores ou apresenta domínio insuficiente do texto dissertativo-argumentativo, não atendendo à estrutura com proposição, argumentação e conclusão.
40	1	Apresenta o assunto, tangenciando o tema, ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais.
0	0	Fuga ao tema/não atendimento à estrutura dissertativo-argumentativa.

Tabela 3.3: competência 3

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista.
160	4	Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista.
120	3	Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista.
80	2	Apresenta informações, fatos e opiniões relacionados ao tema, mas desorganizados ou contraditórios e limitados aos argumentos dos textos motivadores, em defesa de um ponto de vista.
40	1	Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista.
0	0	Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista.

Tabela 3.4: competência 4

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Articula bem as partes do texto e apresenta repertório diversificado de recursos coesivos.
160	4	Articula as partes do texto com poucas inadequações e apresenta repertório diversificado de recursos coesivos.
120	3	Articula as partes do texto, de forma mediana, com inadequações e apresenta repertório pouco diversificado de recursos coesivos.
80	2	Articula as partes do texto, de forma insuficiente, com muitas inadequações e apresenta repertório limitado de recursos coesivos.
40	1	Articula as partes do texto de forma precária.
0	0	Ausência de marcas de articulação, resultando em fragmentação das ideias.

Tabela 3.5: competência 5

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Elabora muito bem proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.
160	4	Elabora bem proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
120	3	Elabora, de forma mediana, proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
80	2	Elabora, de forma insuficiente, proposta de intervenção relacionada ao tema ou não articulada com a discussão desenvolvida no texto.
40	1	Apresenta proposta de intervenção vaga, precária ou relacionada apenas ao assunto.
0	0	Não apresenta proposta de intervenção ou apresenta proposta não relacionada ao tema ou ao assunto.

Em um país de dimensões continentais, a prova do ENEM, com grande quantidade de inscritos, adotada por uma quantidade crescente de instituições de ensino superior como parte do processo de seleção aos cursos de graduação e também como critério para concessão de bolsas, como o Ciência sem Fronteiras, possui números superlativos em relação à logística de correção da redação, representando um grande desafio para os organizadores da avaliação a compatibilização constante dos critérios estabelecidos para a correção, o treinamento e o controle de uma equipe de professores, o que carece então de uma forma moderna de operacionalização do processo, para que este seja seguro, dinâmico, justo e democrático.

O Centro de Seleção e de Promoção de eventos da Universidade de Brasília (Cespe/UnB) integrante do Consórcio Cesgranrio–Cespe, optou, em 2006 pela informatização do processo de correção da redação, o que apresentou como benefícios:

- Com a digitalização prévia de todas as redações, garantiu-se a preservação do original de cada participante.
- Gerou-se automaticamente relatórios acerca do trabalho de cada corretor.
- Possibilitou-se correções múltiplas simultâneas, com geração automática de compatibilização.
- Criou-se de um grupo de controle do processo.

- Controle de eficiência do cumprimento da tarefa pelo corretor.
- Geração de relatórios amostrais por corretor, por equipe, por estado, etc.
- Geração automática de banco de informações para pesquisas futuras.

Cada redação é corrigida por dois corretores, de forma independente, sem que um conheça a pontuação dada pelo outro. Caso haja discrepância no total de pontos atribuídos às competências, a redação passa por uma terceira correção, do supervisor, que faz a compatibilização das duas avaliações. O mesmo deve ocorrer se uma redação for considerada simultaneamente pontuada e desconsiderada pelos corretores.

A comparação das notas dadas pelos corretores é feita instantaneamente pelo sistema eletrônico de correção, gerando compatibilização ou não, de acordo com os critérios pre-estabelecidos.

Quanto à compatibilização por competência, uma vez que o sistema foi configurado para que todas as redações tivessem correção dupla, feita por dois avaliadores distintos, o ajuste de critérios foi feito competência por competência, com a geração de relatórios em tempo real, possibilitando o controle total quanto à qualidade técnica na observância dos critérios de correção definidos pelo INEP, o que não seria possível no processo de correção manual.

Já em relação à compatibilização por situação, o sistema fornece aos supervisores do processo uma tela para monitoramento das situações definidas por um par de avaliadores numa mesma redação. Por exemplo, no caso de apenas um dos corretores indicar que a redação não era uma dissertação, um relatório é gerado automaticamente para o supervisor.

Convém ressaltar que, apesar de a correção da prova de conhecimentos do ENEM ser feita a partir de critérios estabelecidos pela TRI, a redação pauta-se em aspectos da TCT pois se preocupa em explicar o resultado final, isto é, a soma das respostas dadas a uma série de itens como os supracitados, expressa no chamado escore total.

O que representa o escore do respondente? Supostamente ele expressa a magnitude daquilo que o teste queria medir no respondente.

> Contudo, toda e qualquer operação empírica, se sabe, é sujeita a erros. Consequentemente, esse escore bruto do sujeito não pode ser a expressão pura da magnitude daquilo que o teste queria medir no sujeito, mas deve conter igualmente uma porção de erros. (PASQUALI, 2003)

O foco da TCT não é o traço latente, sim o escore em um teste, representando dessa forma uma limitação do modelo.

Hambleton, Swaminathan e Rogers (1991) salientam especialmente quatro limitações teóricas graves que a psicometria tradicional contém:

1. Os parâmetros clássicos dos itens (dificuldade e discriminação) dependem diretamente da amostra de sujeitos utilizada para estabelecê-los.

2. A avaliação das aptidões dos testandos também depende do teste utilizado. Assim, testes diferentes que medem a mesma aptidão irão produzir escores diferentes da mesma aptidão para sujeitos idênticos.

3. A definição do conceito de fidedignidade ou precisão na teoria clássica dos testes constitui também uma fonte de dificuldades. Ela é concebida como a correlação entre os escores obtidos de formas paralelas de um teste ou, mais genericamente, como o oposto do erro de medida.

4. Ela é orientada para o teste total e não para o item individual. Toda a informação do item deriva de considerações do teste geral, não podendo, assim, determinar como o examinando se comportaria diante de cada item individual.

Nesta dissertação, será proposto outro modelo para a correção da redação feita pelo candidato, visando padronizar a utilização da TRI no conjunto da prova do ENEM e apresentar novos dados para a gestão da qualidade dos textos elaborados e sua devolutiva para a comunidade educacional.

3.2 Modelos politômicos

Com o avanço dos estudos sobre aplicações da TRI em diferentes contextos, surgiu a necessidade de introduzir, nos testes psicométricos, respostas que não fossem consideradas exclusivamente dicotômicas, o que proporcionou o desenvolvimento de modelos da TRI de natureza acumulativa para respostas politômicas, nominais ou graduadas, como o Modelo de Resposta Nominal de Bock, o Modelo de Resposta Graduada de Samejina, o Modelo de Crédito Parcial proposto por Masters, entre outros.

Os modelos para respostas politômicas utilizam mais intensamente a informação contidas nos questionários (testes), mas, por consequência, necessitam de um número maior de parâmetros a serem estimados.

Nessa categoria de modelos, estão inclusos os modelos tanto para análise de itens abertos (de resposta livre) quanto para a análise de itens de múltipla escolha avaliados de forma ordenada.

O modelo de resposta gradual de Samejima é uma generalização do modelo logístico de 2 parâmetros, assumindo que as categorias de resposta de um item podem ser ordenadas entre si, como uma escala de Likert, a qual consiste em uma série de cinco proposições, das quais o inquirido deve selecionar uma, podendo estas ser: concorda totalmente, concorda, sem opinião, discorda, discorda totalmente. É efetuada uma cotação das respostas que varia de modo consecutivo: +2, +1, 0, -1, -2 ou utilizando pontuações de 1 a 5.

Embora o uso de escalas com outro número de itens, diferente de cinco, represente uma escala de classificação, quando esta não contiver cinco opções de resposta, não se configura uma escala Likert, mas sim do "tipo Likert". No entanto, como Clason e Dormody (1994) ressaltam, muitos estudos têm usado diversas opções, paralelas à escala tradicional de cinco pontos, obtendo resultados satisfatórios.

Na figura 3.1, apresenta-se a CCI do Modelo de Resposta Gradual de Samejima de um item com quatro categorias de resposta.

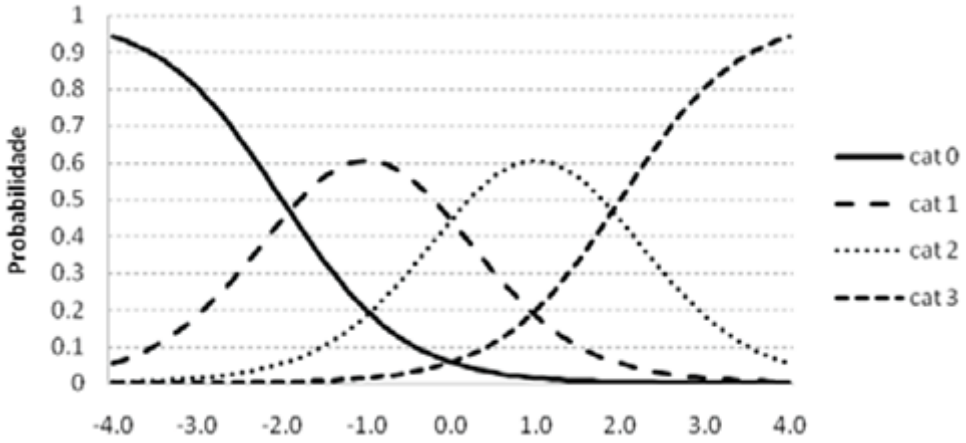

Figura 3.1: CCI do MRG de Samejima.

CCI do Modelo de Resposta Gradual de um item com a= 1,4 e b_1=-2,0 ; b_2=0,0; b_3=2,0.

Observa-se na figura que os respondentes com traço latente até -2,0 têm maior probabilidade de responder a categoria 0. Já os respondentes com traço latente entre -2,0 e 0,0 têm mais chance de alcançarem a categoria 1. Para os respondentes com o traço latente entre 0,0 e 2,0, a maior probabilidade é que respondam à categoria 2, enquanto os respondentes com habilidade acima de 2,0 devem responder à categoria maior, isto é, a categoria 3.

Em 1990, Muraki desenvolveu uma modificação que facilitou o uso do modelo de resposta gradual na análise de questionários nos quais os itens têm o mesmo número de categorias de respostas e estas devem ser igualmente espaçadas.

Há também o modelo de crédito parcial como uma extensão do modelo de Rash para mais de 2 categorias, desenvolvido em 1980 por Masters.

Em 1992, Muraki propôs o modelo de crédito parcial generalizado, flexibilizando o poder de discriminação constante para todos os itens.

Por fim, um quinto modelo é o de escala gradual onde cada item é descrito por um único parâmetro de locação de escala, refletindo o grau de dificuldade de um item.

No quadro a seguir estão representados os modelos estudados da TRI.

Tabela 3.6: Modelos da TRI

Natureza do Item	Dificuldade	Dificuldade e discriminação	Dificuldade e discriminação e "chute"
Respostas dicotômicas	Modelo logístico de 1 parâmetro (modelo de Rasch)	Modelo logístico de 2 parâmetros	Modelo logístico de 3 parâmetros
Respostas politômicas	Modelo de crédito parcial, Modelo de Escala gradual, Modelo de resposta gradual Modificada	Modelo de resposta gradual, Modelo de crédito parcial generalizado	

3.2.1 Modelo de resposta gradual de Samejima

A maior abordagem ao modelo politômico da TRI, distinto do modelo proposto por Rash, é o trabalho de Samejima, construído com base na medida cumulativa de limite de Thurstone. Samejima inicialmente desenvolveu dois modelos para um conjunto de itens em formato de respostas politômicas ordenadas. Muito deste trabalho focou na estimativa de probabilidade máxima para o parâmetro da pessoa (traço).

Os dois modelos tinham uma forma idêntica e se diferenciavam apenas em que um empregava uma ogiva normal para modelar respostas e o segundo usava a função logística para este propósito. Samejima mais tarde expandiu e formalizou seu trabalho anterior para acomodar dados de resposta livre que poderiam concebivelmente incluir um número potencialmente ilimitado de respostas não ordenadas para um item de um teste.

Este modelo, como o modelo de Bock, tenta obter mais informação das respostas dos indivíduos do que simplesmente se eles deram respostas corretas ou incorretas.

Suponha que os escores das categorias de um item i são arranjados em ordem do menor para o maior e denotados por k = 0; 1; ... ;m_i onde (m_i + 1) é o número de categorias do i-ésimo item. A probabilidade de um indivíduo j escolher uma particular categoria ou outra mais alta do item i pode ser dada por uma extensão do modelo logístico de 2 parâmetros:

$$P_{i,k}^{+}(\theta_j) = \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k})}}$$

com i=1,2,3,...,I, j = 1, 2,..., n e k = 0, 1,..., m_i, em que:

- $P_{i,k}^{+}(\theta_j)$ é a probabilidade de um indivíduo j escolher a categoria k ou outra mais alta do item i;
- $b_{i,k}$: é o parâmetro de dificuldade da k-ésima categoria do item i;
- θ_j: representa a habilidade, proficiência (traço latente) do j-ésimo indivíduo;
- a_i é o parâmetro de discriminação (ou de inclinação) do item i, com valor proporcional à inclinação da CCI no ponto;
- D: é o fator de escala constante e igual a 1. Utiliza-se o valor 1,7 quando se deseja que a função logística forneça resultado semelhante ao da função ogiva normal.

O parâmetro de discriminação a varia a cada item, mas é constante dentro dos itens. Essa restrição de igual inclinação em cada categoria tem a finalidade de evitar probabilidades negativas.

Para o parâmetro de dificuldade $b_{i;k}$, por definição, tem-se, sem perda de generalidade, que:

$$b_{i,1} \leq b_{i,2} \leq ... \leq b_{i,m_i}$$

ou seja, devemos ter necessariamente uma ordenação entre o nível de dificuldade das categorias de um dado item, de acordo com a classificação de seus escores. A probabilidade de um indivíduo j receber um escore k no item i é dada então pela expressão:

$$P_{i,k}(\theta_j) = P_{i,k}^{+}(\theta_j) - P_{i,k+1}^{+}(\theta_j)$$

Samejima também define

$$P_{i,0}^{+}(\theta_j)$$

$$e$$

$$P_{i,m_i+1}^{+}(\theta_j)$$

de modo que:

$$P_{i,0}^{+}(\theta_j) = 1$$

$$e$$

$$P_{i,m_i+1}^{+}(\theta_j) = 0$$

Portanto:

$$P_{i,0}(\theta_j) = P_{i,0}^{+}(\theta_j) - P_{i,1}^{+}(\theta_j) = 1 - P_{i,1}^{+}(\theta_j)$$

$$e$$

$$P_{i,m}(\theta_j) = P_{i,m}^{+}(\theta_j) - P_{i,m_i+1}^{+}(\theta_j) = P_{i,m}^{+}(\theta_j)$$

Então temos que:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k+1})}}$$

Note que em um item com (m_i +1) categorias, m_i valores de dificuldade necessitam ser estimados, além do parâmetro de discriminação do item. Assim, para cada item, o número de parâmetros a ser estimado será dado pelo seu número de categorias de resposta. Se, por exemplo, tivermos um teste com I itens, cada um com (m_i+1) categorias de resposta, teremos então $\sum_{i=1}^{I} m_i + I$ parâmetros de item a serem estimados.

O símbolo Theta (θ) representa o valor da variável latente para cada aluno. Este valor é baseado na resposta individual dada a cada um dos indicadores que compõem o índice pelo modelo. Dessa forma, pode-se considerar θ como uma medida do que é avaliado em

uma pesquisa. É importante ressaltar que, devido ao método utilizado, são considerados apenas os itens para os quais os indivíduos forneceram resposta, desconsiderando aquelas em branco.

Analogamente ao modelo de resposta gradual, este também é adequado para itens com categorias de resposta ordenadas. No entanto, é feita uma suposição a mais: a de que os escores das categorias são igualmente espaçados.

Na figura 3.2 temos a representação gráfica do modelo de escala gradual e do modelo de resposta gradual para alguns itens com 4 categorias de resposta. Em todos os itens, o parâmetro a_i foi mantido igual a 1,0. Dessa maneira, podemos verificar a importância dos parâmetros de categoria $b_{i;k}$. Os itens 1 e 4, por terem os parâmetros de categoria igualmente espaçados, podem ser representantes do modelo de escala gradual. Já o modelo de resposta gradual poderia ser representado por qualquer um dos itens acima.

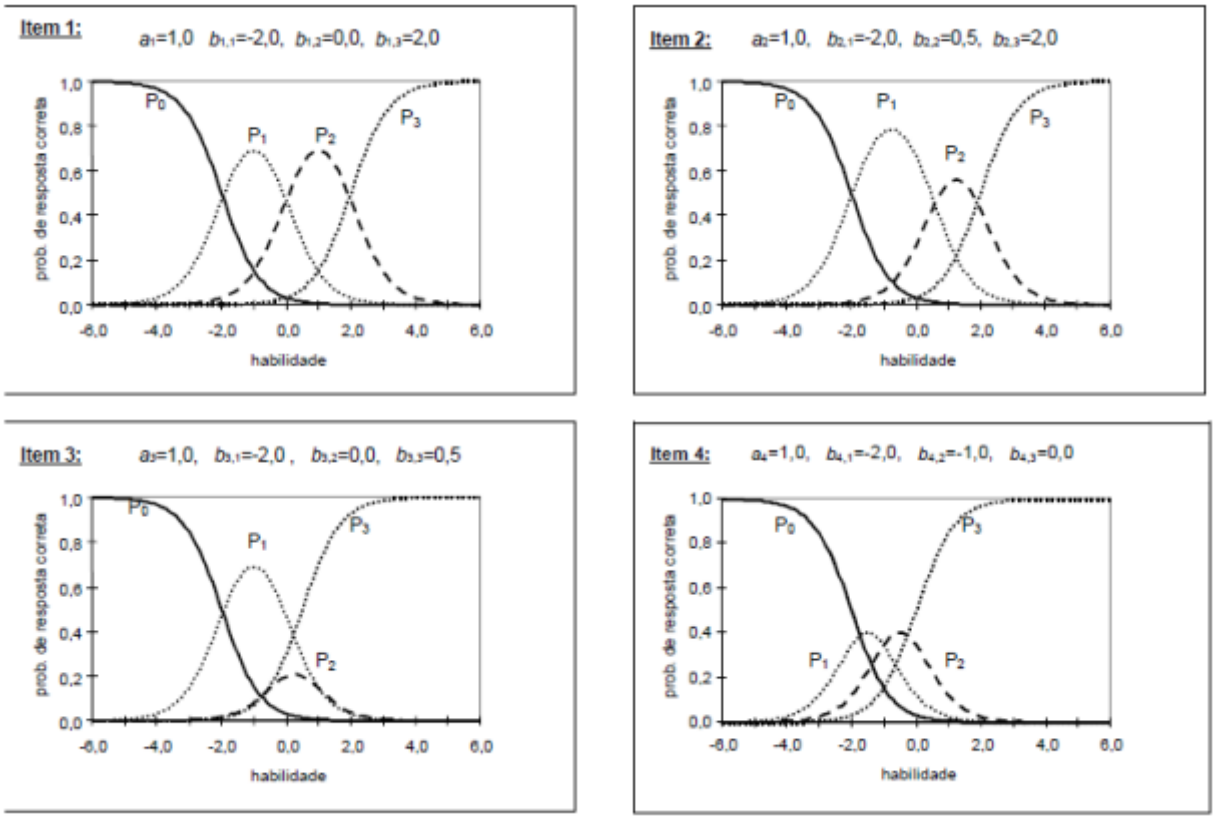

Figura 3.2: Representação gráfica dos modelos de escala gradual e de resposta gradual

Observando o item 1, podemos notar que indivíduos com habilidade até - 2,0 têm maior probabilidade de responder apenas a categoria 0. Já indivíduos com habilidades entre -2,0 e 0,0; têm mais chance de alcançarem a categoria 1.

Para habilidades entre 0,0 e 2,0, a maior probabilidade é que os indivíduos respondam até a categoria 2. Finalmente, indivíduos com habilidade acima de 2,0; devem alcançar a última categoria de resposta (que deverá representar o acerto total).

Note que do item 1 para o 2, a distância entre $b_{i;2}$ e $b_{i;3}$ tornou-se menor. A consequência disto é que aumenta a faixa de habilidade em que os indivíduos deverão responder somente até a categoria 1: de -2,0 a 0,0 no item 1 para -2,0 a 0,5 no item 2. Em outras palavras, a categoria 2 ficou mais difícil de ser alcançada, uma vez que no item 1 indivíduos com habilidades entre 0,0 e 2,0 têm maior probabilidade de conseguir responder a essa categoria do que indivíduos com habilidades entre 0,5 e 2,0 no item 2.

No item 3, praticamente não há chance de os indivíduos responderem até a categoria 2: indivíduos com habilidade entre -2,0 e 0,0 têm mais chance de conseguir responder somente à categoria 1, enquanto indivíduos com habilidade maior do que esse valor já têm maior probabilidade de atingir a última categoria do item.

Finalmente, o item 4 é um exemplo de item onde a maioria dos indivíduos ou responde somente à primeira categoria, ou consegue chegar até a última. Apenas indivíduos com habilidades entre -2,0 e 0,0 apresentam uma chance maior de responderem somente às categorias 1 e 2.

A seguir, analisaremos os resultados de uma pesquisa que, entre outros itens, apresentava aquele que questionava a quantidade de banheiros que a família possuía (as possibilidades de respostas foram as seguintes: nenhum, tem 1, tem 2, tem 3, tem 4 ou mais). Para esse tipo de item, a probabilidade de escolha de cada resposta pode ser modelada através do modelo de respostas graduadas.

Foi obtido o seguinte modelo:

(a_i = 1.511, b_i =0.592, $b_{i,0} = \infty, b_{i,1}$ = 3.81,$b_{i,2}$ = 0.191, $b_{i,3}$ = 1.183, $b_{i,4}$ =1.807), a probabilidade de cada resposta está representada no gráfico seguinte:

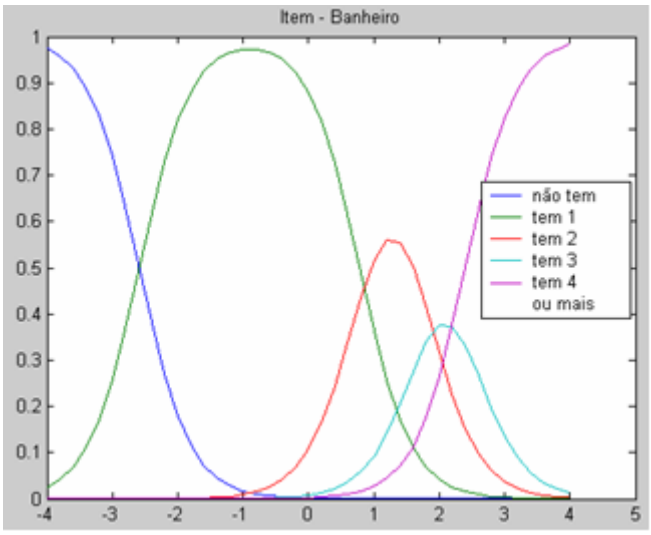

Figura 3.3: Modelo de Respostas Graduadas para o item banheiro.

Podemos extrair, com base na figura 3.3, informações valiosas sobre o item do questionário possuir (um ou mais banheiros) ou não, como:

1. Pessoas com score estimado entre -4 e -3, a probabilidade de não possuir banheiro é superior a 60 %; à medida que aumenta o valor do score, também aumenta a chance de o respondente possuir pelo menos um banheiro.
2. Pessoas com score estimado entre -3 e -2 começam a apresentar um aumento na probabilidade de possuir um banheiro;
3. Já entre o score -2 e 0 podemos afirmar sem perda de generalidade que tal score é caracterizado por aqueles que possuem pelo menos um banheiro.

A construção da interpretabilidade para leitura dos scores estimados, para cada item, segue como descrito acima.

Faz-se necessário chamar a atenção que o valor da probabilidade de uma curva qualquer (para um score dado) é 1 menos o valor da probabilidade de outra curva abaixo da primeira. Por exemplo, exatamente no score 3, temos 80% de possuir quatro banheiros mais, aproximadamente 15% de possuir 3 banheiros mais, 5% (aproximadamente) de possuir 2 banheiros.

3.3 Proposta de correção

A proposta do presente trabalho é apresentar um modelo possível, de acordo com a TRI, para a correção da redação do ENEM. Para tanto, a redação será tratada como

um conjunto de itens de resposta aberta que permite aos examinadores parametrizar a capacidade dos estudantes de, com base nas competências e habilidades desenvolvidas ao longo da formação básica, refletir acerca de uma questão social e apresentar para ela uma proposta de intervenção, demonstrando que recorreram a processos mentais mais complexos na reflexão sobre a questão social. Espera-se que os estudantes comuniquem com clareza os seus processos de tomada de decisão no contexto dado.

No entanto, a configuração da correção a partir de aspectos da TCT não permitem essa parametrização, pois se limita a atribuir um escore em um exame específico, não revelando o traço latente do estudante. Nesse sentido, é preciso alterar o modelo de correção adotando uma outra metodologia.

A proposta do modelo de Samejima atende a essa necessidade na medida em que, além de parametrizar a produção dos candidatos, permite a criação de uma série histórica da aplicação dos exames, o que trará múltiplas vantagens, entre elas o melhor gerenciamento dos dados relativos à qualidade do aprendizado, a possibilidade de avaliar qualitativamente a real situação dos estudantes brasileiros quanto à produção escrita e o aperfeiçoamento de políticas públicas de educação.

Para o novo modelo proposto, continua inalterada a correção com base em cinco competências, cada uma delas nos níveis dados na subseção 3.1 e conforme transcrição a seguir da competência 1.

Tabela 3.7: competência 1

PONTUAÇÃO	NÍVEL	DESEMPENHO
200	5	Demonstra excelente domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizem reincidência.
160	4	Demonstra bom domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com poucos desvios gramaticais e de convenções da escrita.
120	3	Demonstra domínio mediano da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.
80	2	Demonstra domínio insuficiente da modalidade escrita formal da Língua Portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.
40	1	Demonstra domínio precário da modalidade escrita formal da Língua Portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.
0	0	Demonstra desconhecimento da modalidade escrita formal da Língua Portuguesa.

Porém, o corretor não atribuirá ao candidato a pontuação entre 0 a 200, como é no modelo atual, mas sim uma escala do tipo Likert que contemple os seis níveis de desempenho atingido pelos estudantes, de acordo com a tabela descrita a seguir.

Tabela 3.8: competência 1

escala	NÍVEL	DESEMPENHO
Concordo plenamente	5	Demonstra excelente domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizem reincidência.
Concordo parcialmente	4	Demonstra bom domínio da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com poucos desvios gramaticais e de convenções da escrita.
Concordo	3	Demonstra domínio mediano da modalidade escrita formal da Língua Portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.
Discordo	2	Demonstra domínio insuficiente da modalidade escrita formal da Língua Portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.
Discordo parcialmente	1	Demonstra domínio precário da modalidade escrita formal da Língua Portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.
Discordo fortemente	0	Demonstra desconhecimento da modalidade escrita formal da Língua Portuguesa.

Supõe-se que as categorias de um item i encontram-se distribuídos em ordem crescente e denotados por $k = 0; 1; [...]; m_i$, em que (m_i+ 1) compreende o número de categorias do i-ésimo item. Para a redação do ENEM, serão cinco itens, cada um representando uma das competências já expostas.

Logo, o valor i varia de um a cinco. Cada competência terá seis categorias para marcação, de acordo com a tabela acima, sendo assim o valor de k varia de zero a cinco.

A probabilidade do respondente j apontar a categoria de resposta mais alta do item i é dada por:

$$P_{i,k}^{+}(\theta_j) = \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k})}}$$

em que:

- i=1,2,3,4,5. (quantidade de competências);
- j=1,2,3,... (quantidade de estudantes avaliados);

- k=0,1,2,3,4,5;
- $P^{+}_{i,k}(\theta_j)$ é a probabilidade de um indivíduo j escolher a categoria k ou outra mais alta do item i;
- $b_{i,k}$: é o parâmetro de dificuldade da k-ésima categoria do item i;
- θ_j: representa a habilidade, proficiência (traço latente) do j-ésimo indivíduo;
- a_i é o parâmetro de discriminação (ou de inclinação) do item i, com valor proporcional à inclinação da CCI no ponto;
- D: é o fator de escala constante e igual a 1. Utiliza-se o valor 1,7 quando se deseja que a função logística forneça resultado semelhante ao da função ogiva normal.

Segundo Alexandre, Andrade, Vasconcelos, Araujo & Batista (2002), a discriminação de uma categoria específica de resposta é dependente do parâmetro de discriminação a_i, inerente a todas as categorias do item, como da distância das categorias de dificuldade adjacentes.

Assim, necessita-se da existência de uma ordenação do nível de dificuldade das categorias dos itens (modelo politômico), de acordo com a classificação de seus escores, logo $b_{i,1} \leq b_{i,2} \leq ... \leq b_{i,5}$.

A partir do modelo ML2, a probabilidade de um indivíduo j apresentar o escore k no item i, no MRG de Samejima, é determinada pela equação (Bortolotti et al., 2012):

$$P_{i,k}(\theta_j) = P^{+}_{i,k}(\theta_j) - P^{+}_{i,k+1}(\theta_j)$$

Admitindo-se que:

$$P^{+}_{i,0}(\theta_j) = 1$$

$$e$$

$$P^{+}_{i,m_i+1}(\theta_j) = 0$$

Teremos a equação logística do modelo de resposta gradual, conforme a equação:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{-D_{ai}(\theta_j - b_{i,k+1})}}$$

No MRG de Samejima, são estimados dois parâmetros, (a_i e $b_{i,k}$). O parâmetro de discriminação a_i faz referência à inclinação da curva do modelo logístico, compreendendo o poder de segregação de indivíduos por nível de "satisfação" ao aspecto exposto no item.

Logo, quanto maior o valor do parâmetro, maior o poder discriminatório do item nos diferentes níveis de traço latente. Muito embora possa assumir valores orbitando de $-\infty$ a $+\infty$, não são comuns valores negativos (Andrade et al., 2000; Baker, 2001), sendo usualmente trabalhados parâmetros compreendidos no intervalo de 0 a 2 (Hambleton; Swaminathan, Roger, 1991).

O parâmetro de dificuldade $b_{i,k}$ é medido na mesma escala da proficiência θ. Quando b = θ, a probabilidade de um indivíduo escolher uma das categorias do item é igual a 50%.

Capítulo
4

Conclusões

A incorporação da avaliação da escrita nas avaliações educacionais em larga escala revela um avanço nas políticas de avaliação, uma vez que, ao se avaliar a escrita, obtém-se um diagnóstico mais completo da qualidade da educação ofertada pela rede de ensino.

Considerando-se o grande desafio no qual se configura a avaliação do ENEM, a necessidade de aperfeiçoamento constante no modelo de correção e seu potencial *feedback* aos organismos educacionais no tocante à qualidade na verificação do processo de ensino-aprendizagem, o objetivo central desta dissertação é contribuir para este debate, incorporando novos elementos à discussão.

Neste contexto, faz-se necessário um instrumento que contribua com a construção e análise dos resultados obtidos de sorte que se possa ter confiança e garantias que os resultados obtidos foram os mais representativos da verdadeira habilidade dos indivíduos em estudo, refletindo o rendimento deles na avaliação.

Logo, a correção da redação de acordo com o modelo atual pode ser revisto e aperfeiçoado, mudando o foco da análise estatística, desviando-se do teste (que representa o macro) para o item (que representa o micro), sendo que o uso de modelos de resposta gradual, como o de Samejima, para a correção da redação podem ajustar-se de modo razoável e adequadamente aos dados gerados.

Ao incorporar à avaliação da escrita resultados processados pela TRI, permite-se, assim, a construção e a interpretação de uma escala de proficiência da produção escrita dos estudantes no que diz respeito ao desenvolvimento das cinco competências avaliadas.

Naturalmente, outras discussões devem ser aprofundadas, representando um incremento na proposta dada por esta dissertação, como a confiabilidade na pontuação e a severidade do avaliador na construção de uma escala de classificação, o que leva ao estudo

de outros modelos da TRI que complementariam aqueles apresentados, como o modelo multifacetas de Rash.

A análise aqui feita é apenas teórica e também precisaria ser aplicada a um grupo representativo de estudantes, fazendo-se a comparação com a correção feita pelos dois modelos, o oriundo da teoria clássica e o modelo de Samejima.

Enfim, esse é apenas o começo de um caminho que precisa ser trilhado em uma avaliação que já adquiriu grandes proporções no cenário educacional brasileiro.

Referências Bibliográficas

[1] Andrade, D. F.; Tavares, H. R.; Valle, R.C., *Teoria de resposta ao item: conceitos e aplicações. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA*. São Paulo: ABE – Associação Brasileira de Estatística, 2000.

[2] Pasquali,Luiz., *Psicometria: teoria dos testes na psicologia e na educação*. 4.ed. Rio de Janeiro: Vozes, 2011.

[3] Ostini, Remo.;Nering, M.L., *Polytomous Item Response Theory Models. In: Series quantitative applications in the social sciences*, Thousand Oaks, California: SAGE publications, 2006.

[4] Birnbaum, A. *Some latent trait models and their use in inferring an examinee´s ability. In Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick (Eds.). Reading, M. A. Addison-Wesley,1968.

[5] Hambleton, R. K.; SwaminathanW, H.; Rogers, H J. *Fundamentals of Item Response Theory*, Sage Publications: London, 1991.

[6] Lord, F. M.; Novick, M. R, *Statistical Theories of Mental Test Scores*, Addison-Wesley Publishing Company: New Jersey, 1967.

[7] MEC. *Lei de Diretrizes e Bases. Lei n. 9.394*, de 20/12/96.

[8] Muñiz, J. *Introducción a la Teoría de Respuesta a los Items*, Madrid: Pirámide, 1997.

[9] Perrenoud, P. *Avaliação: da excelência à regulação das aprendizagens–Entre Duas Lógicas*, Porto Alegre: Artes Médicas Sul, 1999.

[10] Luckesi, C. C.*Avaliação da aprendizagem escolar: estudos e proposições*, 22.ed. São Paulo:Cortez, 2011.

[11] Rabelo, M. L. *Avaliação Educacional: fundamentos, metodologia e aplicações no contexto brasileiro*, Rio de Janeiro: SBM, 2013.

[12] Andrade, G. G.;Rabelo, M. L.(org) *A produção de textos no ENEM: desafios e conquistas*, Brasília: UnB, 2007.

[13] Oliveira, A. P. M. *Avaliação e regulação da educação: a prova Brasil como política de regulação da rede pública do Distrito Federal*, Brasília: UnB, 2012.

[14] Sandi, F. A.;Chiquito, R. S. *Projeto Marista para Planejamento e Avaliação*, São Paulo: FTD, 2009.

[15] *Diário Oficial da União em 9 de maio de 2013*, (anexo IV página 83).

[16] MEC. *Guia do participante: a redação no ENEM de 2013.*

[17] Mambrini, J. V. M., *Desigualdade em Saúde no Brasil: medida e avaliação*, Belo Horizonte: UFMG/Cedeplar, 2009.

[18] Francisco, R. *APLICAÇÃO DA TEORIA DA RESPOSTA AO ITEM (T.R.I.) NO EXAME NACIONAL DE CURSOS (E.N.C.) DA UNICENTRO.*, Curitiba: UFPR, 2005.

[19] http://www.maxwell.vrac.puc-rio.br/5253/5253$_4$.*PDF*