UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

Metodologia Unificada para Detecção de DIF: uma Aplicação aos Itens do ENEM 2017 para Candidatos com Déficit de Atenção

CHARLES EDUARDO DE ALBUQUERQUE VIEIRA

Orientação: **Profa.Dra.Maria Regina Madruga Tavares**
Coorientação: **Prof.Dr.Héliton Ribeiro Tavares**

Belém
2020

CHARLES EDUARDO DE ALBUQUERQUE VIEIRA

Metodologia Unificada para Detecção de DIF: uma Aplicação aos Itens do ENEM 2017 para Candidatos com Déficit de Atenção

Dissertação apresentada ao Curso de Mestrado em Matemática e Estatística da Universidade Federal do Pará, como pré-requisito para a obtenção do título de Mestre em Estatística.

Orientação: **Profa.Dra.Maria Regina Madruga Tavares**

Coorientação: **Prof.Dr.Héliton Ribeiro Tavares**

Belém

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

V657m Vieira, Charles Eduardo de Albuquerque
Metodologia Unificada para Detecção de DIF: uma Aplicação aos Itens do ENEM 2017 para Candidatos com Déficit de Atenção / Charles Eduardo de Albuquerque Vieira. — 2020.
74 f. : il. color.

Orientador(a): Profª. Dra. Maria Regina Madruga Tavares
Coorientador(a): Prof. Dr. Héliton Ribeiro Tavares
Dissertação (Mestrado) - Programa de Pós-Graduação em Matemática e Estatística, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2020.

1. DIF. I. Título.

CDD 318.115

CHARLES EDUARDO DE ALBUQUERQUE VIEIRA

Metodologia Unificada para Detecção de DIF: uma Aplicação aos Itens do ENEM 2017 para Candidatos com Déficit de Atenção

Esta Dissertação foi julgada e aprovada para a obtenção do grau de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará.

Belém, 10 de Fevereiro de 2020

João Marcelo B Protázio

Prof. Dr. João Marcelo Brazão Protázio
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística – UFPA)

Banca Examinadora

Maria Regina Madruga Tavares

Profa.Dra. Maria Regina Madruga Tavares
PPGME/UFPA
Orientadora

Prof. Dr. Héliton Ribeiro Tavares
PPGME/UFPA
Coorientador

Marinalva C. Maciel

Profª. Dra. Marinalva Cardoso Maciel
FACULDADE DE ESTATÍSTICA/UFPA
Examinadora Externa

Aos meus familiares.

Agradecimentos

À Deus, porque mesmo diante das minhas imperfeições nunca desistiu de mim.

À minha orientadora e ao meu coorientador, profa. Maria Regina Madruga Tavares e prof. Héliton Ribeiro Tavares, pelo auxílio necessário em correções, incentivos, paciência e perfeccionismo, repassando parte da suas valorosas experiências para a construção desse trabalho.

À minha esposa, Luciane Cristina Farias de Aguiar, por acompanhar minhas alegrias e tristezas nessa jornada.

À minha mãe, Maria de Albuquerque Vieira, por me formar na minha personalidade desde pequeno.

À minha filha, Sophia Rosa de Aguiar Vieira por me dá uma enorme felicidade com sua alegria.

Às minhas chefias na PROPLAN/UFPA, Raquel Trindade Borges, Jaciane do Carmo Ribeiro e Maria da Conceição G. Ferreira, por ter me concedido licença qualificação nesses dois anos, sem isso seria improvável conseguir terminar o mestrado, pois estava com alguns problemas de saúde e as disciplinas do mestrado exigem muito tempo.

Aos meus Amigos do PPGME, em especial Thamara Rúbia, Alice Nabiça, Miguel Monteiro, Wilson Rodrigues e Beatriz Cristina, pelo apoio, e à toda minha turma de 2018.

Por fim, quero agradecer à UFPA, como aluno e servidor, porque minhas principais realizações estão ligadas a esta Universidade.

A justiça é a bondade medida ao milímetro.

Emma Andievska

Resumo

Os primeiros estudos sobre funcionamento diferencial de itens (da sigla em inglês, *DIF - Differential Item Functioning*) começaram na década de 60, com o objetivo de resolver problemas de viés em testes. O DIF é uma característica estatística de um item, e ocorre quando dois ou mais grupos de mesma habilidade têm probabilidades diferentes de acertar este item. Vários estudos tratam desse tema, propondo métodos de identificação de itens com DIF e, assim, contribuindo para minimizar injustiças em testes (avaliações) de desempenho.

Neste trabalho são estudados alguns desses métodos para detecção de DIF, e propõe-se uma metodologia unificada para aplicação dos mesmos. A metodologia proposta é aplicada aos dados do ENEM 2017 considerando os itens de duas áreas, Linguagens, Códigos e suas Tecnologias e Matemática e suas Tecnologias, e o desempenho de dois grupos de candidatos: o grupo focal (candidatos que declararam ter déficit de atenção) e o grupo referência (candidatos que declararam resposta negativa em todos os indicadores de atendimento especializado). Nesse estudo foram considerados 1.897 indivíduos no grupo focal e uma amostra de 10.000 indivíduos no grupo referência.

PALAVRAS-CHAVE: funcionamento diferencial do item, metodologia unificada, pacote difR

Abstract

The first studies of Differential Item Functioning (DIF) began in the 1990s. 60, in order to solve test bias problems. DIF is a statistical characteristic of an item, and occurs when two or more groups of the same ability have a different chance of hitting this item. Several studies deal with this theme, proposing methods for identifying items with DIF and thus helping to minimize unfairness in testing performance evaluations.

In this work some of these methods for detection of DIF, and a unified methodology for their application is proposed. The proposed methodology is applied to ENEM 2017 data considering the two-area items, Languages, Codes, and their Technologies and Mathematics and its Technologies, and the performance of two candidate groups: the focal (candidates who reported having attention deficit) and the reference group (candidates who reported a negative response in all indicators specialized service). In this study, 1,897 individuals were considered in the focal group and a sample of 10,000 individuals in the reference group.

Keywords: differential item operation, unified methodology, difR package

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Avaliações Educacionais em Grande Escala

Nesse trabalho o foco são as avaliações em larga escala, que tem por objetivo avaliar o desempenho de candidatos (estudantes), ou seja, estimar habilidades ou proficiências em uma ou mais área do conhecimento. Nessas avaliações há uma grande participação, da ordem de milhões de candidatos, abrangendo grupos com múltiplas realidades do ponto de vista geográfico, racial, com necessidades especiais de natureza física e/ou emocional, entre outras diferentes características.

Os itens (questões) que compõem essas avaliações devem estar estritamente relacionados ao(s) conhecimento(s) que se deseja mensurar, e não podem estar correlacionados com essas diferentes realidades dos candidatos. Assim, espera-se que candidatos de diferentes grupos, mas com mesmas habilidades, tenham iguais probabilidades de acerto aos itens. Ou seja, os itens devem funcionar da mesma forma para todos os grupos de candidatos, exigindo apenas a habilidade do candidato para uma probabilidade correta de resposta. É de suma importância que não ocorra DIF (Funcionamento diferencial do item) nesses itens, pois os resultados dessas avaliações são, geralmente, utilizados para classificar candidatos em processos seletivos, e torna-se imperioso que não ocorram injustiças.

Alguns métodos para detecção de DIF serão utilizados nesse trabalho para avaliar os itens das provas de Linguagens, Códigos e suas Tecnologias, e Matemática e suas Tecnologias, do Exame Nacional do Ensino Médio (ENEM) do ano de 2017, considerando dois grupos de candidatos: os que declararam no questionário do ENEM ter déficit de atenção (grupo focal) e os que declararam resposta negativa em todos os indicadores de atendimento especializado do questionário (grupo controle). É proposta uma metodologia unificada para o uso dos principais indicadores (métodos) para detecção de DIF.

As principais avaliações educacionais no Brasil, a nível federal, são organizadas e aplicadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão ligado ao Ministério da Educação(MEC) e criado em 1937. O INEP tem por missão auxiliar na elaboração de políticas educacionais nas diferentes camadas do governo, colaborando com o desenvolvimento econômico e social do Brasil.

É uma autarquia federal que possui ações nos dois níveis da educação brasileira, Educação Básica e Ensino Superior e, também, internacionalmente. No nível da educação básica organiza o Exame Nacional do Ensino Médio (ENEM), criado em 1998 com o objetivo de avaliar o rendimento escolar no final da educação básica. Participam do ENEM discentes das escolas públicas e particulares do país.

Ainda na educação básica, o INEP gerencia o Sistema Nacional de Avaliação da Educação Básica (SAEB), que teve a primeira edição em 1990 e tem como finalidade analisar a qualidade, equidade e a eficiência da educação no país, produzindo informações e indicadores que auxiliam no acompanhamento de políticas públicas na área da educação. O SAEB gera vários produtos, como o cálculo do índice de Desenvolvimento da Educação Básica (Ideb), tendo como componentes o desempenho dos alunos no SAEB e as informações do fluxo escolar do Censo Escolar.

A nível de Estados, algumas secretarias estaduais de educação realizam as próprias avaliações, com o objetivo de acompanhar o rendimento de seus alunos e definir políticas educacionais voltadas a uma melhor gestão na educação. No Estado de São Paulo há o Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (SARESP), que é um exame realizado todo ano pela Secretaria da Educação do Estado de São Paulo (SEE/SP) para detalhar o perfil educacional. Participam da avaliação estudantes matriculados no 3º, 5º, 7º e 9º anos do ensino fundamental e 3º ano do ensino médio da rede pública e particular, esta, geralmente, por meio de convênio.

Essas avaliações educacionais tem, como principal objetivo, medir o conhecimento/habilidade dos estudantes em alguma área do conhecimento, que é uma variável latente (não pode ser observada diretamente). Do ponto de vista estatístico, essa medida de conhecimento pode ser estimada com a Teoria Clássica dos Testes (TCT) ou com os modelos da Teoria da Resposta ao Item (TRI). Em ambas abordagens, considera-se um conjunto de itens

(questões) na composição do teste, que estão diretamente relacionados à área do conhecimento que se deseja avaliar.

Na TCT estima-se a medida de conhecimento como o total de acertos do estudante (escore total), que dependerá do particular conjunto de itens usados no teste. Assim, as análises estarão relacionadas ao teste, e não será possível comparar diretamente as medidas de conhecimento entre estudantes que não foram submetidos ao mesmo teste.

Nos modelos da TRI é estabelecida uma relação funcional entre a probabilidade do estudante dar uma certa resposta a um item e seu conhecimento (variável ou traço latente). Na TRI os elementos centrais são os itens, e adota-se uma métrica para a estimação da medida de conhecimento. Portanto, é possível comparar o desempenho de estudantes de diferentes populações que realizaram testes distintos, mas com alguns itens comuns; ou, ainda, comparar estudantes de uma mesma população, submetidos a testes totalmente distintos (ANDRADE; TAVARES; VALLE, 2000). Essa é uma grande vantagem da TRI sobre a TCT.

A TRI é usada pelo INEP em várias avaliações, como o ENEM e o SAEB, e é utilizada em vários países nas avaliações educacionais: no *National Assessment of Educational Progress* (NAEP), organizado nos Estados Unidos pelo *Educational Testing Service* (ETS), no *GAOKAO*, um tipo de vestibular para acesso ao ensino superior realizado na China, no *Programme for International Student Assessment* (PISA), coordenado pela Organização para Cooperação e Desenvolvimento Econômico (OCDE) e aplicado em 72 países (no Brasil, a coordenação local é do INEP), entre outras.

Neste trabalho, o foco será propor uma metodologia unificada para detecção de DIF, e aplicar aos dados do ENEM 2017, identificando se há itens com DIF ao compararmos o desempenho de estudantes com déficit de atenção e estudantes em um grupo controle. Na próxima seção serão descritas brevemente algumas características associadas ao protocolo de elaboração dos itens dessa importante avaliação.

1.2 Elaboração dos Itens no ENEM

O ENEM é realizado anualmente em 2 domingos, geralmente no mês de novembro, com a participação de milhões de candidatos. O exame contempla

180 questões em 4 áreas de conhecimento: Linguagens, Códigos e suas Tecnologias; Ciências Humanas e suas Tecnologias; Ciências da Natureza e suas Tecnologias; e Matemática e suas Tecnologias.

O processo de organização dos espaços para elaboração e armazenamento das provas é sigiloso, realizado em um ambiente físico integrado e seguro, um espaço de segurança máxima em Brasília. A proteção inclui identificação biométrica, em que poucos servidores do INEP e colaboradores tem acesso.

As questões são elaboradas por técnicos capacitados e qualificados na respectiva área do conhecimento. Após a elaboração os itens são revisados, e um técnico deverá assegurar a qualidade técnico-pedagógico do item; se o item não for considerado bom pelo técnico ele é devolvido para o elaborador, iniciando novamente o ciclo.

Posteriormente, o item passará por outra revisão, chamada de forma dos padrões. Com o aval do elaborador o item será pré-testado em um estudo piloto feito por aproximadamente 400 estudantes para análise inicial.

Com o arcabouço de todos esses testes, por meio de critérios nacionais são estimados os parâmetros dos vários itens (suas características) pré-testados, formando assim um banco de itens para as 4 áreas. Esse banco possui questões elaboradas no decorrer de anos, verificadas em cerca de dez etapas antes de serem consideradas relevantes para apresentação em alguma edição do ENEM.

1.3 Déficit de Atenção

No cenário escolar, o aluno que apresenta Transtorno do Déficit de Atenção com Hiperatividade (TDAH) é, além de um desafio, um grande aprendizado para os educadores na missão de lecionar. Trata-se, na maioria das vezes, de um aluno inquieto, agitado, e com pouca atenção para as tarefas escolares.

Em geral, o diagnóstico de TDAH (ou, simplesmente, déficit de atenção), depende de 18 perguntas feitas pelo especialista ao educador sobre o comportamento do aluno. As 9 primeiras identificam a desatenção, e as outras a impulsividade e a hiperatividade. Para ser diagnosticado com uma dessas características, ou as duas, o educador deve responder positivamente a pelo menos 6 das 9 perguntas de cada bloco, de acordo com Associação Brasileira do Déficit de Atenção (ABDA).

As perguntas a serem respondidas sobre o comportamento do aluno são apresentadas abaixo:
1. Não consegue se concentrar muito nos detalhes ou comete erros por distração nos trabalhos da escola ou tarefas;
2. Fica confuso ao realizar tarefas ou atividades de lazer;
3. Parece não estar escutando quando se conversa apenas com ele;
4. Em relação às tarefas, à escola, e às obrigações segue comandos parcialmente e não consegue concluir;
5. Tem resistência ao organizar tarefas e atividades;
6. Participa sem desejar, evita ou não gosta de tarefas que requerem esforço mental extenso;
7. Perde recursos essenciais para atividades (p. ex: brinquedos, deveres da escola, lápis ou livros);
8. Impulsos externos causam distração;
9. Esquece de atividades do dia-a-dia;
10. Sacode os pés, mãos e se move na cadeira;
11. Não consegue ficar sentado na sala de aula quando é necessário;
12. É incômodo ao correr de um lado para o outro e subir excessivamente em objetos;
13. Não tem tranquilidade ao brincar ou ao realizar atividades de lazer;
14. Não consegue ficar quieto e fica com frequência a "mil por hora";
15. Fala muito;
16. Responde perguntas de forma apressada antes de terem sido finalizadas;
17. Dificilmente aguarda a vez;
18. Intervém ou interrompe conversas (p.ex. mete-se nas conversas/jogos).

Um diagnóstico rápido é de suma importância, pois o TDAH acomete de 5% a 10% das crianças e adolescentes e pode prejudicar o rendimento escolar, especialmente em relação à leitura, à escrita e à matemática. O desenvolvimento de práticas pedagógicas que envolvam um acompanhamento multidisciplinar ajuda a diminuir esses prejuízos, e torna o ambiente escolar mais atraente e confortável para essas crianças.

1.4 Justificativa e Importância da Dissertação

A presença de viés em pesquisas, seja de natureza qualitativa ou quantitativa, pode provocar erros que modificam as estatísticas e análises feitas para a

seleção e classificação (ZUMBO, 1999). Isto pode comprometer processos seletivos, que levam em consideração o rendimento do estudante numa prova, gerando, consequentemente, políticas públicas educacionais equivocadas. Este trabalho tem relevância nas áreas social, estatística, educacional, psicométrica, saúde, etc.

Os itens (ou questões) que compõem um instrumento de medida devem estar correlacionados apenas ao construto (ou traço latente) de interesse, sem apresentar qualquer tipo de viés ou funcionamento diferencial. Por exemplo, a resposta a itens de matemática devem exigir apenas o conhecimento em matemática, e não outro traço secundário (BUZICK; STONE, 2011).

Com objetivo de promover a igualdade e diminuir o viés no exame para determinados grupos, o INEP permite um atendimento especializado com uma hora a mais de prova. Entre esses grupos está os candidatos com déficit de atenção.

Para além do atendimento especializado, é preciso criar conjecturas para examinar possíveis dificuldades destes candidatos. De acordo com Camilli e Shepard (1994) é possível haver vício ao obter as estimativas das habilidades (notas) e dos parâmetros dos itens. E torna-se imperativo o estudo e uso de técnicas que detectem a presença desses vieses, pois o ENEM é um dos maiores processos seletivos do mundo, com a nota é possível ingressar no ensino superior e participar de programas de financiamento.

Além das várias instituições de ensino superior que fazem a seleção com a nota do ENEM, destacam-se também os programas :

• PROUNI (Programa Universidade para Todos): é um projeto de incorporação de estudantes do país, elaborado em 2004 pelo Ministério da Educação (MEC) tem como intenção beneficiar estudantes brasileiros que não possuem recursos para custear o preço de uma universidade particular a ter acesso ao ensino superior mediante a concessão de descontos que pode chegar até 100% do valor das mensalidades;

• FIES (Fundo de Financiamento Estudantil): é um programa do governo federal criado em 1999, tem como objetivo principal facilitar o ingresso ao ensino superior em instituições de ensino privada, por meio do financiamento do curso de graduação;

• SISU (Sistema de Seleção Unificada): sistema criado em 2009 com a participação de diversas instituições de ensino, que disponibilizam vagas (por

meio de um sistema eletrônico) para estudantes que fizeram a prova do ENEM;

Nesse contexto, é importante o desenvolvimento de uma ferramenta que possibilite detectar a ocorrência ou não de DIF nos itens das avaliações.

1.5 Objetivos

1.5.1 Objetivo Geral

Propor e implementar uma metodologia unificada para utilizar os principais métodos de detecção de DIF, avaliando a metodologia proposta nos dados do ENEM 2017.

1.5.2 Objetivos Específicos

- Estudo e levantamento bibliográfico dos principais métodos para detecção de Funcionamento Diferencial do Item (DIF), verificando o estado do conhecimento sobre o assunto abordado;
- Organizar e definir critérios para a composição dos dois grupos de candidatos: grupo focal e grupo controle (de referência);
- Propor uma metodologia unificada, tendo como base alguns métodos já consolidados na literatura para detecção de DIF;
- Identificar e caracterizar a ocorrência do DIF na edição de 2017 do ENEM, nas áreas de Linguagens, Códigos e suas Tecnologias (LC), e Matemática e suas Tecnologias (MT), em relação aos candidatos dos dois grupos estudados.

1.6 Organização da Dissertação

Este trabalho encontra-se dividido em 6 capítulos, a saber:

- O Capítulo 1 aborda as avaliações em grande escala, elaboração dos itens, justificativa e objetivos do trabalho.
- O Capítulo 2 aborda alguns modelos da Teoria da Resposta ao Item (TRI), modelos logísticos de 1, 2 e 3 parâmetros, aspectos gerais, e o funcionamento diferencial do item.

- O Capítulo 3 aborda os principais métodos para detecção de DIF, com detalhes sobre classificações e métodos mais atuais.
- O Capítulo 4 propõe uma metodologia unificada em que são usados vários métodos de detecção de DIF.
- O Capítulo 5 apresenta os resultados da aplicação da metodologia proposta aos dados do ENEM 2017.
- O Capítulo 6 apresenta as principais conclusões e considerações obtidas no decorrer do trabalho.

Capítulo 2

Modelos da Teoria da Resposta ao Item (TRI)

2.1 Aspectos Gerais

Em problemas envolvendo a obtenção de medidas educacionais e/ou psicológicas, a variável estudada é de conhecimento empírico, por isso esses fatores são chamados de não observáveis ou traços latentes (ou construtos). Os diferentes modelos da TRI são indicados para a estimação dessas medidas, e podem ser divididos em relação ao tipo de item: dicotômicos (do tipo certo ou errado) e não dicotômicos (resposta com várias categorias).

Entre os modelos para itens não dicotômicos há aqueles para itens com categorias de respostas nominais e com categorias graduais. Em relação aos itens dicotômicos, os modelos mais conhecidos são os modelos logísticos, que se diferenciam pela quantidade de parâmetros dos itens, 1, 2 ou 3 parâmetros.

O modelo logístico de 3 parâmetros será apresentado na próxima seção, pois é o caso mais geral. A partir dele tem-se como casos particulares os modelos logísticos de 1 e 2 parâmetros.

2.2 Modelo Logístico de 3 Parâmetros (ML3)

Esse modelo é utilizado em várias avaliações de larga escala, em particular no ENEM, e tem como característica principal a presença de 3 características para cada item: dificuldade, poder de discriminação e probabilidade de acerto casual.

A Função de Resposta do Item (FRI) para o ML3, considerando um exame com I itens e n respondentes, é dada por

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad i = 1, \cdots, I; \ j = 1, \cdots, n. \tag{2.1}$$

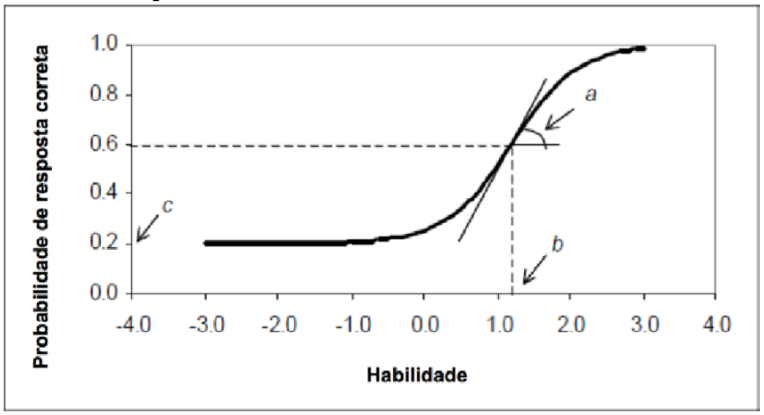
Em que

- U_{ij} é uma variável dicotômica, que assume o valor 1 quando o indivíduo j responde corretamente o item i, ou 0, quando o indivíduo j não responde corretamente o item i;
- θ_j é a habilidade (ou traço latente) do indivíduo j;
- b_i é o parâmetro que mede a dificuldade do item i;
- a_i é o parâmetro que mede o poder de discriminação do item i;
- c_i representa a probabilidade de acerto casual do item i;
- D é uma constante com valor igual a 1. Adota-se o valor 1,7 quando se deseja resultados do ajuste logístico semelhantes àqueles da ogiva normal.

Os modelos logísticos de 1 e 2 parâmetros são um caso particular do ML3: no modelo logístico de 2 parâmetros (ML2) toma-se $c_i = 0$, considerando-se a dificuldade e a discriminação para cada item; já no modelo logístico de 1 parâmetro (ML1) considera-se apenas a dificuldade do item ($c_i = 0$ e $a_i = 1$).

O modelo em (2.1) pode ser representado graficamente por meio da chamada *Curva Característica do Item* (CCI). A Figura 2.1 mostra um exemplo de uma CCI no ML3.

Figura 2.1 *Exemplo de uma Curva Característica do Item no ML3.*



Fonte: Extraído de Andrade, Tavares e Valle (2010)

Observa-se que quanto maior a habilidade do respondente maior é a pro-

babilidade de acerto ao item. O parâmetro a mede a inclinação da curva, e itens com maior inclinação tem maior poder para discriminar os respondentes de menor e maior habilidade. O parâmetro b é medido na mesma escala da habilidade, e itens mais "difíceis"(maiores valores de b) são aqueles em que apenas os respondentes de maior habilidade tem alta probabilidade de acerto. Já os itens "fáceis"são aqueles em que todos os respondentes tem alta probabilidade de acertá-los. O parâmetro c representa a probabilidade de acerto para indivíduos de baixa habilidade ("acerto no chute"). Na Figura 2.1 tem-se $c = 0,2$, considerando que o item tem 5 alternativas de resposta.

Para o ajuste do modelo em (2.1) é necessário estimar as habilidades (escores ou notas) para cada um dos n respondentes, caso os parâmetros (a_i, b_i, c_i) dos I itens já sejam conhecidos. Caso sejam desconhecidos, os mesmos também devem ser estimados juntamente com as habilidades. Maiores detalhes sobre os métodos de estimação podem ser vistos em Andrade, Tavares e Valle (2000).

2.3 Funcionamento Diferencial do Item (DIF)

O funcionamento diferencial de um item (do inglês, *Differential Item Functioning* - DIF) ocorre quando grupos de indivíduos de mesma habilidade apresentam desempenhos diferentes neste item. Assim, a detecção de DIF em um item acontece quando combina-se grupos de respondentes de mesma habilidade, e observa-se um grupo com maior (ou menor) probabilidade de acerto ao item. Nestes casos, há alguma característica secundária, diferente do traço latente de interesse, influenciando a resposta dada a este item.

Algumas vezes, essa característica secundária pode estar associada ao perfil sociodemográfico do respondente (gênero, raça, idade, nível sócio-econômico, ocupação etc ...), ou a alguma incapacidade física, cognitiva ou emocional. Em qualquer caso, é importante que este item seja revisado ou até excluído do banco de itens, pois ele pode causar distorções na estimativa das habilidades.

Os estudos para detecção de DIF consideram dois grupos de respondentes, um grupo de referência (controle) e um grupo focal, para o qual se deseja verificar diferenças no desempenho. Geralmente, a presença do DIF leva a diferentes estimativas do parâmetro de dificuldade deste item, para um grupo o item é mais difícil do que para o outro grupo. Podem ocorrer diferenças nas estimativas dos parâmetros de discriminação e de acerto casual também, mas

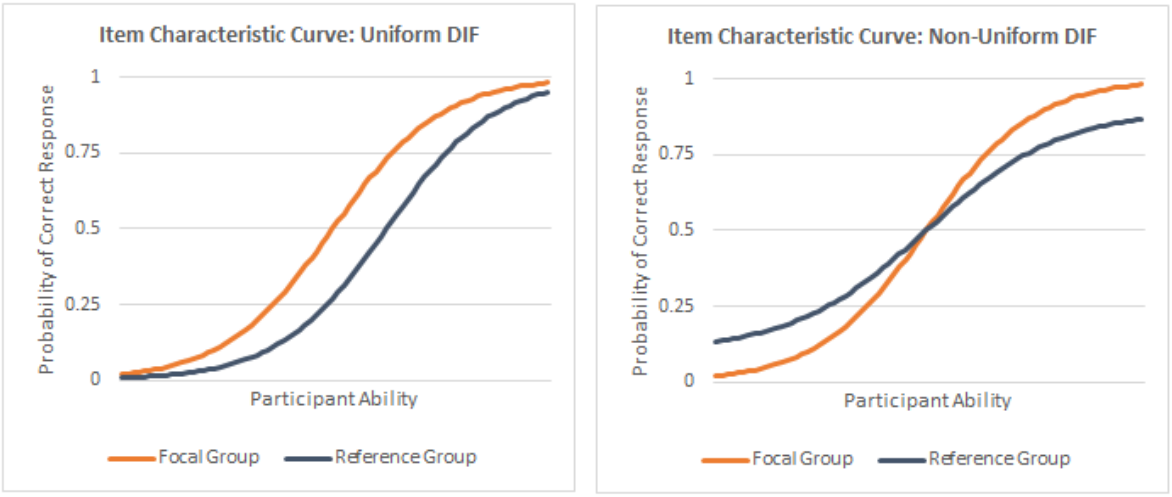
não é muito comum. Uma forma de visualizar a presença de DIF em um item é plotando a curva característica (CCI) para cada grupo.

Quando ocorrem diferenças apenas nas estimativas do parâmetro de dificuldade b, o DIF é classificado como *Uniforme*. Neste caso, as $CCI's$ se apresentam de forma paralela para os dois grupos, ou seja, têm a mesma inclinação (valor do parâmetro a) e mesmo valor do parâmetro c, e as diferenças na probabilidade de acerto ao item se mantém constante em torno do valor de b, com um grupo sempre tendo maior probabilidade de acertar o item, para uma mesma habilidade fixada.

Quando há diferenças nas estimativas dos parâmetros a e b, as $CCI's$ para o dois grupos se cruzam, e o DIF é classificado como *Não-Uniforme*. Neste caso, as diferenças nas probabilidades de acerto mudam ao longo da escala de habilidade, invertendo os grupos com maior probabilidade de acertar o item.

A Figura 2.2 mostra um exemplo de CCI para um item com DIF uniforme e a CCI para um item com DIF não-uniforme, em um modelo de 2 parâmetros.

Figura 2.2 *Curvas características para itens com DIF uniforme e não-uniforme em um modelo logístico de 2 parâmetros.*



Fonte: Extraído de (FOSSEY, 2014), acesso em 17/12/2019

No próximo capítulo serão apresentados os principais métodos e técnicas para detecção de DIF, que serão usados na proposta de metodologia unificada

Capítulo 3

Principais Métodos para Detecção de DIF

3.1 Classificação dos Métodos

Segundo Andriola (2001), o pressuposto de padronização nas condições de aplicação de instrumentos de medida (testes, questionários, etc...) é fundamental para se evitar injustiças com algum grupo de candidatos. Daí surge a necessidade dos estudos envolvendo métodos para identificar a presença de DIF.

Alguns métodos para a detecção de DIF usam a teoria clássica do testes (TCT), ou seja, consideram o escore total do candidato na construção das $CCI^{'}s$ e de algumas medidas, enquanto outros usam o escore obtido via o ajuste de algum modelo da TRI. Segundo Hambleton, Swaminathan e Rogers (1991) (*apud* Andriola (2001)) "a TRI oferece um marco apropriado ao estudo do DIF".

Para se detectar a presença de DIF com base nas $CCI^{'}s$, é preciso medir as diferenças entre as curvas, e avaliar se essas diferenças são estatisticamente significativas. Há vários métodos propostos na literatura da área para a detecção de DIF, que podem ser classificados em duas categorias, segundo Whitmore e Schumacker (1999):

— *Métodos que utilizam um critério interno*: o escore ou a pontuação auferida no teste ou grupo de itens analisados;

— *Métodos que utilizam um critério externo*: um critério externo ao teste ou grupo de itens, por exemplo, o escore em outros testes (CLAUSER; NUNGESTER; SWAMINATHAN, 1996).

Outra classificação, devido a Mellenbergh (1989), Van Der Flier et al. (1984), como citado em Andriola (2001), é dada por:

— *Métodos incondicionais*: Supõe que o grupo de indivíduos e itens apresentam algum tipo de interação;

— *Métodos condicionais*: fundamentado no suposto de que os parâmetros do item são distintos para os indivíduos com a mesma magnitude na variável latente, oriundos de diferentes grupos. Esta suposição está baseado na ideia de que a dificuldade de um item tem dois elementos: um intrínseco (as características do item, tais como, tipo — aberto ou fechado; tamanho do enunciado e das alternativas; sinais utilizados — verbal, numérico, abstrato, etc.) e um extrínseco (as características dos indivíduos, tais como, gênero, nível socioeconômico, raça, idade, etc...). Neste contexto, a dificuldade de um item expressa a interação entre os dois elementos (SCHEUNEMAN; GERRITZ, 1990).

Para Van Der Flier et al. (1984): os métodos condicionais são preferíveis, pois condicionam a probabilidade de resposta certa a um nível de habilidade fixado. Outra classificação de métodos condicionais foi proposta por Millsap e Everson (1993), *Métodos de Invariância Condicional Observada*, que usam o escore total (via TCT) para comparar as diferenças nas $CCI^{'}s$; e *Métodos de Invariância Condicional Não Observada*, que usam a habilidade estimada via TRI (podemos destacar, o Qui-quadrado de Lord, métodos da medida da área entre as $CCI^{'}s$, comparação das probabilidades, comparação dos parâmetros dos itens).

No âmbito da literatura estatística, os métodos condicionais estão baseados no chamado Paradoxo de Simpson (DORANS; HOLLAND, 1992), que evidencia a importância de se comparar os desempenhos dos grupos condicionando no mesmo grau ou magnitude do traço latente.

Andriola (2001) mostra um exemplo do paradoxo de Simpson, citado em Holland e Thayer (1988), que apresenta a proporção de acerto de dois grupos, A e B, a um item hipotético. Neste exemplo observa-se uma maior proporção de acertos ao item no grupo A, sem condicionar na habilidade dos respondentes. Porém, ao se recalcular a proporção de acerto, condicionando em 3 diferentes níveis de habilidade, o grupo B apresenta maior proporção em todos os níveis.

Este paradoxo mostra a importância de se comparar o desempenho dos grupos no item condicionando no mesmo nível de habilidade. Há diversos métodos para analisar o DIF baseados no Paradoxo de Simpson. Aqui apresentaremos métodos condicionais que, de uma forma geral, tem por objetivo

testar a hipótese nula de que o item não apresenta DIF contra a hipótese alternativa de que há presença de DIF no item sob estudo.

3.2 Método das Dificuldades Transformadas dos Itens (TID)

O Método das Dificuldades Transformadas dos Itens (TID), também chamado de *Método Delta de Angoff* (ANGOFF, 1982; ANGOFF; FORD, 1973) detecta a presença de DIF uniforme, com base no desempenho do respondente via escore total. Na TCT a dificuldade de um item é dada pela proporção de acertos, ou seja, os itens fáceis são aqueles que apresentaram alta proporção de acertos no teste, e os difíceis são os que apresentaram baixa proporção de acertos.

Este método consiste em fazer uma transformação linear na medida de dificuldade clássica do item para uma nova escala, chamada de escala Delta. Para cada item i calcula-se a proporção de acertos p_i, que é inicialmente transformada (usando a tabela da distribuição normal padrão) para um desvio normal z_i e, em seguida, transformada linearmente para a escala Delta, $\Delta_i = 4z_i + 13$, com valores no intervalo de 1 a 25 (maiores detalhes em Pasquali (2017)).

Os valores Delta para cada item, Δ_{iR} e Δ_{iF}, são obtidos para o grupo de referência e o grupo focal, respectivamente. Geralmente, o gráfico de dispersão dos pontos $(\Delta_{iR}, \Delta_{iF},\ i = 1, 2, \cdots, I)$, mostrará os pontos em forma de elipse indo do canto inferior esquerdo ao canto superior direito.

Mas, se os 2 grupos são de uma mesma população, o gráfico de dispersão será próximo de uma reta, geralmente apresentando uma correlação alta, em torno de 0,98 ou 0,99. Se os grupos são distintos, com relação ao desempenho (habilidade), os pontos ainda cairão em uma elipse estreita, deslocada vertical ou horizontalmente, dependendo de qual grupo tem melhor desempenho. Pontos que caem distantes da nuvem de pontos podem indicar itens com DIF (possível interação *item* $\times$ *grupo*).

Para resumir aspectos significativos no gráfico, os autores propuseram obter a equação do eixo maior da elipse e calcular as distâncias perpendiculares D_i de cada ponto a essa linha. O desvio padrão dessas distâncias, σ_{D_i}, é uma função da interação *item* $\times$ *grupo*, e a correlação $r_{\Delta_R\Delta_F}$ dos pontos na elipse

representa o grau ao qual os itens tem a mesma ordem de dificuldade nos dois grupos (representação inversa da interação *item* × *grupo*).

A equação para o eixo principal da elipse é dada por $Y = AX + B$, em que

$$A = \frac{\sigma_R^2 - \sigma_F^2\sqrt{(\sigma_R^2 - \sigma_F^2)^2 + 4r_{\Delta_R\Delta_F}^2\sigma_F^2\sigma_R^2}}{2r_{\Delta_R\Delta_F}\sigma_F\sigma_R} \tag{3.1}$$

e

$$B = M_R - AM_F \tag{3.2}$$

com σ_G^2 e M_G representando a variância e a média dos valores Δ no grupo G $(G = R, F)$.

3.3 Método Qui-quadrado de Lord

O método Qui-quadrado de Lord (LORD, 1980) usa modelos da TRI para a medida de habilidade, e permite detectar itens com DIF uniforme e não-uniforme.

Segundo Andriola (2002), esse método propõe a comparação simultânea dos parâmetros de discriminação (a) e de dificuldade (b) nos dois grupos, por meio de uma estatística qui-quadrado com dois graus de liberdade, dada por

$$\chi^2 = X\Sigma^{-1}X' \tag{3.3}$$

Em que

• $X = (a_R - a_F \quad b_R - b_F)$ é o vetor 1×2 das diferenças entre as estimativas dos parâmetros a e b dos grupos referência e focal;

• $\Sigma^{-1} = (\Sigma_R + \Sigma_F)^{-1}$ é a inversa da soma das matrizes 2×2 de variâncias-covariâncias de X;

• $\Sigma_R = \begin{pmatrix} Var(a_R) & Cov(a_R, b_R) \\ Cov(a_R, b_R) & Var(b_R) \end{pmatrix}$ e $\Sigma_F = \begin{pmatrix} Var(a_F) & Cov(a_F, b_F) \\ Cov(a_F, b_F) & Var(b_F) \end{pmatrix}$

• X' é o vetor transposto de X.

Se o modelo logístico de 1 parâmetro da TRI for usado, então apenas o parâmetro de dificuldade será comprado entre os grupos. Neste caso, a estatística do χ^2 de Lord se resume a

$$\chi^2 = \frac{b_F - b_R}{Var(b_F) - Var(b_R)} \tag{3.4}$$

A significância da estatística χ^2, quando comparada ao valor crítico, levará à decisão de rejeitar ou não a hipótese nula de presença de DIF.

Uma extensão da estatística χ^2 em (3.3) para a comparação de mais de dois grupos foi proposta por Kim, Cohen e Park (1995), chamado de *Método* χ^2 *de Lord Generalizado*.

3.4 Método Padronizado

Esse método propõe uma medida do índice de discrepância entre os desempenhos dos dois grupos no item. Segundo Andriola (2002), este índice é muito utilizado pelo *Educational Testing Service* (ETS), e considera a TCT para a medida de desempenho. O índice é obtido por

$$STD = \frac{\sum_k L_k(p_{Fk} - p_{Rk})}{\sum L_k} \tag{3.5}$$

Em que L_k são os pesos obtidos para os grupos analisados. De acordo com Holland e Thayer (1988) (*apud* Andriola (2002)), alguns valores possíveis para L_k são:

- $L_k = N_{tk}$ é o número total de indivíduos no nível k da habilidade;
- $L_k = N_{Rk}$ é o número total de indivíduos do grupo de referência, no nível k da habilidade;
- $L_k = N_{Fk}$ é o número total de indivíduos do grupo focal, no nível k da habilidade;
- $L_k =$ a frequência relativa de indivíduos de qualquer um dos grupos, com o nível k de habilidade;
- p_{Rk} e p_{Fk} são as proporções de indivíduos que acertaram o item nos grupos de referência e focal, respectivamente.

Este índice varia de -1 a 1, um valor positivo indica que o item é mais fácil para o grupo focal, e quando negativo, é mais fácil para o grupo de referência. Para valores pequenos, de –0,05 a +0,05, o DIF é desprezível; nos valores entre –0,06 a -0,10 e +0,06 e +0,10, o DIF é moderado; e para valores acima de +0,10 e menores que -0,10, o DIF é severo (HOLLAND; THAYER, 1988) *apud* (ANDRIOLA, 2002).

3.5 Método da Área de Raju

O método proposto por Raju tem por objetivo estimar a área entre as $CCI's$ nos dois grupos. Raju (1990) desenvolveu expressões para a área exata entre as duas curvas, nos modelos logísticos de 1, 2 e 3 parâmetros (neste caso, considerando o mesmo valor para o parâmetro c). As expressões foram obtidas para a área com sinal (considerando a diferença) e para a área sem sinal (considerando o valor absoluto da diferença).

Porém, como as expressões dependiam das estimativas dos parâmetros dos itens, as áreas obtidas estavam sujeitas a flutuações aleatórias. E como essa variabilidade não é controlada, não era possível saber se as $CCI's$ diferiam apenas por erro amostral.

Assim, Raju (1990) propôs uma modificação, e apresentou o método para estimação da área com base na distribuição amostral assintótica (médias e variâncias) das áreas estimadas, com e sem sinal, para os 3 modelos logísticos da TRI, considerando o parâmetro c fixado.

Sejam $F_1(\theta)$ e $F_2(\theta)$ as funções de resposta do item (FRI) para um mesmo item, nos grupos 1 e 2, respectivamente. Considerando o ML3, temos

$$F_1 = F_1(\theta) = c_1 + (1 - c_1)P_1$$

$$F_2 = F_2(\theta) = c_2 + (1 - c_2)P_2$$

Em que

$P_1 = \frac{\exp(Da_1(\theta - b_1))}{1+\exp(Da_1(\theta - b_1))}$

$P_2 = \frac{\exp(Da_2(\theta - b_2))}{1+\exp(Da_2(\theta - b_2))}$

Segundo Raju (1990), com as $FRI's$ estimadas a partir das estimativas dos parâmetros dos itens e das habilidades, as áreas com e sem sinal entre as curvas são dadas, respectivamente, por

$$\text{Área com sinal} = SA_{kl} = \int_{-\infty}^{+\infty} (\hat{F}_1 - \hat{F}_2)\, d\theta \tag{3.6}$$

e

$$\text{Área sem sinal} = UA_{kl} = \int_{-\infty}^{+\infty} |\hat{F}_1 - \hat{F}_2|\, d\theta \tag{3.7}$$

Em que k representa o número de parâmetros do modelo logístico da TRI, e

$$l = \begin{cases} 0 & (a_i \text{ são iguais ou diferentes, índice com sinal}) \\ 1 & (a_i \text{ são iguais, índice sem sinal}) \\ 2 & (a_i \text{ são diferentes, índice sem sinal}) \end{cases}$$

Raju (1990) desenvolveu expressões para as médias e variâncias assintóticas das áreas entre as curvas, e obteve estimativas para as áreas em cada um dos 3 modelos logísticos da TRI. Essas estimativas são apresentadas abaixo:

- *Modelo ML1 (ou Modelo de Rasch)*

$$SA_{10} = \hat{b}_2 - \hat{b}_1 \tag{3.8}$$
$$UA_{11} = |\hat{b}_2 - \hat{b}_1| \tag{3.9}$$

- *Modelo ML2*

$$SA_{20} = \hat{b}_2 - \hat{b}_1 \tag{3.10}$$
$$UA_{21} = |\hat{b}_2 - \hat{b}_1|, \text{quando } \hat{a}_1 = \hat{a}_2 \tag{3.11}$$
$$UA_{22} = \left| \frac{2(\hat{a}_2 - \hat{a}_1)}{D\hat{a}_1\hat{a}_2} \ln \left\{ 1 + \exp \left[\frac{D\hat{a}_1\hat{a}_2(\hat{b}_2 - \hat{b}_1)}{\hat{a}_2 - \hat{a}_1} \right] \right\} - (\hat{b}_2 - \hat{b}_1) \right|, \text{quando } \hat{a}_1 \neq \hat{a}_2 \tag{3.12}$$

- *Modelo ML3*

$$SA_{30} = (1 - c)\, SA_{20} \tag{3.13}$$
$$UA_{31} = (1 - c)\, UA_{21}, \text{quando } \hat{a}_1 = \hat{a}_2 \tag{3.14}$$
$$UA_{32} = (1 - c)\, UA_{22}, \text{quando } \hat{a}_1 \neq \hat{a}_2 \tag{3.15}$$

Para testar a significância das áreas com sinais entre as $CCI's$ assume-se que essas áreas são normalmente distribuídas, cujas expressões das médias e variâncias já foram obtidas (RAJU, 1990), e utiliza-se a estatística de teste

$$Z = \frac{SA - 0}{\sigma(SA)}, \tag{3.16}$$

que tem distribuição normal padrão.

A estatística acima não pode ser usada para testar a significância das áreas

sem sinais, uma vez que a suposição de normalidade não é válida nestas áreas. Mas, a normalidade pode ser assumida para a variável

$$H = \frac{2(\hat{a}_2 - \hat{a}_1)}{D\hat{a}_1\hat{a}_2} \ln \left\{ 1 + \exp \left[\frac{D\hat{a}_1\hat{a}_2(\hat{b}_2 - \hat{b}_1)}{\hat{a}_2 - \hat{a}_1} \right] \right\} - (\hat{b}_2 - \hat{b}_1),$$

e uma vez que a área sem sinal em (3.12) é dada por $|H|$, então tem-se uma distribuição *half-Normal*. Assim, uma solução para o problema da não-normalidade é testar a significância de H e não de $|H|$, com a estatística

$$Z = \frac{H - 0}{\sigma(H)}, \tag{3.17}$$

que tem distribuição normal padrão.

3.6 Método da Regressão Logística

Esse método, proposto por Swaminathan e Rogers (1990), consiste em ajustar um modelo de regressão logística para prever a probabilidade de acerto ao item, considerando como covariáveis a proficiência (θ) e o grupo (G).

O modelo a ser ajustado é dado por

$$P(u = 1|\theta_j) = \frac{exp(z)}{1 + exp(z)}, \tag{3.18}$$

com

$$z = b_o + b_1\theta + b_2G + b_3(\theta G)$$

Segundo Andriola (2002), tem-se

- G é o grupo (referência ou focal) ao qual o indivíduo pertence;
- θ é a habilidade do indivíduo;
- b_0 é o ponto de interseção da reta de regressão com o eixo das abscissas;
- b_1 é a inclinação da reta de regressão;
- b_2 mede a diferença entre o rendimento dos grupos no item em questão;
- b_3 é o indicador paramétrico da possível interação entre θ e G;

Assim, um item possui DIF não uniforme se $b_3 \neq 0$ (com ou não $b_2 = 0$), e um item terá DIF uniforme se $b_2 \neq 0$ e $b_3 = 0$. Essas hipóteses podem ser testadas pelo teste de Wald ou da verossimilhança generalizada.

Uma extensão deste método para o caso de múltiplos grupos focais, chamado de *Método da Regressão Logística Generalizado*, foi proposto por Magis et al. (2011)

3.7 Método SIBTEST

O método SIBTEST, proposto por Shealy e Stout (1993), considera um modelo multidimensional da TRI para detectar viés/DIF simultaneamente para vários itens. Assim, o traço latente é um vetor composto pela habilidade de interesse θ e por ruídos determinantes $\eta_1, \eta_2, \cdots$. Por simplicidade, vamos considerar o caso bidimensional, mas extensões para vários ruídos determinantes são diretas e o procedimento SIBTEST funciona igualmente bem nessas configurações.

Uma amostra aleatória de indivíduos é selecionada de cada grupo, grupo focal (F) e de referência (R), e os mesmos são submetidos a um teste com N itens. Geralmente, suspeita-se que uma parte do teste é viesada contra o grupo focal, e esse subteste é o alvo do estudo de viés/DIF. Para um indivíduo selecionado aleatoriamente, denota-se vetor de respostas dicotômicas por $U = (U_1, \cdots, U_N)$

Para o modelo da TRI bidimensional, cada indivíduo em ambos os grupos tem vetor latente (θ, η), e a FRI do item i é representada por $P_i(\theta, \eta)$. Assume-se que todos os itens dependem de θ, e um ou mais dependem de η; para aqueles que dependem somente de θ, a FRI é dada por $P_i(\theta)$.

Marginalizando a FRI bidimensional $P_i(\theta, \eta)$ para o grupo g (R ou F), com respeito a habilidade θ, tem-se

$$T_{ig}(\theta) = \int_{-\infty}^{+\infty} P_i(\theta, \eta) f_g(\eta|\theta)\, d\eta, \tag{3.19}$$

e o viés no item i contra o grupo focal, na habilidade θ, ocorre se $T_{iF}(\theta) < T_{iR}(\theta)$; caso contrário, o viés é contra o grupo de referência.

Por convenção, suponha que os itens $n+1$ até N são os itens que compõe o *subteste de estudo* para viés/DIF. Os demais itens serão chamados de *subteste válido*. Seja

$$T_{Sg}(\theta) = \sum_{i=n+1}^{N} T_{ig}(\theta) \tag{3.20}$$

a função de resposta do subteste de estudo para o grupo g. Então, o viés do teste (em θ) contra o grupo focal ocorre se $T_{SF}(\theta) < T_{SR}(\theta)$.

Se o viés total (em θ) contra o grupo focal, causado pelo subteste de estudo, é dado por

$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta),$$

pode-se notar que $B(\theta)$ também satisfaz

$$B(\theta) = T_{SR}(\theta) - T_{SF}(\theta) = \sum_{i=1}^{N} T_{iR}(\theta) - \sum_{i=1}^{N} T_{iF}(\theta), \tag{3.21}$$

ou seja, o somatório pode ser sobre todos os itens, uma vez que $T_{iF}(\theta) = T_{iR}(\theta) = P_i(\theta)$ para os itens que compõem o subteste válido.

O SIBTEST propõe testar *viés unidirecional*, no sentido de que se existe viés no teste contra um mesmo grupo para todo θ, então pode-se dizer que o viés é unidirecional contra este grupo. Assim, o viés unidirecional ocorre se $B(\theta) > 0$ para todo θ, ou $B(\theta) < 0$ para todo θ. Um caso particular de viés unidirecional é o DIF uniforme, em que T_{SF} é o T_{SR} deslocado horizontalmente.

Um índice de viés unidirecional global contra o grupo focal é dado por

$$\beta(U) = \int_{\theta} B(\theta) f_F(\theta)\, d\theta, \tag{3.22}$$

Em que $f_F(\theta)$ é a função densidade de probabilidade de θ para o grupo focal. E deseja-se testar

$$H_0 : \beta(U) = 0 \quad \text{contra} \quad H_1 : \beta(U) > 0.$$

A hipótese alternativa é unilateral porque deseja-se testar o viés contra o grupo focal (mas, o SIBTEST disponibiliza uma opção para a alternativa bilateral). A estatística para o teste baseia-se em uma estimativa de $\beta(U)$, normalizada para ter variância unitária, e obtida a partir do escore total.

Seja $Y = \sum_{i=n+1}^{N} U_i$ o escore do subteste de estudo, e $X = \sum_{i=1}^{n} U_i$ o escore do subteste válido. Assim, indivíduos de mesmo escore X tem aproximadamente a mesma habilidade θ, e podem ser diretamente comparáveis na avaliação do viés. Seja $\bar{Y}_{gk}$ a média de Y para todos os indivíduos no grupo g que tiveram escore $X = k$.

As diferenças

$$\bar{Y}_{Rk} - \bar{Y}_{Fk}, \qquad k = 0, 1, \cdots, n, \tag{3.23}$$

fornecem uma medida do viés contra o grupo focal, e eles propõe a seguinte estimativa para $\beta(U)$,

$$\hat{\beta}_U = \sum_{k=0}^{n} \hat{p}_k(\bar{Y}_{Rk} - \bar{Y}_{Fk}), \tag{3.24}$$

Em que $\hat{p}_k$ é a proporção de indivíduos no grupo focal que obtiveram escore $X = k$. Se J_{gk} é o número de indivíduos no grupo g com escore $X = k$, então $\hat{p}_k = \frac{J_{Fk}}{\sum_{j=0}^{n} J_{Fj}}$.

A estatística para o teste de viés/DIF é dada por

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)}, \tag{3.25}$$

com

$$\hat{\sigma}(\hat{\beta}_U) = \left(\sum_{k=0}^{n} \hat{p}_k^2 \left(\frac{1}{J_{Rk}} \hat{\sigma}^2(Y|k, R) + \frac{1}{J_{Fk}} \hat{\sigma}^2(Y|k, F) \right) \right)^{1/2},$$

e $\hat{\sigma}^2(Y|k, g)$ sendo a variância amostral dos escores Y para os indivíduos do grupo g com escore $X = k$. Pode-se mostrar (detalhes em Shealy e Stout (1993)) que esta estatística tem, sob H_0 e supondo distribuições iguais para a habilidade θ nos dois grupos, distribuição Normal padrão aproximada.

Uma extensão do método SIBTEST para detectar DIF não uniforme, denominada de *Crossing SIBTEST* (CSIBTEST), foi proposta por Li e Stout (1996), e Chalmers (2018).

3.8 Método de Mantel-Haenszel

O método de Mantel-Haenszel (MH) é um dos mais utilizados para detecção de DIF em estudos com itens dicotômicos. Foi inicialmente desenvolvido para uso em pesquisas epidemiológicas (MANTEL; HAENSZEL, 1959), e depois aplicado em estudos para detecção de DIF por Holland e Thayer (1988).

Neste método são criados estratos de indivíduos em ambos os grupos (Focal (F) e Referência (R)), com base em m níveis do escore total escolhidos pelo pesquisador. Para cada nível k ($k = 1, \cdots, m$), constrói-se uma tabela de contingência 2×2, contendo as frequências de acertos e erros ao item sob

estudo em cada grupo. Assim, para cada item sob estudo, haverão m tabelas como a apresentada na Tabela 3.1.

Tabela 3.1 *Respostas ao item sob estudo no nível k do escore total, segundo os grupos.*

Grupo	Acertos	Erros	Total
Referência	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

Assim, $O_{Rk} = A_k/B_k$ é a chance (*odds*) de resposta correta ao item sob estudo, no nível k, no grupo de referência, e $O_{Fk} = C_k/D_k$ é a respectiva chance no grupo focal. A razão entre as chances (*odds ratio*), no nível k, é dada por

$$\alpha_k = \frac{O_{Rk}}{O_{Fk}} \tag{3.26}$$

O estimador de Mantel–Haenszel da razão de chances comum para os m estratos é dado por

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} \frac{A_k D_k}{T_k}}{\sum_{k=1}^{m} \frac{B_k C_k}{T_k}} \tag{3.27}$$

Esse estimador pode ser usado para medir DIF, pois $\hat{\alpha}_{MH} = 1$ indicará que o desempenho no item é o mesmo para os dois grupos, ou seja, o item não apresenta DIF. Por outro lado, valores maiores que um indicarão que o grupo de referência tem um desempenho melhor no item.

A estatística χ^2 de Mantel–Haenszel para testar as hipóteses $H_0 : \hat{\alpha}_{MH} = 1$ (não há DIF) contra $H_1 : \hat{\alpha}_{MH} > 1$ (presença de DIF), é dada por

$$\chi^2_{MH} = \frac{\{|\sum_{k=1}^{m} A_k - \sum_{k=1}^{m} E(A_k)| - 0,50\}^2}{\sum_{k=1}^{m} Var(A_k)}, \tag{3.28}$$

Em que

$$E(A_k) = \frac{n_{Rk}\, m_{1k}}{T_k} \quad \text{e} \quad Var(A_k) = \frac{n_{Rk} n_{Fk} m_{1k} m_{0k}}{T_k^2 (T_k - 1)}.$$

A estatística χ^2_{MH} tem aproximadamente distribuição χ^2 com um grau de liberdade.

O método MH descrito acima é indicado para comparar o desempenho de dois grupos em itens com respostas dicotômicas. Uma extensão para vários grupos e

com itens de resposta politômica é denominada de *Método de Mantel–Haenszel Generalizado* (LANDIS; HEYMAN; KOCH, 1978; PENFIELD, 2001).

3.9 Método de Breslow-Day

O teste de Breslow-Day (BD) foi inicialmente proposto por Breslow, Day e Heseltine (1980) como um teste de homogeneidade da razão de chances em uma tabela de contingência 2×2, para um estudo epidemiológico do tipo caso-controle envolvendo pacientes com câncer. Semelhantemente ao método MH, o teste de homogeneidade BD também é mencionado para análise de tabelas 2×2 em m estratos de uma terceira variável (AGRESTI, 1990; HOSMER; LEMESHOW; STURDIVANT, 1989).

Apesar de muito usado nos estudos epidemiológicos, o método BD tem pouca discussão na literatura nos estudos para análise de DIF. Nesse contexto, podemos destacar os trabalhos de Aguirre (2004) e Penfield (2003).

Para a definição da estatística do teste BD consideraremos a notação apresentada na Tabela 3.1, e que os indivíduos dos 2 grupos são submetidos a um teste com I itens. Os estratos são definidos pelos $m = I - 1$ escores totais, aqui não são considerados os escores extremos (0 e I) pois o desempenho nos dois grupos seria o mesmo nesses casos. Assim, tem-se os níveis de escore $k = 1, 2, \cdots, I - 1$.

A estatística para testar a hipótese H_0 (não há DIF) contra a hipótese H_1 (há DIF), é dada por

$$BD = \sum_{k=1}^{I-1} \frac{(A_k - E(A_k))^2}{Var(A_k)} \tag{3.29}$$

A soma em (3.29) não inclui os casos (estratos) envolvendo tabelas com frequências marginais nulas, uma vez que nestes casos a $Var(A_k)$ seria zero. Sob a hipótese nula, a estatística BD tem aproximadamente distribuição χ^2 com $m^{'} - 1$ graus de liberdade, sendo $m^{'}$ o número de tabelas efetivamente consideradas.

O valor esperado e a variância da frequência A_k são obtidos sob a suposição de homogeneidade da razão de chances; $E(A_k)$ é obtida como uma solução da equação quadrática

$$E(A_k) = (\hat{\alpha}(n_{Rk} + m_{1k}) + (n_{Fk} - m_{1k}) \pm \{[\hat{\alpha}(n_{Rk} + m_{1k}) + (n_{Fk} - m_{1k})]^2 - [4(\hat{\alpha} - 1)\hat{\alpha}(n_{Rk} m_{1k})]\}^{1/2})/2(\hat{\alpha} - 1),$$

Em que $\hat{\alpha}$ é uma estimativa da razão de chances comum, e

$$Var(A_k) = \left(\frac{1}{E(A_k)} + \frac{1}{n_{Rk} - E(A_k)} + \frac{1}{m_{1k} - E(A_k)} + \frac{1}{n_{Fk} - m_{1k} + E(A_k)} \right)^{-1}.$$

Capítulo 4

Metodologia Unificada

Por conta da existência de vários métodos para detecção de DIF, e por nem sempre os resultados destes métodos serem coincidentes na indicação da presença ou não de DIF para um mesmo item, surge a necessidade de combinar estes diferentes resultados e apresentar uma única classificação para cada item.

Neste trabalho foi desenvolvido um critério síntese para identificação ou não de DIF no item, com base em alguns métodos apresentados no Capítulo 3 e implementados no pacote *difR versão 5.0*, disponível no software R (R Development Core Team, 2008). Com base nos resultados (presença ou ausência de DIF) apontados por 8 métodos, **Dificuldades Transformadas dos Itens (TID), Padronizado (Std), Regressão Logística (R-Log), SIBTEST, Mantel-Haenszel (MH), Breslow-Day (BD), Lord** e **Raju**, a metodologia proposta unificará esses resultados em uma estatística de teste, que identificará se o item possui ou não DIF. O *tipo de DIF* será identificado com 3 valores: 1 (uniforme), 2 (não uniforme), 3 (uniforme e/ou não uniforme). Os detalhes serão apresentados nas próximas seções.

4.1 O pacote difR

O pacote *DifR versão 2.2* (MAGIS et al., 2010) forneceu uma ferramenta para análise de DIF com base em nove métodos padrão para detectar o funcionamento diferencial em itens dicotômicos. Alguns desses métodos podiam identificar DIF uniforme e/ou não uniforme em dois grupos de respondentes (referência e focal), com base no escore observado ou na nota obtida via TRI. Apesar de não ser o interesse neste trabalho, também há métodos para comparar um grupo de referência e dois ou mais grupos focais. Na *versão 5.0*, por Magis, Beland e Raiche (2018), o pacote *difR* incorporou mais 3 métodos para detecção de DIF, e passou a disponibilizar um total de 12 métodos.

A maioria dos métodos disponíveis na literatura tem software próprio

(ferramenta computacional) para implementação individual, como é o caso do *DICHODIF* (ROGERS; SWAMINATHAN; HAMBLETON, 1993) para implementar o método MH; *IRTDIF* (COHEN et al., 1992) para o método da área de Raju; *IRTLRDIF* (THISSEN, 2001) para o método da razão de verossimilhanças; *SIBTEST* (STOUT et al., 1994) para o método SIBTEST; entre outros pacotes.

A proposta do *difR* é apresentar uma ferramenta que permita a execução de vários métodos para detectar DIF em itens dicotômicos em um único pacote. Os comandos tem uma estrutura similar a de todos os métodos para DIF, e o usuário pode escolher entre métodos baseados na TCT e métodos baseados na TRI. Alguns parâmetros específicos de alguns métodos também estão disponíveis. A Tabela 4.1 mostra os métodos disponíveis no *difR versão 5.0* e as características quanto à abordagem usada (TCT ou TRI), tipo de DIF (U - Uniforme; NU - Não Uniforme) a ser detectado e o número de grupos a serem comparados.

Tabela 4.1 *Métodos para detecção de DIF disponíveis no difR versão 5.0 e suas características.*

	Método	Comando no *R*	Abordagem	Tipo de DIF	*Nº* de grupos
1.	TID	*difTID*	TCT	U	2
2.	Padronizado	*difStd*	TCT	U	2
3.	MH	*difMH*	TCT	U/NU	2
4.	Reg. Logística	*difLogistic*	TCT	U/NU	2
5.	BD	*difBD*	TCT	NU	2
6.	SIBTEST	*difSIBTEST*	TCT	U/NU	2
7.	χ^2 de Lord	*difLord*	TRI	U/NU	2
8.	Raju	*difRaju*	TRI	U/NU	2
9.	Razão de Ver. generalizada	*difLRT*	TRI	U/NU	> 2
10.	MH generalizado	*difGMH*	TCT	U/NU	> 2
11.	Reg. Logística generalizado	*difGenLogistic*	TCT	U/NU	> 2
12.	χ^2 de Lord generalizado	*difGenLord*	TRI	U/NU	> 2

Fonte: Adaptado de Magis et al. (2010)

Apesar de cada um dos métodos apresentados na Tabela 4.1 ter um comando para execução individual, o pacote *difR* também disponibiliza a função *dichoDIF* para a execução simultânea de dois ou mais métodos, tornando possível comparar os resultados para um mesmo item. Esta função permite a execução simultânea apenas dos métodos 1 a 9 apresentados na Tabela 4.1.

Um aspecto importante nos estudos para detecção de DIF é evitar o chamado *problema de confundimento*, que ocorre quando um ou vários itens com DIF

podem levar a uma identificação errônea de DIF em outros itens. Isto provocará um aumento indesejado do Erro Tipo I, ou seja, aumento nos resultados *falso-positivos* associado ao método utilizado.

Em particular, se itens com DIF forem incluídos no conjunto de *itens âncoras* (aqueles que a priori não têm DIF), eles afetarão a medida de habilidade seja nos métodos que utilizam a TCT, uma vez que o escore total será influenciado por estes itens com DIF; seja nos métodos que utilizam a TRI, uma vez que esses itens com DIF influenciarão as estimativas dos parâmetros dos itens, que afetarão a escala da habilidade.

Segundo Magis et al. (2010), alguns autores sugeriram uma eliminação iterativa dos itens com DIF, que é um processo chamado de *purificação de itens*, e baseia-se nos seguintes passos:

1. Teste todos os itens, um por um, assumindo que eles não tem DIF.
2. Defina um conjunto de itens com DIF baseado nos resultados do passo 1.
3. Se o conjunto de itens com DIF estiver vazio após a primeira iteração, ou se este conjunto for idêntico ao obtido na iteração anterior, vá para o passo 6. Caso contrário, vá para o passo 4.
4. Teste todos os itens um por um, omitindo os itens do conjunto obtido no passo 2, exceto quando o item com DIF em questão estiver sendo testado.
5. Defina um conjunto de itens com DIF com base nos resultados do passo 4 e vá para o passo 3.
6. Pare.

O pacote *difR* possui argumentos para executar a purificação dos itens em todos os métodos disponíveis. Para os métodos que usam a TRI, o passo 4 do algoritmo acima é executado descartando-se os itens com DIF no redimensionamento dos parâmetros do item para uma métrica comum. Para os métodos que usam a TCT, o passo 4 é executado descartando-se os itens com DIF do cálculo do escore total do teste.

Como não há garantias de se atingir a regra de parada do algoritmo, ou seja, obter dois conjuntos sucessivos de itens idênticos, o *difR* também disponibiliza um argumento para se definir o número máximo de iterações do processo.

4.2 A Estatística T

Na implementação da metodologia unificada é necessário definir uma estatística de teste a ser usada para a tomada de decisão no confronto das hipóteses: H_0 : *o item não tem DIF* e H_1 : *o item tem DIF*.

Na realização de um teste de hipóteses deve-se controlar a probabilidade α de cometer *o erro de tipo I* (ou nível de significância α do teste), que neste caso corresponde a probabilidade de um resultado *falso-positivo*. Geralmente, na execução de um teste adota-se $\alpha = 0,05$, e define-se o conjunto de valores da estatística do teste (*região crítica*) que produz (pelo menos) esse nível adotado (BOLFARINE; SANDOVAL, 2010). Ou seja, os valores da estatística que apontarão a presença de DIF (rejeição de H_0) no item sob estudo.

Assim, com base nos resultados (presença ou ausência de DIF) apontados pelos 8 métodos: TID, Std, R-Log, SIBTEST, MH, BD, Lord e Raju, propomos uma estatística T representando o número de métodos (índices) entre os 8 que apontam presença de DIF no item, fixado o nível de significância $\alpha = 0,05$. Considerando os $M = 8$ métodos previamente ordenados, seja $I_m(u)$ a variável indicadora de presença de DIF do tipo u pelo método m, ou seja,

$$I_m(u) = \begin{cases} 1, & \text{se há indicação de DIF tipo } u \text{ pelo método } m \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

Portanto, para um item sob estudo, a estatística T representa o total de métodos que apontaram resultado positivo para DIF do tipo u, ou seja, tem-se

$$T = \sum_{m=1}^{M} I_m(u) \tag{4.2}$$

A partir da definição da estatística T, surge uma importante pergunta: a partir de qual valor t^* de T podemos inferir que há DIF com nível de significância α?

Para obter esse valor t^* que minimiza a probabilidade do erro Tipo I, foram realizadas simulações para $\alpha = 0,05$, cujos detalhes serão apresentados a seguir.

4.2.1 Estudo das Probabilidades Associadas aos Valores de T

O uso da estatística T servirá para uma tomada de decisão usual em teste de hipóteses na abordagem unificada, ou seja, de aceitar ou rejeitar a presença de DIF em um particular item i. Neste estudo, foi adotado um nível de significância comum, α, para a indicação de DIF em cada teste. A probabilidade $1 - \alpha$ indica a probabilidade de rejeitar DIF quando realmente não existe, e o estabelecimento do valor de T neste estudo está diretamente associado a essa probabilidade.

Mesmo que os dados sejam gerados sem qualquer tipo de DIF, sabe-se que alguns testes poderão indicar a presença de DIF ao nível de significância adotado. Como são 8 testes envolvidos no estudo, e para cada um espera-se uma probabilidade de rejeitar DIF (pois ela não existirá), a probabilidade associada à $\{T = 0\}$ seria $(1 - \alpha)^8$ se as decisões fossem independentes, mas dificilmente o serão. Por conta disso a simulação pode apresentar valores mais realistas, sem modelar a independência. A estratégia será olhar para as probabilidades $P(T \leq t)$ e obter o valor de t em que tal probabilidade acumulada mais se aproximar de $1 - \alpha$.

Para a obtenção das estimativas das probabilidades associadas aos valores da estatística T foram geradas respostas para 2 grupos com 10.000 indivíduos cada um, respondendo a 45 itens idênticos (sem DIF). Os parâmetros dos itens seguiram a dinâmica:

- Dificuldade (b_i): valores não aleatórios, variando de -2 até 2, com mesma amplitude (1/11).
- Discriminação (a_i): aleatórios, segundo uma $U(0,5; 2,5)$, compatíveis com a métrica normal ($D = 1,702$).
- Acerto Casual (c_i): aleatórios, segundo uma $U(0; 0,3)$.

As habilidades de ambos os grupos foram geradas segundo uma $N(0, 1)$. Com isso, os dados para cada indivíduo foram gerados segundo o Modelo logístico de 3 Parâmetros (3PL) da TRI, que é o adotado no ENEM. Para o processo de simulação foi construída uma sintaxe pelo autor na linguagem R.

A Tabela 4.2 foi obtida por meio do processo de simulação baseada em 200 réplicas e mostra a probabilidade estimada $P(T \leq t)$, para um nível de significância $\alpha = 0,05$, de uma quantidade t de índices estarem apontando

a presença de DIF do tipo 3, ou seja, qualquer DIF, seja uniforme ou não uniforme, ou ambos. Busca-se o valor de t para o qual a coluna correspondente tenha valores mais próximos de $1-\alpha = 0,95$. As colunas relativas a $t = 6, 7, 8$ foram excluídas por baixa ou nenhuma frequência.

Tabela 4.2 *Probabilidades simuladas de não cometer o erro Tipo I segundo o valor da estatística T para $\alpha = 0,05$.*

	Valores estimados de $P(T \leq t)$					
	0	1	2	3	4	5
Item 1	0,886	0,935	0,985	0,995	0,995	1,000
Item 2	0,866	0,940	0,965	0,985	1,000	1,000
Item 3	0,886	0,915	0,945	0,980	0,995	1,000
Item 4	0,871	0,940	0,955	0,985	0,995	1,000
Item 5	0,851	0,935	0,965	0,975	0,985	1,000
Item 6	0,851	0,915	0,935	0,970	0,990	1,000
Item 7	0,861	0,920	0,950	0,980	0,995	1,000
Item 8	0,821	0,920	0,955	0,975	0,995	1,000
Item 9	0,876	0,940	0,960	0,980	0,990	1,000
Item 10	0,836	0,910	0,940	0,965	0,985	1,000
Item 11	0,821	0,915	0,960	0,975	0,985	0,995
Item 12	0,861	0,945	0,970	0,990	1,000	1,000
Item 13	0,831	0,930	0,945	0,990	0,990	1,000
Item 14	0,841	0,930	0,965	0,980	0,995	1,000
Item 15	0,856	0,925	0,960	0,985	0,995	1,000
Item 16	0,861	0,910	0,950	0,975	0,985	1,000
Item 17	0,826	0,930	0,975	0,990	1,000	1,000
Item 18	0,791	0,920	0,940	0,965	0,995	1,000
Item 19	0,826	0,891	0,930	0,960	0,980	1,000
Item 20	0,796	0,925	0,950	0,965	0,995	1,000
Item 21	0,796	0,896	0,960	0,990	0,995	1,000
Item 22	0,816	0,915	0,965	0,985	0,995	1,000
Item 23	0,786	0,896	0,930	0,955	0,975	1,000
Item 24	0,846	0,915	0,940	0,945	0,975	1,000
Item 25	0,836	0,920	0,950	0,965	0,990	1,000
Item 26	0,866	0,920	0,955	0,980	0,995	1,000
Item 27	0,826	0,945	0,955	0,975	0,985	1,000
Item 28	0,821	0,945	0,970	0,985	1,000	1,000
Item 29	0,746	0,881	0,965	0,985	0,995	1,000
Item 30	0,811	0,935	0,965	0,990	1,000	1,000
Item 31	0,816	0,920	0,960	0,985	0,995	1,000
Item 32	0,781	0,886	0,950	0,975	1,000	1,000
Item 33	0,811	0,915	0,960	0,985	0,985	1,000
Item 34	0,841	0,905	0,940	0,965	0,975	1,000
Item 35	0,761	0,896	0,960	0,980	1,000	1,000
Item 36	0,786	0,935	0,970	0,990	0,995	1,000
Item 37	0,821	0,910	0,950	0,985	0,990	1,000
Item 38	0,836	0,896	0,965	0,990	0,990	1,000
Item 39	0,811	0,900	0,965	0,980	0,995	1,000
Item 40	0,756	0,891	0,965	0,990	0,995	0,995
Item 41	0,761	0,896	0,960	0,980	1,000	1,000
Item 42	0,821	0,960	0,985	0,985	0,995	1,000
Item 43	0,796	0,930	0,980	0,985	0,995	1,000
Item 44	0,751	0,876	0,930	0,970	0,980	0,995
Item 45	0,766	0,886	0,945	0,975	0,985	1,000

Percebe-se que em 8 dos 45 itens o valor $t = 1$ gera $P(T \leq t)$ mais próximo de 0,95; para 30 itens o valor $t = 2$ é o que mais aproxima, enquanto para $t = 3$ apenas 3 itens melhor aproximaram de 0,95.Portando, (se mais de 2 dos 8 índices detectarem DIF) tem-se o valor mais próximo do $1 - \alpha$ adotado,

de forma que este será o critério adotado nesta dissertação para a tomada de decisão em favor da hipótese H_0. Portanto, rejeitaremos H_0 se $T > 2$, ou seja, um particular item i será considerado ter DIF se pelo menos 3 dos 8 métodos o acusarem.

Uma outra questão importante é se a taxa de detecções corretas de DIF está associada ao nível de dificuldade do item, o que influenciaria a escolha do valor de t. O ideal é que não tenha associação entre o valor do parâmetro de dificuldade e a taxa de detecção, ou seja, que a correlação seja nula, equivalente a uma reta de regressão linear horizontal. Nas Figuras 4.1 a 4.3 apresenta-se as estimativas de $P(T \leq t)$ para $t = 0, 1, 2$. Nota-se que se a decisão fosse adotar $T = 0$ ou $T = 1$ para a rejeição de DIF, tais critérios estariam associados ao valor do parâmetro de dificuldade, no sentido que quanto mais difícil o item menor seria a taxa de rejeição de DIF. Para $T = 2$ temos a correlação praticamente nula, e o coeficiente linear foi 0,9567, bem próximo do valor $1 - \alpha$ desejado.

Figura 4.1 *Estimativas de* $1 - \alpha$ *em função do parâmetros de dificuldade para* $T = 0$.

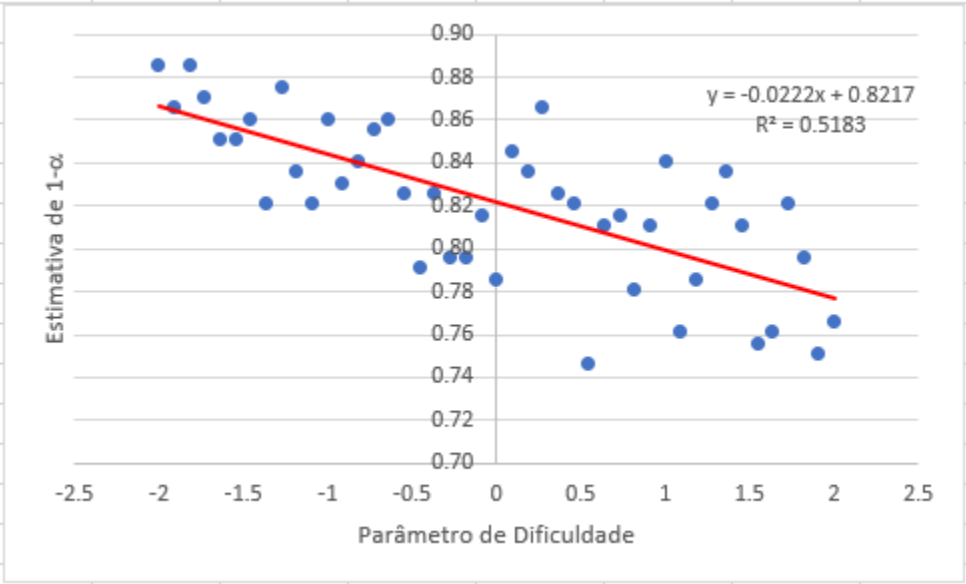


Figura 4.2 *Estimativas de* $1-\alpha$ *em função do parâmetros de dificuldade para* $T=1$.

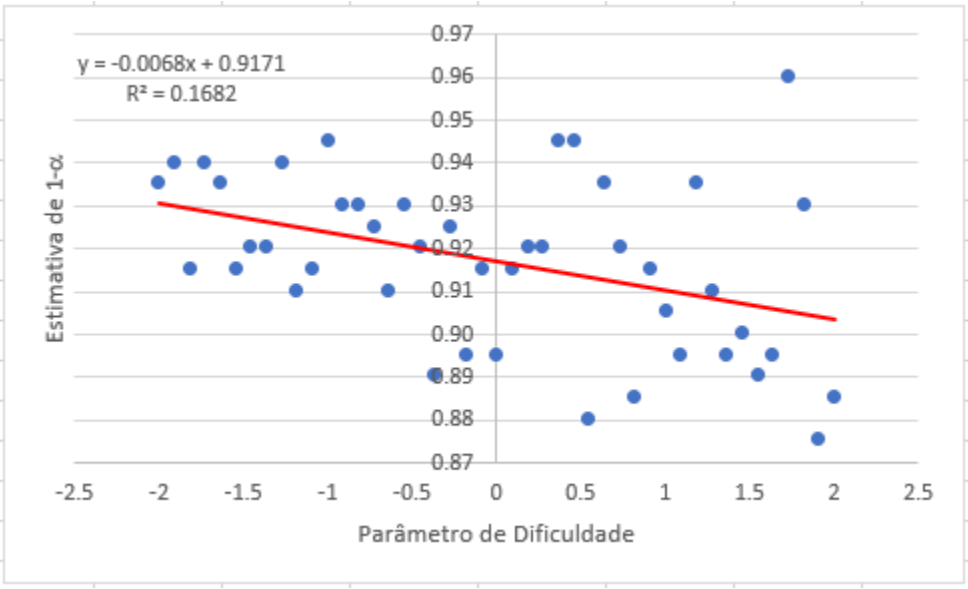
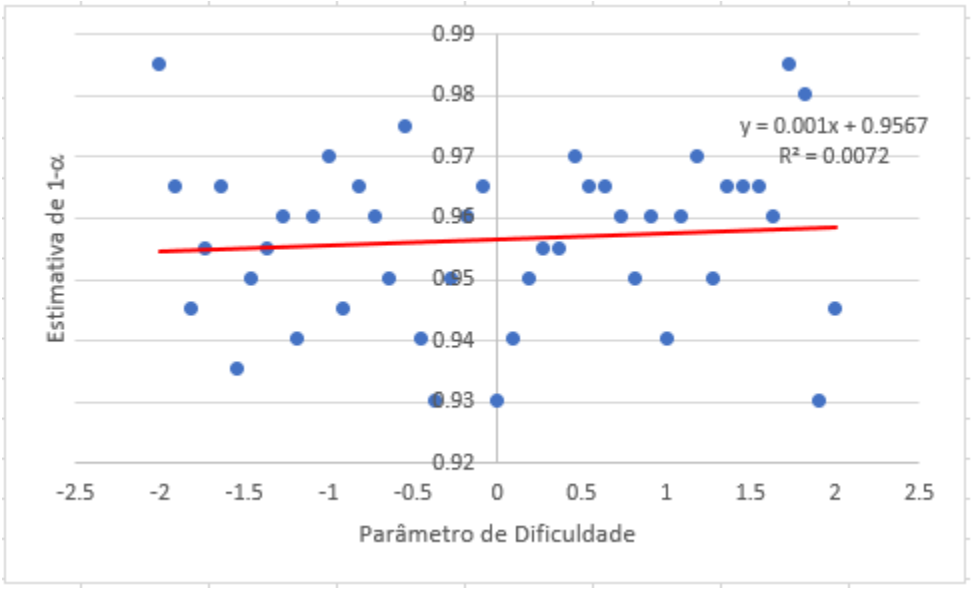


Figura 4.3 *Estimativas de* $1-\alpha$ *em função do parâmetros de dificuldade para* $T=2$.



Capítulo 5

Aplicações

Neste capítulo serão realizadas aplicações da metodologia proposta ao ENEM-2017 com os dados atualizados pelo INEP em 15 de março de 2019, restringindo-se às respostas dadas nas provas das áreas de Linguagens, Códigos e Suas Tecnologias (LC) e Matemática e Suas Tecnologias (MT).

Para a aplicação da metodologia unificada na detecção de DIF nos itens dessas provas, foi necessário inicialmente compor os dois grupos a serem comparados, o grupo focal e o grupo de referência. Como deseja-se comparar o desempenho nos itens entre candidatos com e sem déficit de atenção, foram definidos alguns filtros no banco de dados a fim de controlar fatores externos que poderiam vir a influenciar o desempenho dos candidatos dos dois grupos.

5.1 Caracterização dos Dados

O banco de dados disponibilizado pelo INEP continha o registro de 6.731.341 candidatos. Dentre esses, 63,5% já haviam concluído o Ensino Médio, 26,5% estavam cursando e concluiriam em 2017, 8,9% declararam que concluiriam após 2017 (são os chamados "treineiros"), e 1,1% declararam não terem concluído e nem estar cursando o Ensino Médio.

No preenchimento do questionário sócio-econômico, os candidatos informaram sobre pedidos de atendimento especializado, respondendo de forma dicotômica a alguns indicadores de condições físicas e/ou emocionais. Entre esses indicadores, destacaram-se com maior frequência de respostas autodeclaradas positivas: deficiência física (12.424), déficit de atenção (7.789), baixa visão (7.307), e deficiência auditiva (4.390).

Segundo o INEP (divulgado em 27/10/2017):

- Foram aprovadas 35.653 solicitações de atendimento especializado.

- A maioria dos participantes (24.878) com direito a recurso declararam não precisar de nenhum apoio para realização das provas.

- Os recursos mais solicitados foram sala de fácil acesso (8.758), tempo adicional (8.584), auxílio para leitura (4.902), auxílio para transcrição (4.611) e prova ampliada (4.117).

Para este estudo houve um recorte no banco de dados original do ENEM 2017, e foram realizados os seguintes filtros para a composição dos grupos de estudo:

- candidatos presentes (não faltaram e nem foram eliminados) nas provas das áreas de LC e MT do ENEM 2017;
- candidatos que concluiriam o Ensino Médio em 2017 em instituições de ensino regular;
- foram excluídos os candidatos que estudavam no exterior;
- candidatos que responderam os 4 cadernos principais de prova, na área de LC e de MT: cadernos nas cores azul, amarela, rosa e branca para LC e azul, amarela, rosa e cinza para MT(não foram consideradas as provas adaptadas e de reaplicação);
- foram excluídos candidatos com dados faltantes;

Após o recorte no banco de dados, atendendo os critérios acima, restaram 1.897 candidatos que declararam ter pelo menos déficit de atenção (estes compuseram o grupo focal), e 1.272.429 candidatos que se autodeclararam sem nenhum tipo de deficiência e necessidades especiais e que não solicitaram nenhum atendimento especial (uma amostra de 10.000 candidatos foi selecionada entre eles para compor o grupo referência).

A seleção da amostra de 10.000 candidatos para o grupo referência foi realizada de forma estratificada, a fim de ter uma composição semelhante a do grupo focal. Os estratos foram definidos pela *dependência administrativa da escola onde o candidato cursava o Ensino Médio* (Federal, Estadual, Municipal ou Privada) e o *tipo de escola* (Pública ou Privada). As Tabelas 5.1 e 5.2 mostram a composição do grupo focal e do grupo referência, respectivamente, segundo os estratos considerados.

Em cada uma das áreas, LC e MT, os 4 cadernos de prova considerados neste estudo continham os mesmos itens, havendo apenas alteração na ordem de apresentação dos mesmos no caderno. Neste estudo foi tomado como referência o caderno azul, e o desempenho dos 2 grupos nos itens foram

comparados segundo a ordem apresentada no caderno de referência. Não foram considerados os 5 itens de Língua estrangeira da prova de LC para o estudo do DIF. Assim, foram avaliados para detecção de DIF 40 itens de LC e 45 itens da prova de MT.

Nas próximas seções serão apresentados os resultados referentes às duas áreas incluídas neste estudo. Inicialmente os itens serão comparados nos grupos por meio da probabilidade de acerto (medida clássica de dificuldade do item) e por meio das suas Curvas características. Em seguida, será aplicada a metodologia unificada para detecção de DIF.

Tabela 5.1 *Composição do Grupo Focal segundo a dependência administrativa e o tipo de escola.*

Tipo de Escola	Dependência Administrativa da Escola				Total
	Federal	Estadual	Municipal	Privado	
Pública	24	434	7	0	465
Privada	0	0	0	1.432	1.432
Total	24	434	7	1.432	1.897

Tabela 5.2 *Composição do Grupo Referência segundo a dependência administrativa e o tipo de escola.*

Tipo de Escola	Dependência Administrativa da Escola				Total
	Federal	Estadual	Municipal	Privado	
Pública	126	2.288	37	0	2.451
Privada	0	0	0	7.549	7.549
Total	127	2.288	37	7.549	10.000

5.2 Linguagens, Códigos e Suas Tecnologias (LC)

5.2.1 Proporções de Acertos - Itens da Prova de LC

Um dos primeiros sinais de presença de DIF é uma diferença razoável entre as proporções de acertos dos grupos em análise. No entanto, é possível ter diferença nula e mesmo assim haver DIF, uma vez que a proporção está sendo obtida sem condicionar nas habilidades. A Tabela 5.3 apresenta esses resultados para os itens da prova de LC. Observa-se que 17 itens apresentaram diferença entre 5% e 11%, todos favorecendo o grupo de referência. Apenas 6

itens apresentaram uma pequena diferença negativa, favorecendo levemente o grupo focal.

Tabela 5.3 *Proporção de acertos nos 40 itens da prova de Linguagem, Códigos e suas Tecnologias do ENEM 2017, segundo os grupos focal e referência.*

Item	Grupo Referência (R)	Grupo Focal (F)	Diferença (R - F)
1	0,1121	0,0928	0,0193
2	0,7147	0,6552	0,0595
3	0,4588	0,4070	0,0518
4	0,4282	0,3579	0,0703
5	0,5756	0,5177	0,0579
6	0,3756	0,3548	0,0208
7	0,7813	0,7022	0,0791
8	0,2997	0,2810	0,0187
9	0,2712	0,2804	-0,0092
10	0,5396	0,4976	0,0420
11	0,5055	0,4892	0,0163
12	0,4776	0,4470	0,0306
13	0,5878	0,5098	0,0780
14	0,3119	0,2562	0,0557
15	0,6695	0,5598	0,1097
16	0,2214	0,2267	-0,0053
17	0,7235	0,6658	0,0577
18	0,4977	0,4660	0,0317
19	0,6983	0,6083	0,0900
20	0,4485	0,4091	0,0394
21	0,1735	0,1950	-0,0215
22	0,4815	0,4449	0,0366
23	0,8242	0,7290	0,0952
24	0,6752	0,6410	0,0342
25	0,2071	0,2172	-0,0101
26	0,4108	0,3706	0,0402
27	0,6333	0,5577	0,0756
28	0,5492	0,4802	0,0690
29	0,1424	0,1492	-0,0068
30	0,8741	0,7907	0,0834
31	0,3153	0,3094	0,0059
32	0,2620	0,2710	-0,0090
33	0,6404	0,6073	0,0331
34	0,4712	0,4112	0,0600
35	0,2235	0,2235	0,0000
36	0,5813	0,5061	0,0752
37	0,2861	0,2762	0,0099
38	0,6045	0,5614	0,0431
39	0,6270	0,5630	0,0640
40	0,5548	0,5403	0,0145

5.2.2 Curva Característica Empírica - Itens da Prova de LC

Nas Figuras 5.1 a 5.5 são apresentadas as curvas características dos itens da prova de LC. Elas foram obtidas a partir das respostas dos indivíduos e das proficiências estimadas pelo INEP, que foram categorizadas em intervalos de amplitude 50 e para cada intervalo foi obtida a proporção de acerto em cada item, por cada grupo, considerando os indivíduos com proficiências no respectivo intervalo. As proporções de acertos foram plotadas nos pontos médios dos intervalos (isto é, em 275, 325...), mas apenas os intervalos com pelo menos 5 indivíduos aparecerão nos gráficos. Os itens são apresentados seguindo a ordem do caderno azul da prova.

Podemos observar em praticamente todos os gráficos, quando aumenta a habilidade aumenta também a proporção de acertos ao item, como esperado. No entanto, alguns itens apresentam alguma variação entre os desempenhos dos grupos. Com o uso da metodologia unificada na próxima seção, será possível identificar os itens com DIF.

Figura 5.1 *Curva Característica dos itens 1 a 8 da prova de LC.*

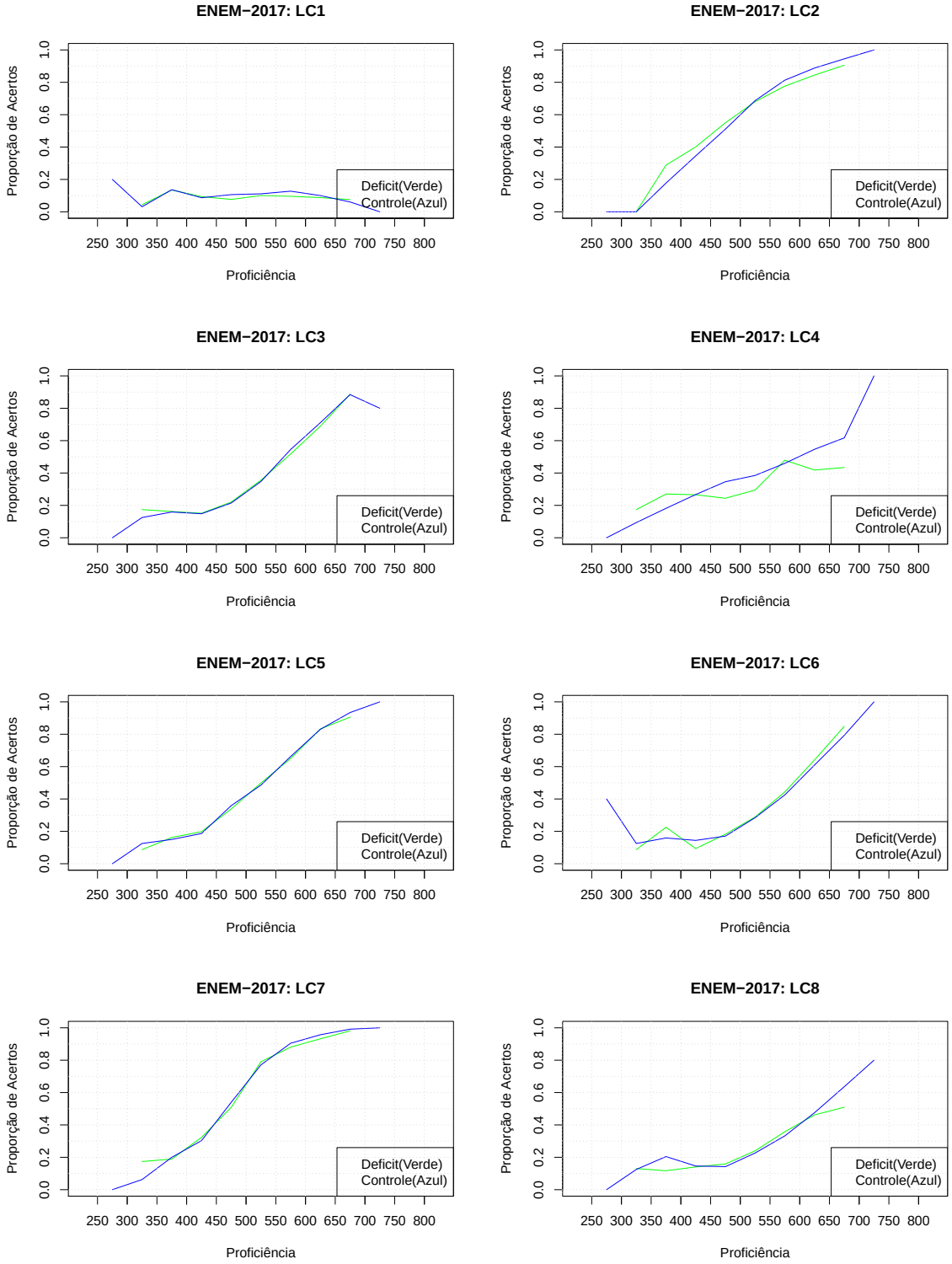


Figura 5.2 *Curva Característica dos itens 9 a 16 da prova de LC.*

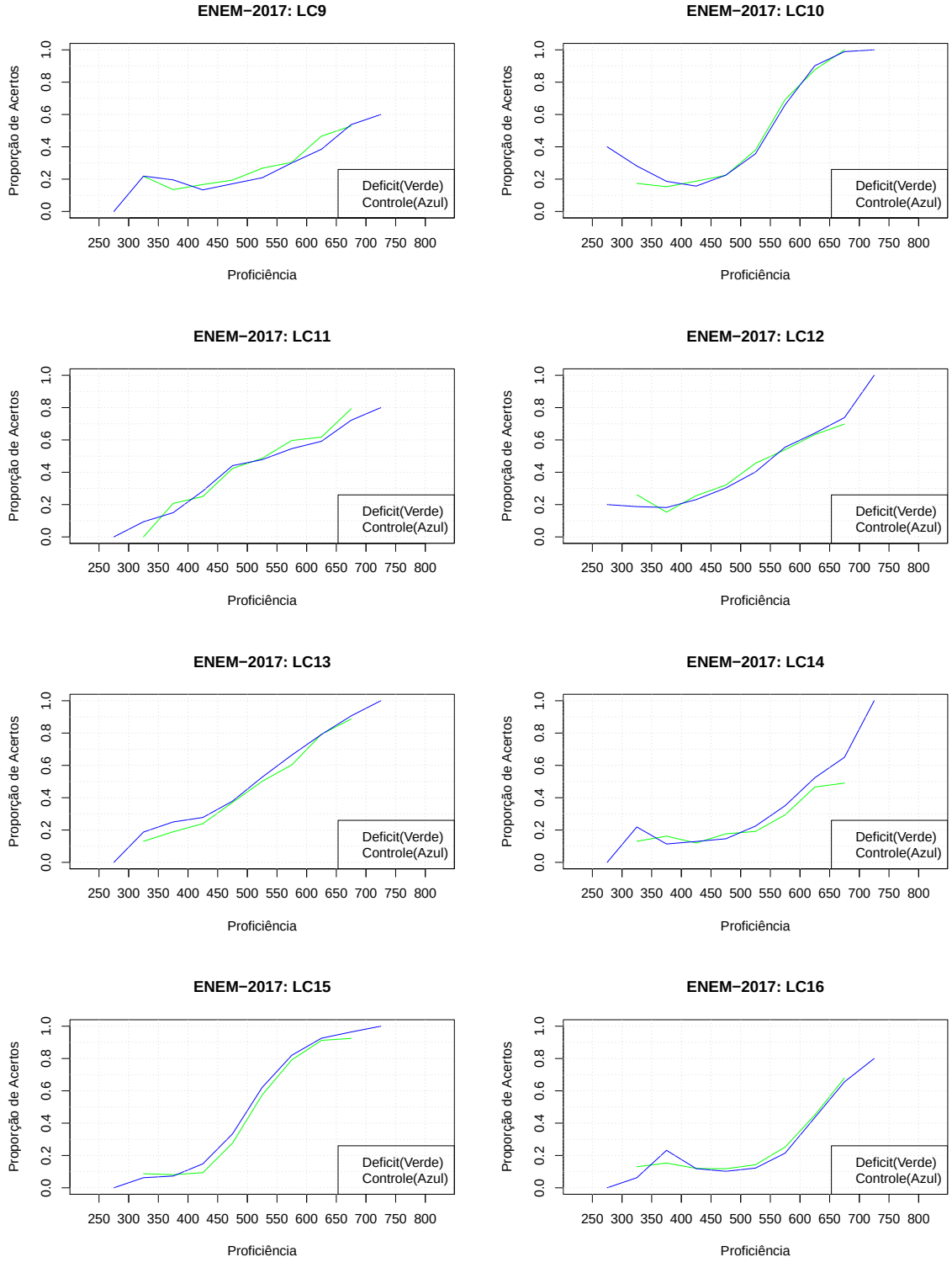


Figura 5.3 *Curva Característica dos itens 17 a 24 da prova de LC.*

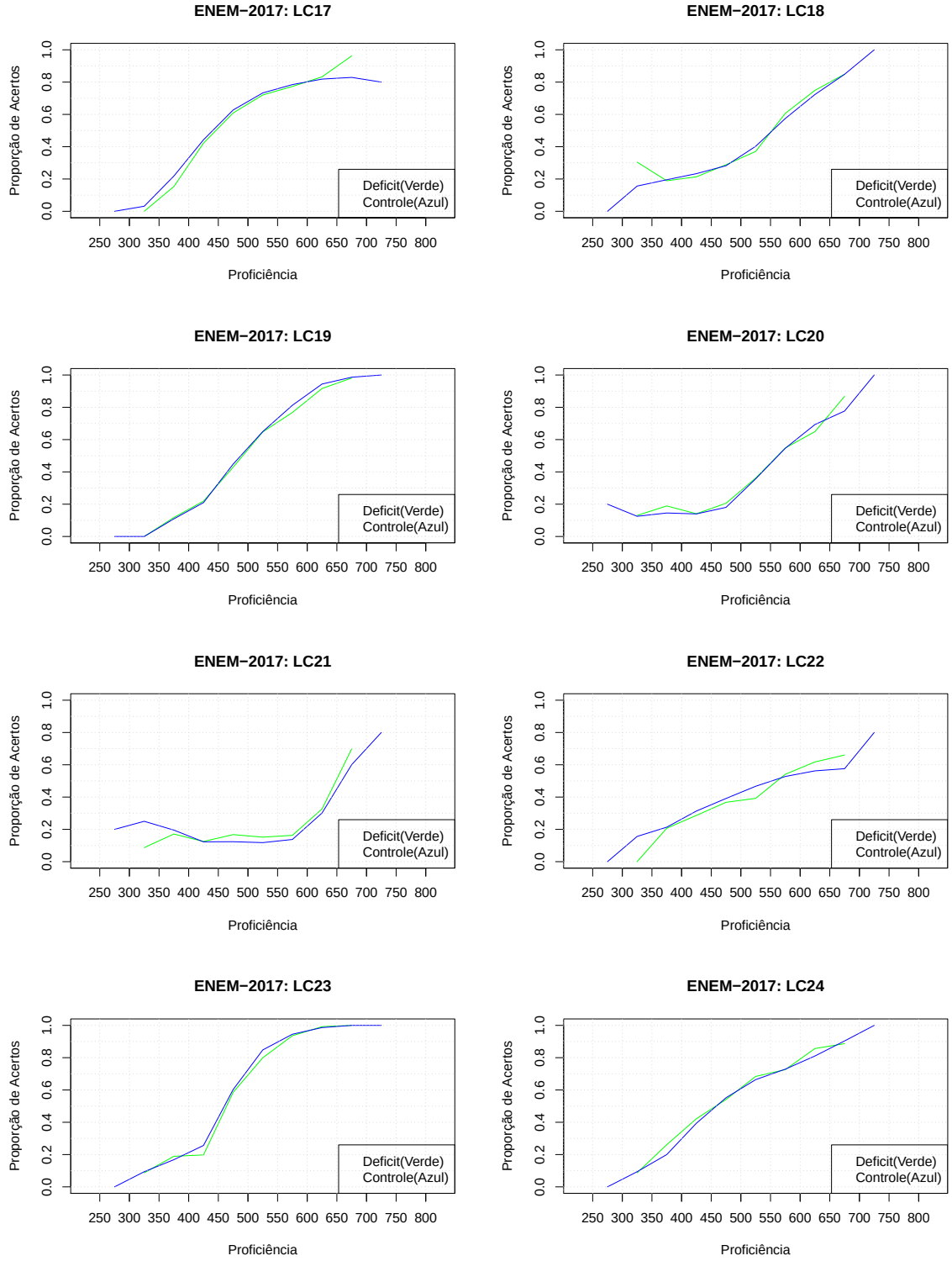


Figura 5.4 *Curva Característica dos itens 25 a 32 da prova de LC.*

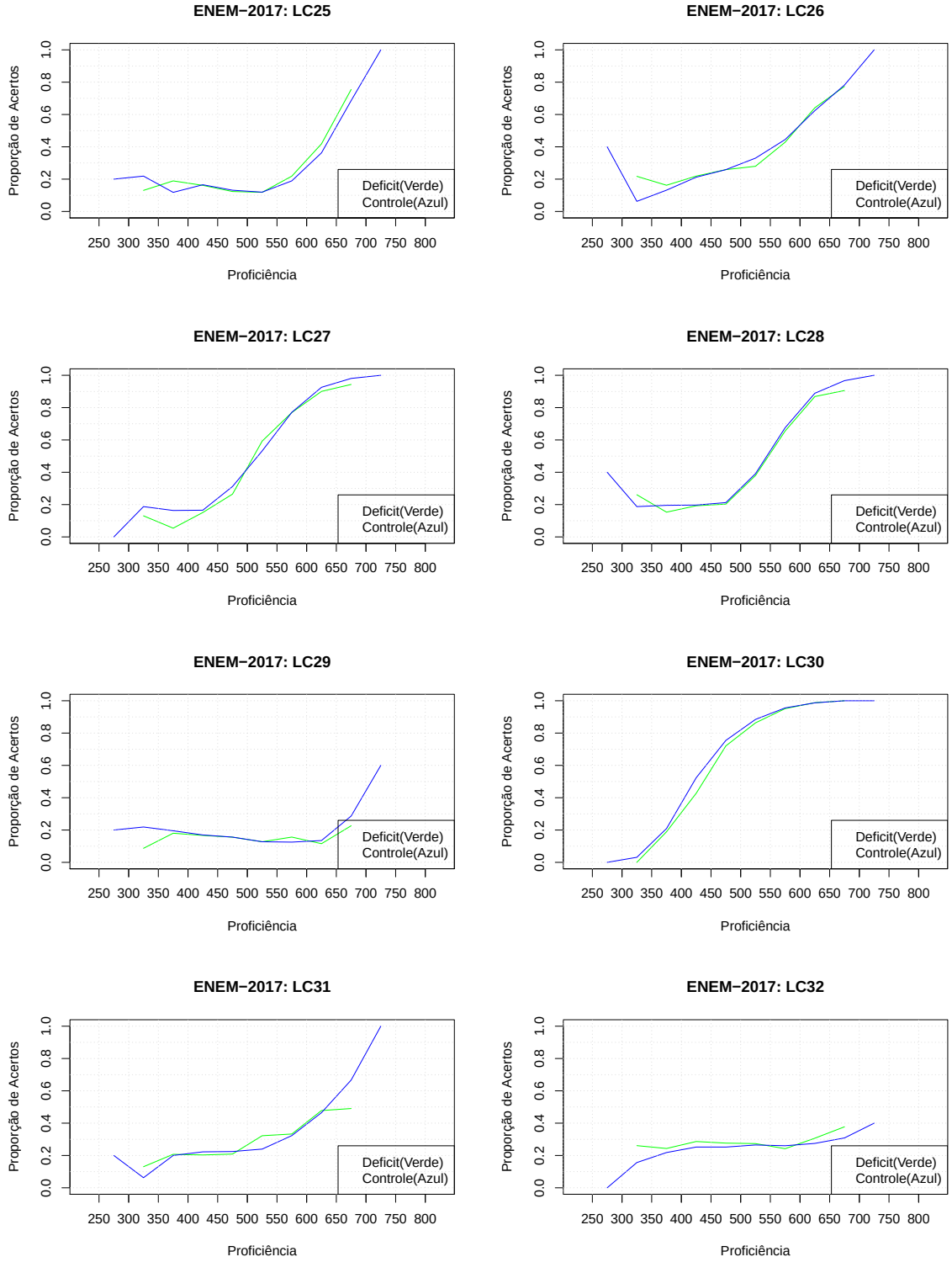
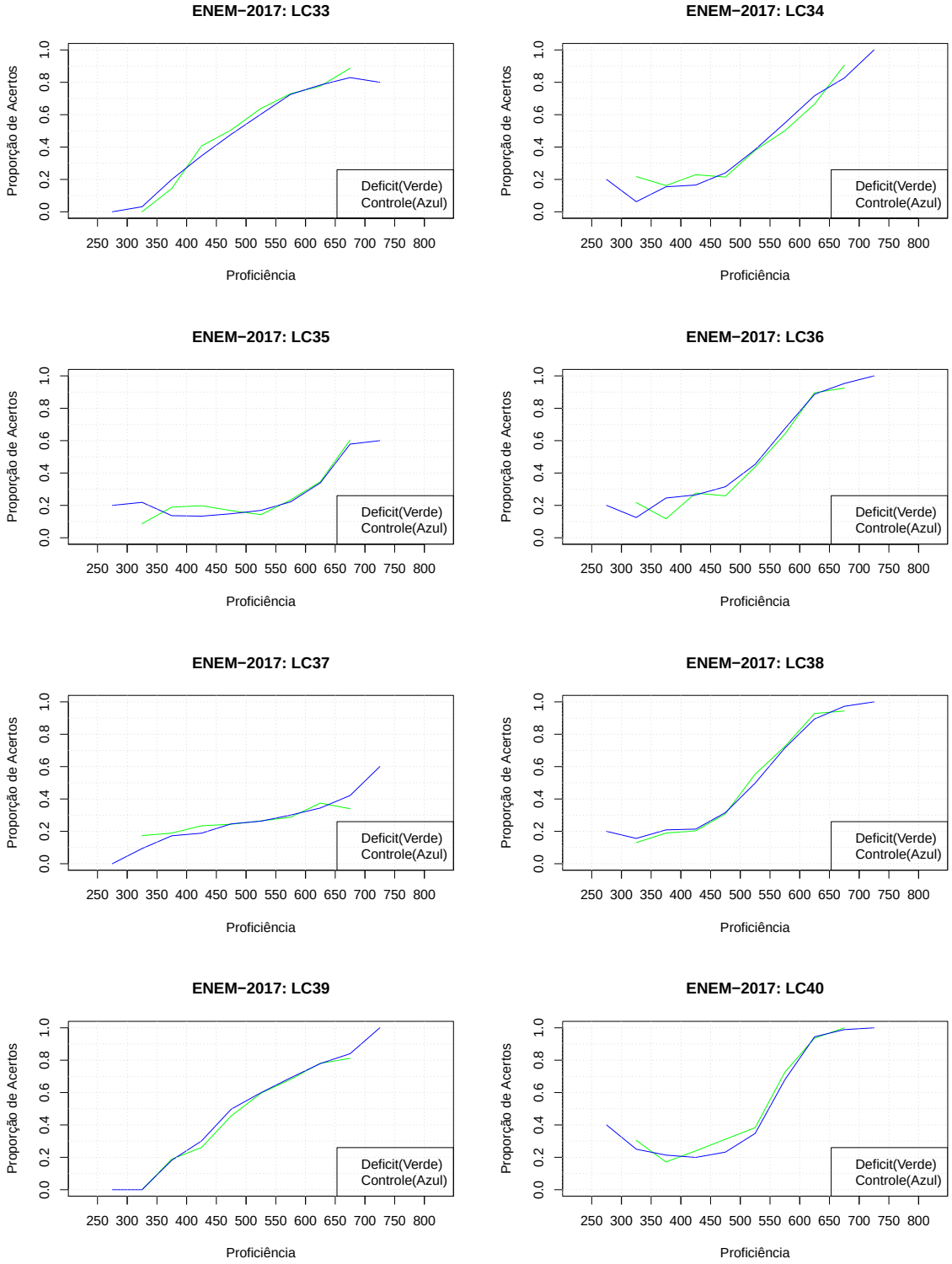


Figura 5.5 *Curva Característica dos itens 33 a 40 da prova de LC.*



5.2.3 Metodologia Unificada - Itens da Prova de LC

Nos resultados que seguem, ressalta-se que os métodos das dificuldades transformadas e padronizado não indicaram DIF em qualquer item da prova de LC. Por outro lado, o método de Mantel-Haenszel apontou 14 itens com DIF, o método da Regressão Logística indicou 21 itens com DIF, o Breslow-Day indicou 2 itens com DIF, o SIBTEST indicou 17 itens com DIF, Lord indicou 37 itens com DIF e o Raju indicou 29 itens com DIF.

Na Tabela 5.4 apresenta-se o resultado da metodologia unificada para todos os itens da prova de LC. Para cada item indica-se se o método detectou (marcado com "x") ou não a presença de DIF, e o valor da estatística T (total de métodos que indicaram DIF), observando que tais itens apontaram haver DIF uniforme e/ou não uniforme.

Usando o critério anteriormente apresentado de aceitar a presença de DIF se $T \geq 3$, foram identificados 22 itens com DIF na prova de LC do ENEM 2017. Estes são: 2, 4, 7, 9, 10, 11, 13, 14, 15, 16, 17, 19, 21, 22, 23, 24, 25, 30, 34, 36, 39 e 40.

Tabela 5.4 *Resultado da metodologia unificada para detecção de DIF nos itens da prova de Linguagens, Códigos e suas Tecnologias do ENEM-2017, por método, e o valor da estatística T.*

	Método								
Item	TID	Std	MH	R-Log	Lord	BD	SIBTEST	Raju	T
1							x		1
2				x	x	x		x	**4**
3					x			x	2
4			x	x	x		x	x	**5**
5					x			x	2
6					x			x	2
7			x	x	x			x	**4**
8					x				1
9			x	x	x				**3**
10					x		x	x	**3**
11				x	x			x	**3**
12					x			x	2
13			x	x	x		x	x	**5**
14			x	x	x		x	x	**5**
15			x	x	x		x	x	**5**
16			x	x	x		x		**4**
17				x	x			x	**3**
18					x			x	2
19			x	x	x		x	x	**5**
20					x			x	2
21			x	x			x		**3**
22				x	x		x	x	**4**
23			x	x	x		x	x	**5**
24				x	x			x	**3**
25			x	x	x		x		**4**
26					x			x	2
27					x			x	2
28					x			x	2
29					x				1
30			x	x	x		x	x	**5**
31					x				1
32						x			1
33					x			x	2
34					x		x	x	**3**
35				x	x				2
36			x	x	x		x	x	**5**
37					x				1
38					x			x	2
39				x	x		x	x	**4**
40			x	x	x		x	x	**5**
Total	0	0	14	21	37	2	17	29	—

5.3 Matemática e Suas Tecnologias

5.3.1 Proporções de Acertos - Itens da Prova de MT

A Tabela 5.5 apresenta as proporções de acertos nos dois grupos para os 45 itens da prova de MT. Observa-se que apenas 3 itens apresentaram diferenças aproximadamente entre 5% e 9%, com esses itens favorecendo o grupo de

referência. Apesar de apresentar um valor pequeno na diferença, tem-se 20 itens favorecendo levemente o grupo focal (com diferença negativa).

Tabela 5.5 *Proporção de acertos nos 45 itens da prova de Matemática e suas Tecnologias do ENEM 2017, segundo os grupos referência e focal.*

Item	Grupo Referência (R)	Grupo Focal (F)	Diferença (R-F)
1	0,1893	0,1977	-0,0084
2	0,1858	0,1919	-0,0061
3	0,4388	0,4549	-0,0161
4	0,5175	0,5082	0,0093
5	0,1143	0,1365	-0,0222
6	0,1720	0,2072	-0,0352
7	0,2609	0,2404	0,0205
8	0,0854	0,0944	-0,0090
9	0,3854	0,3959	-0,0105
10	0,2775	0,2710	0,0065
11	0,3182	0,3300	-0,0118
12	0,3193	0,3110	0,0083
13	0,3883	0,3732	0,0151
14	0,4090	0,3880	0,0210
15	0,5441	0,5050	0,0391
16	0,3118	0,3089	0,0029
17	0,4296	0,4138	0,0158
18	0,2093	0,2198	-0,0105
19	0,7510	0,6595	0,0915
20	0,6229	0,5672	0,0557
21	0,2402	0,2372	0,0030
22	0,5075	0,4618	0,0457
23	0,2058	0,2256	-0,0198
24	0,3492	0,3221	0,0271
25	0,0960	0,1123	-0,0163
26	0,3007	0,2899	0,0108
27	0,1343	0,1360	-0,0017
28	0,3226	0,3142	0,0084
29	0,5499	0,4913	0,0586
30	0,2327	0,2493	-0,0166
31	0,2031	0,1882	0,0149
32	0,1840	0,1940	-0,0100
33	0,3471	0,3231	0,0240
34	0,1718	0,1903	-0,0185
35	0,4359	0,3938	0,0421
36	0,3579	0,3469	0,0110
37	0,1507	0,1645	-0,0138
38	0,3273	0,3158	0,0115
39	0,3527	0,3479	0,0048
40	0,2677	0,2420	0,0257
41	0,2658	0,2720	-0,0062
42	0,3850	0,3875	-0,0025
43	0,7206	0,6747	0,0459
44	0,2863	0,3047	-0,0184
45	0,2106	0,2109	-0,0003

5.3.2 Curva Característica Empírica - Itens da Prova de MT

Os critérios adotados para construção das curvas características empíricas dos itens da prova de MT são os mesmos anteriormente apresentados para os itens de LC. As Figuras 5.6 a 5.11 se referem às questões de 1 a 45 da prova de Matemática e suas Tecnologias, seguindo a ordem do caderno de prova azul.

Figura 5.6 *Curva Característica dos itens 1 a 8 da prova de MT.*

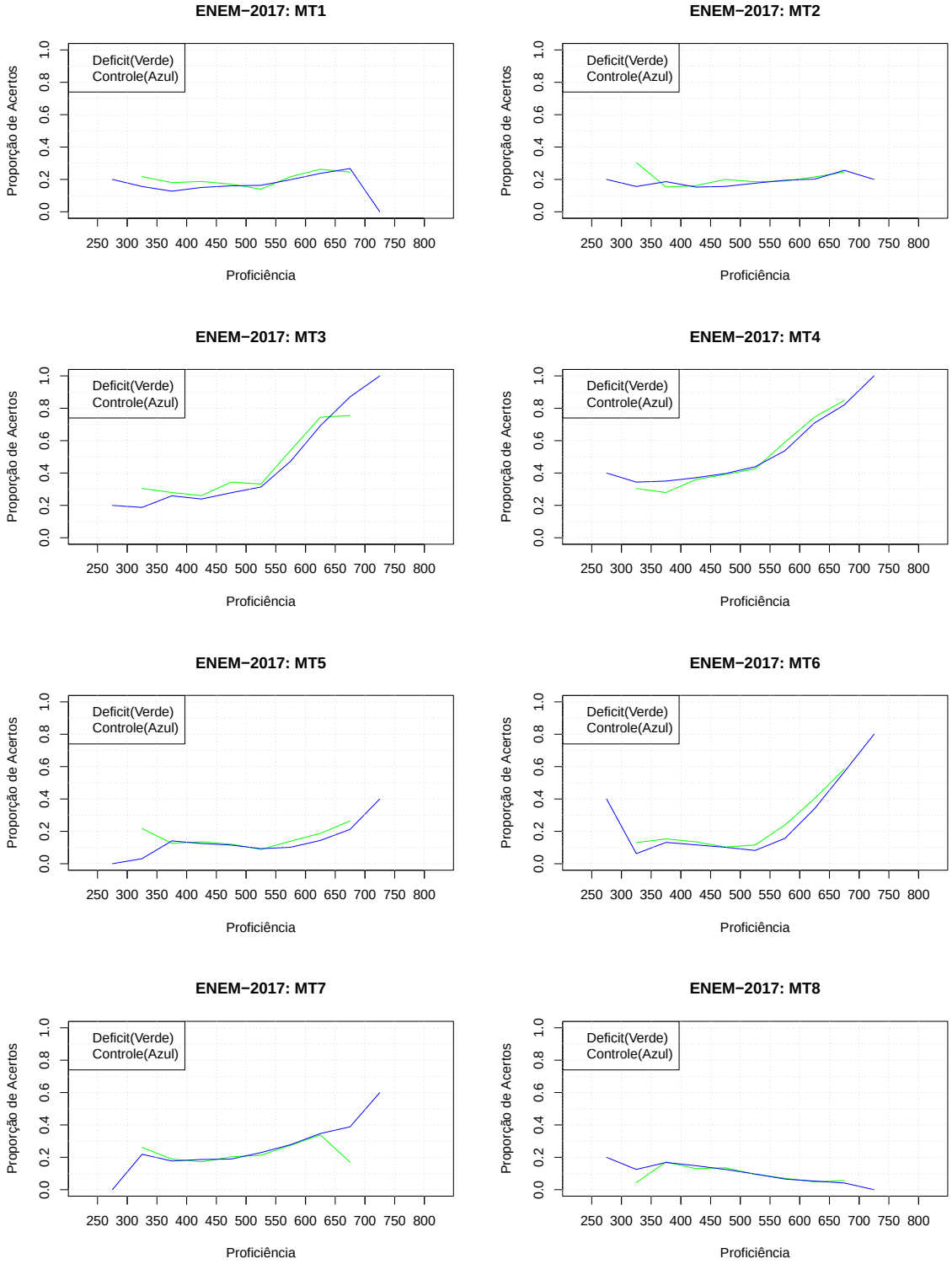


Figura 5.7 *Curva Característica dos itens 9 a 16 da prova de MT.*

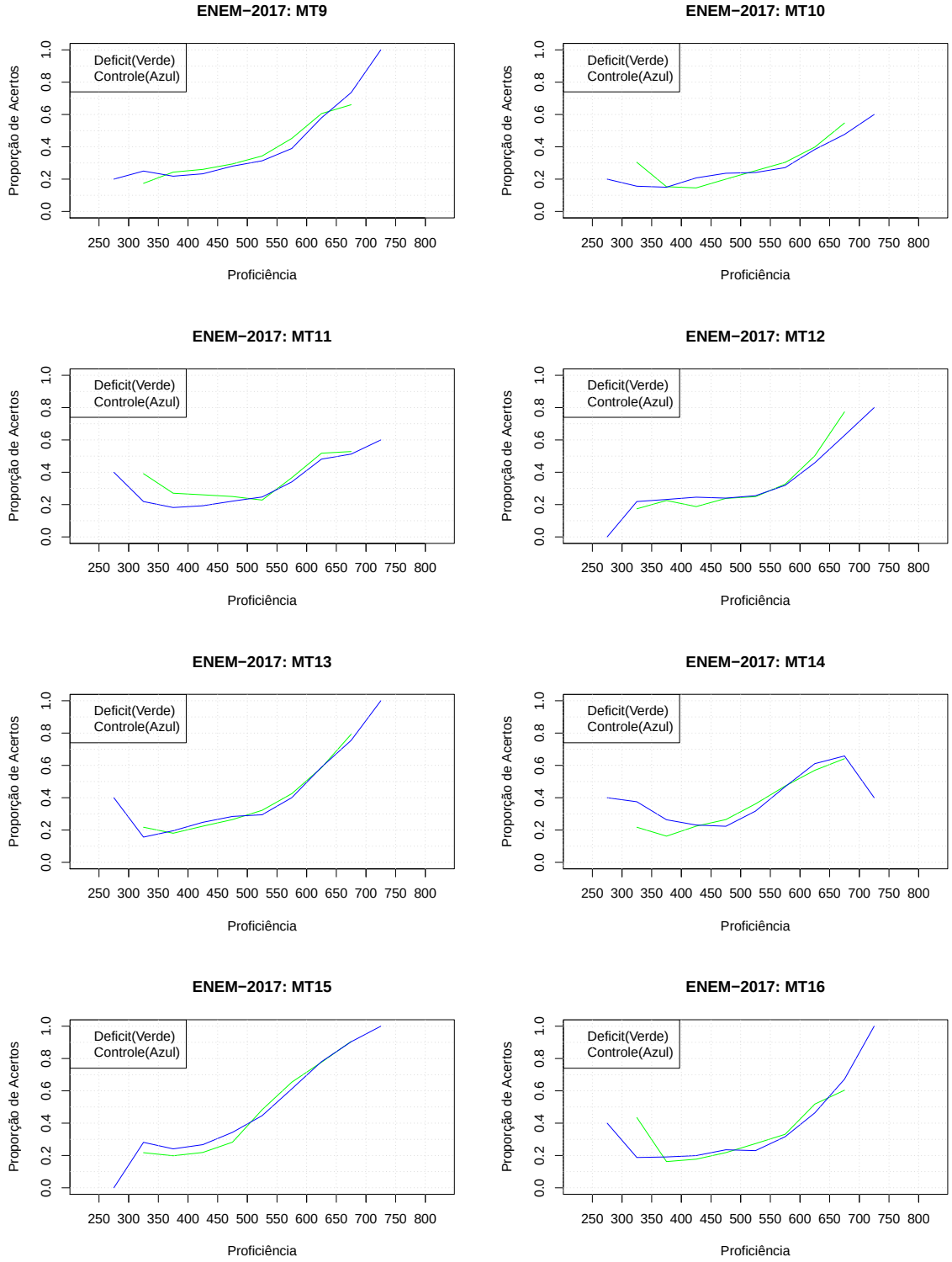


Figura 5.8 *Curva Característica dos itens 17 a 24 da prova de MT.*

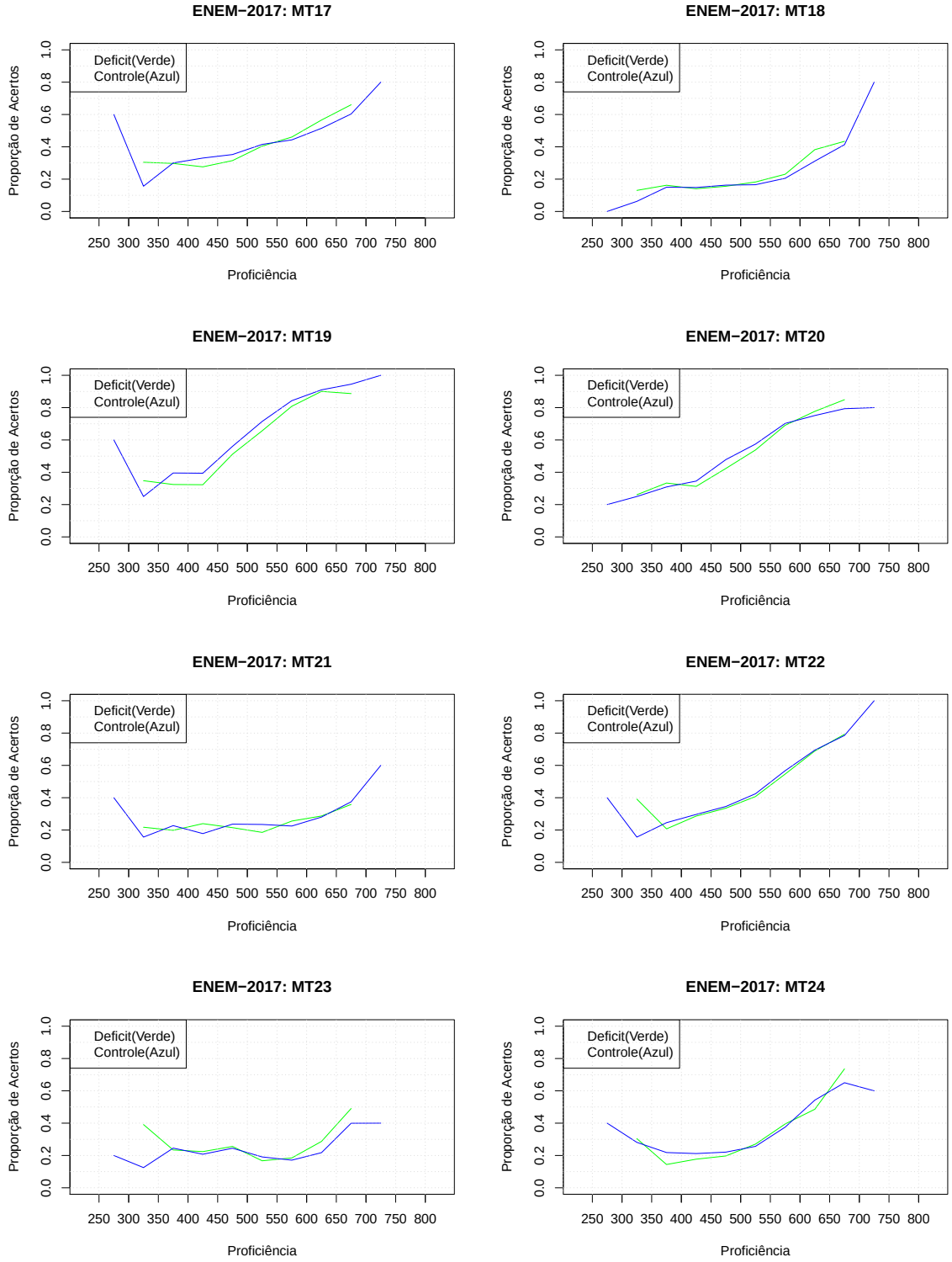


Figura 5.9 *Curva Característica dos itens 25 a 32 da prova de MT.*

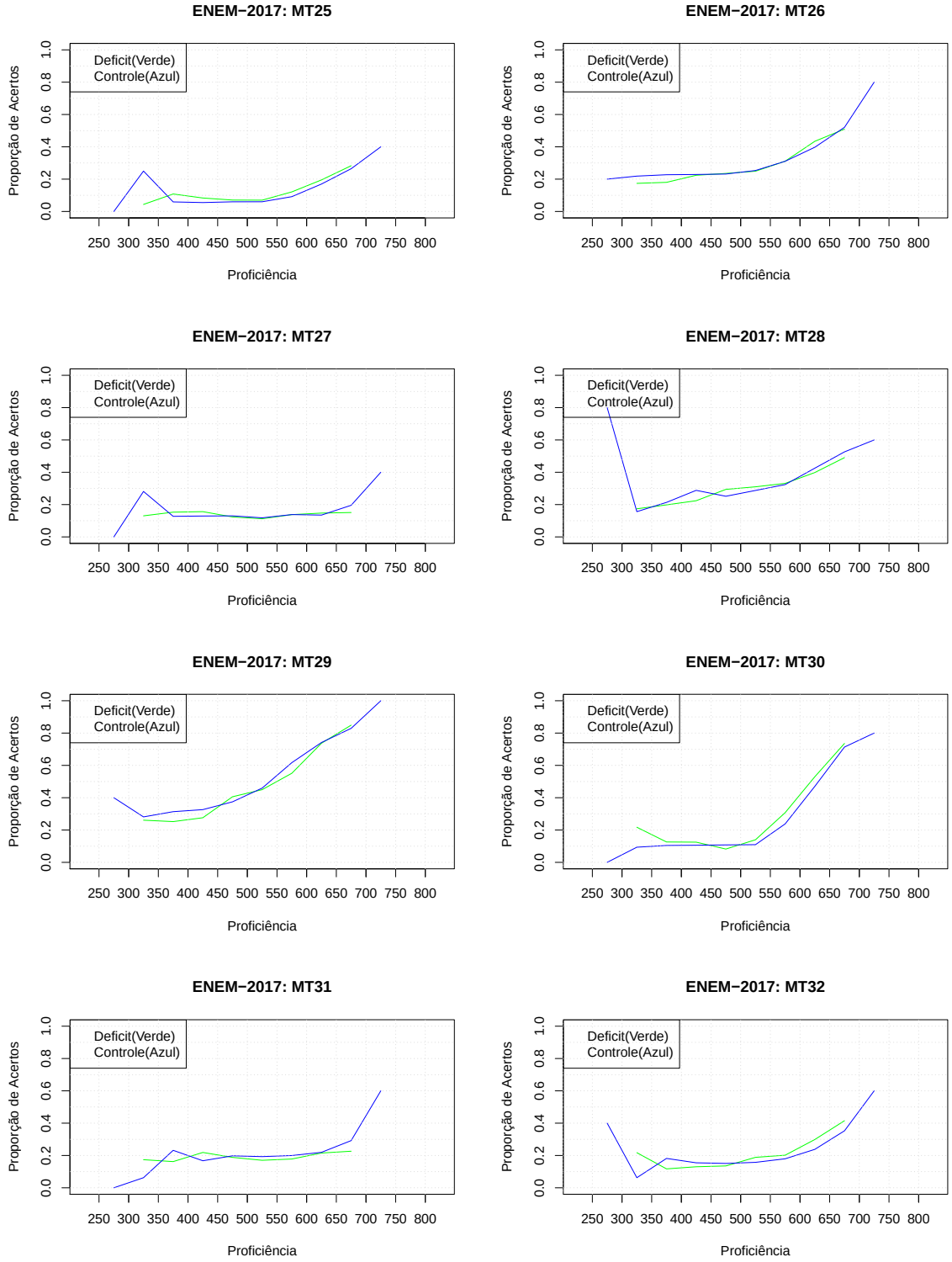


Figura 5.10 *Curva Característica dos itens 33 a 40 da prova de MT.*

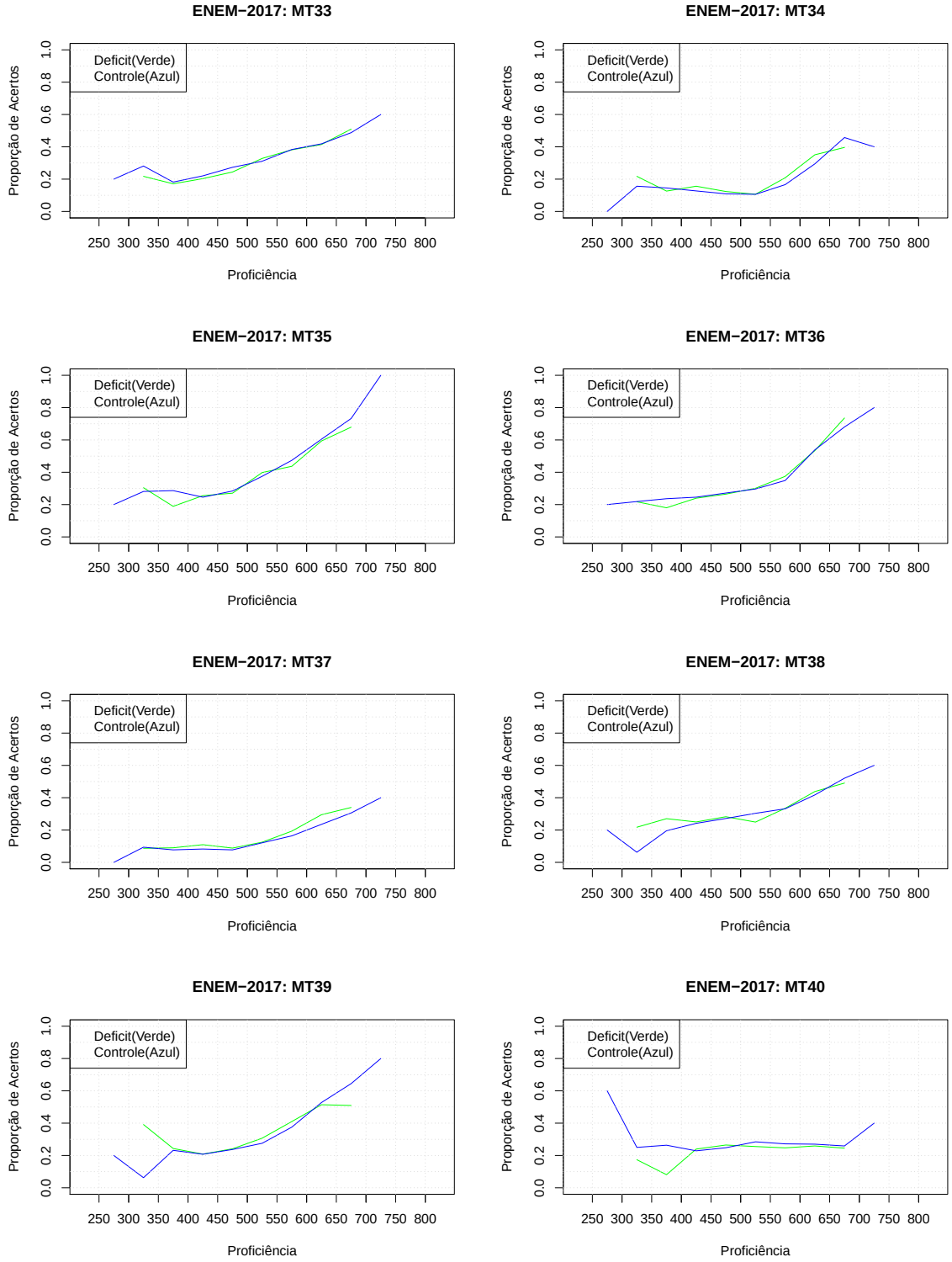
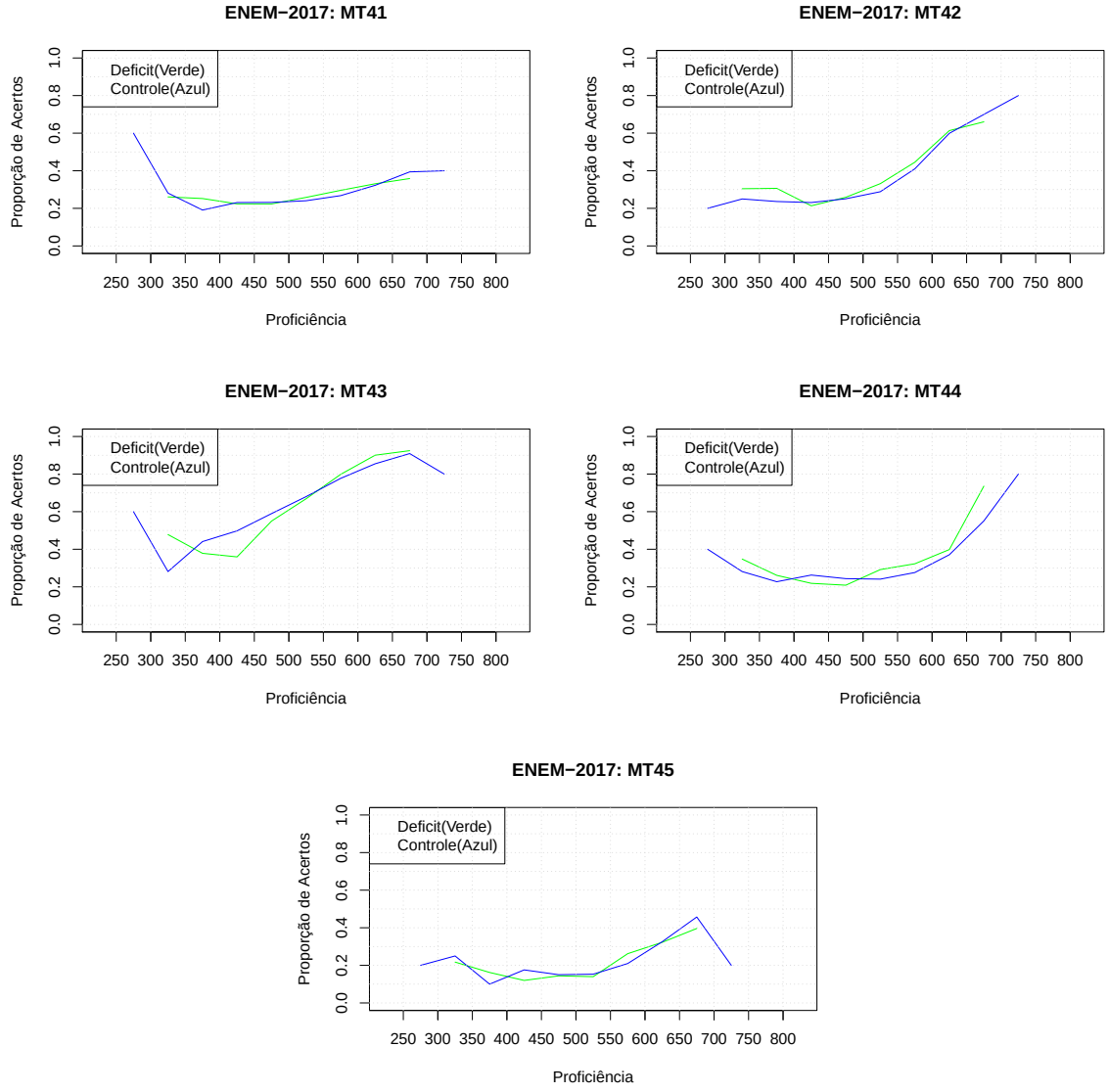


Figura 5.11 *Curva Característica dos itens 41 a 45 da prova de MT.*



5.3.3 Metodologia Unificada - Itens da Prova de MT

Para os itens da prova de MT novamente os métodos das dificuldades transformadas e padronizado não indicaram DIF. O método de Mantel-Haenszel apontou 11 itens com DIF, o método da Regressão Logística indicou 16 itens com DIF, o Breslow-Day indicou 2 itens com DIF, o SIBTEST indicou 17 itens com DIF, o método Lord indicou 43 itens e o Raju indicou 43 itens com DIF.

Na Tabela 5.6 apresenta-se o resultado da metodologia unificada para todos

os itens da prova de MT. Para cada item indica-se se o método detectou (marcado com "x") ou não a presença de DIF, e o valor da estatística T (total de métodos que indicaram DIF), observando que tais itens apontaram haver DIF uniforme e/ou não uniforme.

Usando o critério anteriormente apresentado de aceitar a presença de DIF se $T \geq 3$, foram identificados 21 itens com DIF na prova de MT do ENEM 2017. Estes são: 4, 5, 6, 7, 9, 12, 14, 15, 16, 17, 19, 20, 22, 24, 26, 29, 33, 35, 40, 42 e 43.

5.4 Resumo Geral dos Resultados para os Itens das Provas de LC e MT

Com o uso da metodologia unificada foi possível observar que os métodos TID, Std e BD são pouco sensíveis para detectar DIF. Já os métodos de Lord e Raju apresentaram alta taxa de detecção de DIF. A Tabela 5.7 apresenta uma síntese dessa taxas por método e por área considerada.

Com a regra de decisão adotada na metodologia proposta foram detectados 22 itens com DIF na prova de Linguagem, Códigos e suas Tecnologias, e 21 itens com DIF na prova de Matemática e suas tecnologias. Em ambas as áreas, alguns desses itens favorecem o grupo referência e outros favorecem o grupo focal. Aqui não houve interesse em identificar o grupo favorecido, mas alertar que alguns cuidados adicionais precisam ser tomados na elaboração dos itens do ENEM.

Tabela 5.6 *Resultado da metodologia unificada para detecção de DIF nos itens da prova de Matemática e suas Tecnologias do ENEM-2017, por método, e o valor da estatística T.*

	Método								
Item	TID	Std	MH	R-Log	Lord	BD	SIBTEST	Raju	T
1					x			x	2
2					x				1
3					x			x	2
4					x		x	x	**3**
5			x	x	x		x	x	**5**
6			x	x	x		x	x	**5**
7				x	x			x	**3**
8					x			x	2
9					x			x	**3**
10					x			x	2
11					x			x	2
12				x	x		x	x	**4**
13					x			x	2
14				x	x			x	**3**
15			x	x	x		x	x	**5**
16					x	x		x	**3**
17				x	x		x	x	**4**
18					x			x	2
19			x	x	x		x	x	**5**
20			x	x	x		x	x	**5**
21					x			x	2
22			x	x	x		x	x	**5**
23					x			x	2
24			x	x	x		x	x	**5**
25					x			x	2
26					x		x	x	**3**
27					x			x	2
28					x			x	2
29			x	x	x		x	x	**5**
30					x			x	2
31							x	x	2
32					x			x	2
33				x	x		x	x	**4**
34					x			x	2
35			x	x	x		x	x	**5**
36					x			x	2
37					x			x	2
38					x			x	2
39					x			x	2
40			x	x			x		**3**
41					x			x	2
42					x	x		x	**3**
43			x	x	x		x	x	**5**
44					x			x	2
45					x			x	2
Total	0	0	11	16	43	2	17	43	—

Tabela 5.7 *Número de Itens Detectados com DIF por cada Método e Área.*

Métodos	LC	MT
TID	0	0
MH	14	11
Logístico	21	16
Padronizado (Std)	0	0
Breslow-Day (BD)	2	2
SIBTEST	17	17
Lord	37	43
Raju	29	43

Capítulo 6

Conclusões e Considerações Gerais

Este trabalho teve por objetivo apresentar as principais técnicas disponíveis na literatura para detecção de DIF (Funcionamento Diferencial de Itens) em itens que compõem um instrumento de avaliação. Algumas dessas técnicas surgiram na década de 60, e desde então muitas contribuições foram dadas ao tema.

Neste estudo foram apresentados 8 métodos para detecção de DIF em itens dicotômicos, considerando 2 grupos de respondentes, grupo referência e grupo focal. Foram eles: Método das Dificuldades Transformadas dos Itens (TID), Método Qui-quadrado de Lord, Método Padronizado (Std), Método de Raju, Método da Regressão Logística (R-Log), Método SIBTEST, Método de Mantel-Haenszel (MH) e Método de Breslow-Day (BD).

Esses 8 métodos, ao serem usados conjuntamente para detecção de DIF em um mesmo item, podem apresentar resultados distintos. Por conta disto, foi proposta uma metodologia unificada neste trabalho. Essa metodologia unifica os resultados dos 8 métodos relacionados acima em uma estatística de teste resumo T, que conta quantos métodos indicaram haver DIF para um particular item. Foram realizadas simulações para definir a regra de decisão (ou região crítica do teste) e obteve-se, para um nível de significância $\alpha = 0,05$, que o item sob estudo apresenta DIF se $T \geq 3$ (pelo menos 3 métodos entre os 8 indicarem DIF).

A metodologia proposta foi aplicada, com o uso do *pacote difR*, aos dados do ENEM 2017, considerando 40 itens da prova de Linguagens, Códigos e suas Tecnologias (LC) e 45 itens de Matemática e suas Tecnologias (MT). Com um recorte adequado na base de dados do ENEM 2017, foram considerados apenas candidatos concluintes do ensino médio no ano de 2017 em escola brasileira de ensino regular, que realizaram as provas de LC e MT nos cadernos principais (cores azul, amarela, rosa e branca), e que não apresentavam dados faltantes.

Após o recorte, o grupo focal foi composto pelos 1.897 candidatos que

declararam ter pelo menos déficit de atenção. Entre os 1.272.429 candidatos que se autodeclararam sem nenhum tipo de deficiência e necessidades especiais, foi retirada uma amostra estratificada de 10.000 candidatos para compor o grupo de referência, de forma que tivesse uma composição semelhante ao do grupo focal. Os estratos foram definidos pela dependência administrativa da escola onde o candidato cursava o Ensino Médio (Federal, Estadual, Municipal ou Privada) e o tipo de escola (Pública ou Privada).

A metodologia proposta detectou 22 itens da prova de LC e 21 itens da prova de MT com DIF. Alguns desses itens favorecem o grupo referência e outros favorecem o grupo focal. Foi possível observar que os métodos TID, Std e BD são pouco sensíveis para detectar DIF. Já os métodos de Lord e Raju apresentaram alta taxa de detecção de DIF.

Recomenda-se para trabalhos futuros:

- Avaliar se itens identificados com DIF estão associados a uma competência de área específica da matriz de Referência do ENEM.
- Construir uma versão da estatística T^* ponderada, em que os testes mais precisos na detecção de DIF terão maior peso.
- Realizar estudos de simulação em que as distribuições das proficiências simuladas reproduzam as distribuições dos dados reais para cada área avaliada.
- Disponibilizar uma versão do pacote $difR$ com saídas apropriadas para as avaliações brasileiras e com a implementação da estatística T.

Bibliografia

AGRESTI, A. Models for matched pairs. **Symmetry Models: Categorical Data Analysis. New York: John Wiley & Sons**, p. 409–454, 1990.

AGUIRRE, L.A. Introdução à Identificação de Sistemas: Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais, 2 edn, Editora UFMG, Belo Horizonte, BH. Identificação e Estimação dos Parâmetros de Modelos. **Revista Brasileira de Estatıstica**, v. 59, n. 212, p. 25–51, 2004.

ANDRADE, D.F.; TAVARES, H.R.; VALLE, R.C. Teoria da Resposta ao Item: conceitos e aplicações. **ABE, Sao Paulo**, 2000.

ANDRIOLA, W.B. Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (DIF). **Psicologia: Reflexão e Crıtica**, SciELO Brasil, v. 14, n. 3, p. 643–652, 2001.

_____. **Detección del funcionamiento diferencial del ıtem (DIF) en tests de rendimiento**. 2002. Tese (Doutorado) – Universidad Complutense de Madrid.

ANGOFF, W.H. Summary and derivation of equating methods used at ETS. **Test equating**, Academic Press New York, p. 55–69, 1982.

ANGOFF, W.H.; FORD, S.F. Item-race interaction on a test of scholastic aptitude 1. **Journal of Educational Measurement**, Wiley Online Library, v. 10, n. 2, p. 95–105, 1973.

BOLFARINE, H.; SANDOVAL, M.C. Introdução a inferência estatıstica, 2010.

BRESLOW, N.E.; DAY, N.E.; HESELTINE, E. **Statistical methods in cancer research**. [S.l.]: International Agency for Research on Cancer Lyon, 1980. v. 1.

BUZICK, H.; STONE, E. Recommendations for conducting differential item functioning (DIF) analyses for students with disabilities based on previous DIF studies. **ETS Research Report Series**, Wiley Online Library, v. 2011, n. 2, p. i–26, 2011.

CAMILLI, G.; SHEPARD, L.A. MMSS: Methods for identifying biased test items, 1994.

CHALMERS, A.F. **Qu'est-ce que la science?: récents développements en philosophie des sciences: Popper, Kuhn, Lakatos, Feyerabend**. [S.l.]: La découverte, 2018.

CLAUSER, B.; NUNGESTER, R.J; SWAMINATHAN, H. Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. **Journal of Educational Measurement**, Wiley Online Library, v. 33, n. 4, p. 453–464, 1996.

COHEN, A.R. et al. A modified transfusion program for prevention of stroke in sickle cell disease, 1992.

DORANS, N.J.; HOLLAND, P.W. DIF Detection and description: Mantel-Haenszel and standardization 1, 2. **ETS Research Report Series**, Wiley Online Library, v. 1992, n. 1, p. i–40, 1992.

FOSSEY, A. **Item Analysis – Differential Item Functioning (DIF)**. 2014. Disponível em: <`https://https://www.questionmark.com/item-analysis-differential-item-functioning-dif/`>.

HAMBLETON, R.K.; SWAMINATHAN, H.; ROGERS, H.J. **Fundamentals of item response theory**. [S.l.]: Sage, 1991.

HOLLAND, P.W.; THAYER, D.T. Differential item functioning and the Mantel-Haenszel procedures. **Test Validity. Hillsdale, NJ: Lawrence Erlbaum**, p. 129–145, 1988.

HOSMER, D.W.; LEMESHOW, S.; STURDIVANT, R.X. Logistic regression for matched case-control studies. **Applied logistic regression**, Wiley New York, v. 2, p. 223–259, 1989.

KIM, S.; COHEN, A.S.; PARK, T.K. Detection of differential item functioning in multiple groups. **Journal of Educational Measurement**, Wiley Online Library, v. 32, n. 3, p. 261–276, 1995.

LANDIS, J.R.; HEYMAN, E.R.; KOCH, G.G. Average partial association in three-way contingency tables: a review and discussion of alternative tests. **International Statistical Review/Revue Internationale de Statistique**, JSTOR, p. 237–254, 1978.

LI, H.H.; STOUT, W. A new procedure for detection of crossing DIF. **Psychometrika**, Springer, v. 61, n. 4, p. 647–677, 1996.

LORD, F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ, Lawrence Erlbaum Ass, 1980.

MAGIS, D.; BELAND, S.; RAICHE, G. Package 'difR', 2018.

MAGIS, D. et al. A general framework and an R package for the detection of dichotomous differential item functioning. **Behavior research methods**, Springer, v. 42, n. 3, p. 847–862, 2010.

MAGIS, D. et al. Central modulation in cluster headache patients treated with occipital nerve stimulation: an FDG-PET study. **BMC neurology**, Springer, v. 11, n. 1, p. 25, 2011.

MANTEL, N.; HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. **Journal of the national cancer institute**, Oxford University Press, v. 22, n. 4, p. 719–748, 1959.

MELLENBERGH, G.J. Item bias and item response theory. **International journal of educational research**, Elsevier, v. 13, n. 2, p. 127–143, 1989.

MILLSAP, R.; EVERSON, H.T. Methodology review: Statistical approaches for assessing measurement bias. **Applied psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 17, n. 4, p. 297–334, 1993.

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação**. [S.l.]: Editora Vozes Limitada, 2017.

PENFIELD, R.D. Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of nonuniform DIF. **Alberta Journal of Educational Research**, v. 49, n. 3, 2003.

PENFIELD, R.D. Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. **Applied Measurement in Education**, Taylor & Francis, v. 14, n. 3, p. 235–259, 2001.

RAJU, N.S. Determining the significance of estimated signed and unsigned areas between two item response functions. **Applied Psychological Measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 14, n. 2, p. 197–207, 1990.

ROGERS, H.J.; SWAMINATHAN, H.; HAMBLETON, R.K. DICHODIF A FORTRAN program for DIF analysis of dichotomously scored item response data [A computer program]. **Amherst, MA: University of Massachusetts**, 1993.

SCHEUNEMAN, J.D.; GERRITZ, K. Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. **Journal of Educational Measurement**, Wiley Online Library, v. 27, n. 2, p. 109–131, 1990.

SHEALY, R.; STOUT, W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. **Psychometrika**, Springer, v. 58, n. 2, p. 159–194, 1993.

STOUT, R.D. et al. Staphylococcal glycocalyx activates macrophage prostaglandin E2 and interleukin 1 production and modulates tumor necrosis factor alpha and nitric oxide production. **Infection and immunity**, Am Soc Microbiol, v. 62, n. 10, p. 4160–4166, 1994.

SWAMINATHAN, H.; ROGERS, H.J. Detecting differential item functioning using logistic regression procedures. **Journal of Educational measurement**, Wiley Online Library, v. 27, n. 4, p. 361–370, 1990.

THISSEN, D. IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. **Chapel Hill, NC: LL Thurstone Psychometric Laboratory**, 2001.

VAN DER FLIER, H. et al. An iterative item bias detection method. **Journal of educational measurement**, Wiley Online Library, v. 21, n. 2, p. 131–145, 1984.

WHITMORE, M.L.; SCHUMACKER, R.E. A comparison of logistic regression and analysis of variance differential item functioning detection methods. **Educational and Psychological Measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 59, n. 6, p. 910–927, 1999.

ZUMBO, B.D. A handbook on the theory and methods of differential item functioning (DIF). **Ottawa: National Defense Headquarters**, 1999.