Avaliação dos métodos para detecção de DIF
(Differential Item Functioning)

Aluno: Luís Otávio Marques Fernandes
Prof. : Marcos Antônio da Cunha Santos (Orientador)

1. Introdução

Exames escolares, acadêmicos ou concursos com o formato de questões de múltipla escolha, podem ser avaliados pela Teoria Clássica ou pela Teoria de Resposta ao Item (TRI). A teoria clássica é o método mais usual, o qual a nota final é simplesmente a porcentagem de acerto do estudante. Já a TRI leva em consideração outros fatores para a nota final do exame, como a discriminação, dificuldade de cada item e a probabilidade de acerto ao acaso. O que torna o exame mais justo, no sentido de avaliar o aluno com mais exatidão. Outra vantagem é poder avaliar os itens, notar o grau de dificuldade de cada item, notar se o item discrimina bem os candidatos e qual é a probabilidade do indivíduo acertar o item sem ter conhecimento para tal ("chute").

Os itens devem ser padronizados para não favorecerem nenhum grupo. Este grupo pode ser um grupo ético, pessoas que moram em determinada região ou até mesmo mulheres ou homens. É interessante que estudantes com mesma habilidade tenham a mesma probabilidade de acertar um item. Porém, isto não acontece se o item privilegiar certo grupo. Neste caso o item não é interessante e existe a presença de DIF (Differential Item Functioning). O DIF ocorre quando estudantes com a mesma habilidade têm diferentes probabilidades de acertar um item.

Vários métodos são utilizados para detectar a presença de DIF, os principais são: Paradoxo de Simpson: Fundamento para Detecção do DIF, Comparação das Probabilidades de Acertar o Item, Comparação dos Parâmetros dos Itens, Qui-quadrado de Pearson ou Total, Qui-quadrado de Lord, Cálculo da Área entre as Curvas Características dos itens, Método Logístico Interativo, Método Padronizado, Método Mantel-Haenszel, Regressão Logística e Qui-quadrado de Scheuneman.

Grande parte dos métodos citados acima apresenta falhas e não são eficientes. Os principais métodos são o de Mantel Haenszel e o Método Padronizado, ambos usados pelo ETS (Educational Testing Service).

2. Proposta do Projeto

Para os métodos citados acima se pretende estudar alguns deles e avaliar a precisão ao detectar o DIF. Para isso os dados necessários serão simulados e avaliada a proporção de vezes que o teste detecta a presença de DIF, existindo ou não. Também poderá se notar qual a diferença necessária entre os parâmetros para que cada teste detecte a presença do DIF.

Nesta primeira parte serão avaliados o Teste de Mantel Haenszel e o Teste Padronizado.

3. Metodologia de Trabalho

Para avaliar os métodos de detecção de DIF, as simulações foram efetuadas com o software R. Em um estudo preliminar, foram simuladas as habilidades de 4000 alunos, divididos em dois grupos de mesmo tamanho. As habilidades têm distribuição normal com média 0 e desvio-padrão 1, para vários itens (65). A discriminação de cada um deles foi atribuído uma distribuição log-normal(0,1). Para a dificuldade de cada um dos grupos, foi utilizado uma distribuição normal (0,2). O parâmetro de acerto ao acaso é a média da distribuição Beta(5,17). O valor de Alfa determina a diferença de dificuldade entre os grupos para um determinado item. Os indivíduos foram separados em níveis de habilidade para a aplicação dos testes, já que o conceito de DIF envolve a diferença da probabilidade de acerto para indivíduos com o mesmo nível de habilidade.

Em uma simulação inicial, as dificuldades de cada item foram simuladas separadamente, com Alfa constante para cada item. Todos os itens são avaliados pelos dois testes citados. Após as simulações foi obtido a proporção de vezes que cada teste detecta a presença de DIF para o item.

3.1 Teste de Mantel Haenszel

Para cada nível de habilidade dos alunos tem-se a tabela abaixo. O teste busca comparar as freqüências observadas e esperadas, tanto para acertos quanto para os erros dos grupos.

Tabela de frequências para o nível j de um item

	Acertos	Erros	Totais
Grupo 1	A_j	B_j	n_{R_j}
Grupo 2	C_j	D_j	n_{F_j}
Total	m_{1_j}	m_{0_j}	T_j

O teste apresenta como hipótese nula que os erros e acertos são independentes para cada grupo respondente (não ocorre DIF). Foram considerados casos de detecção de DIF quando o teste apresentou o p-valor < 0.05.

3.2 Teste Padronizado

Calcula um índice de diferença entre os itens baseado nas probabilidades de acerto amostrais para cada nível de habilidade.

$$STD\ pDIF = \frac{\sum_m w_m (P_{fm} - P_{rm})}{\sum_m w_m}$$

Para os cálculos foi considerado W_m igual ao número total de indivíduos no nível m, outros valores podem ser atribuídos a esta peso. P_{fm} é a probabilidade de acerto estimada para o grupo focal de nível m, P_{rm} é a probabilidade correspondente ao grupo de referência.

O índice pode variar entre -1 e 1. Foram considerados casos de detecção de DIF quando o índice teve valor absoluto maior que 0,10. O que é considerado um DIF severo (Dorans & Holland, 1993).

4. Exemplo de Aplicação

Este trabalho foi baseado em estudos preliminares, com os testes citados acima avaliados através de técnicas de simulação. Como exemplo, no Gráfico 1 podemos ver as curvas de poder estimadas para cada um dos testes. Alfa é a diferença entre as dificuldades dos itens. Nota-se que para grandes valores de Alfa temos grandes percentuais de detecção do DIF. Para valores próximos de 0 os testes não detectam a presença do DIF, conforme esperado.

Sabe-se que quando Alfa é positivo o item favorece o Grupo 1, neste caso os dois testes tem percentuais diferentes para a detecção do DIF. É notável que principalmente quando o Grupo 2 é favorecido (Alfa negativo) os testes tem percentuais de detecção distintos, o teste padronizado se mostra mais poderoso.

Item 1

A primeira análise é de um item que favorece o Grupo 1, pois tem um nível de dificuldade maior para o Grupo 2. O parâmetro de discriminação A é razoavel e o item apresenta baixa probabilidade de acerto ao acaso (C) . O item apresenta um grande comportamento diferencial (DIF), podemos notar pela diferença das dificuldades (Alfa=0,957). Devido a este valor o teste padronizado apresentou um índice de 0,14, o que significa DIF severo. Coerentemente, o teste de Mantel Haenszel apresentou P-Valor <0,05, o que leva a rejeição da hipótese nula de independência dos acertos e erros dos alunos, ou seja, comportamento diferencial do item.

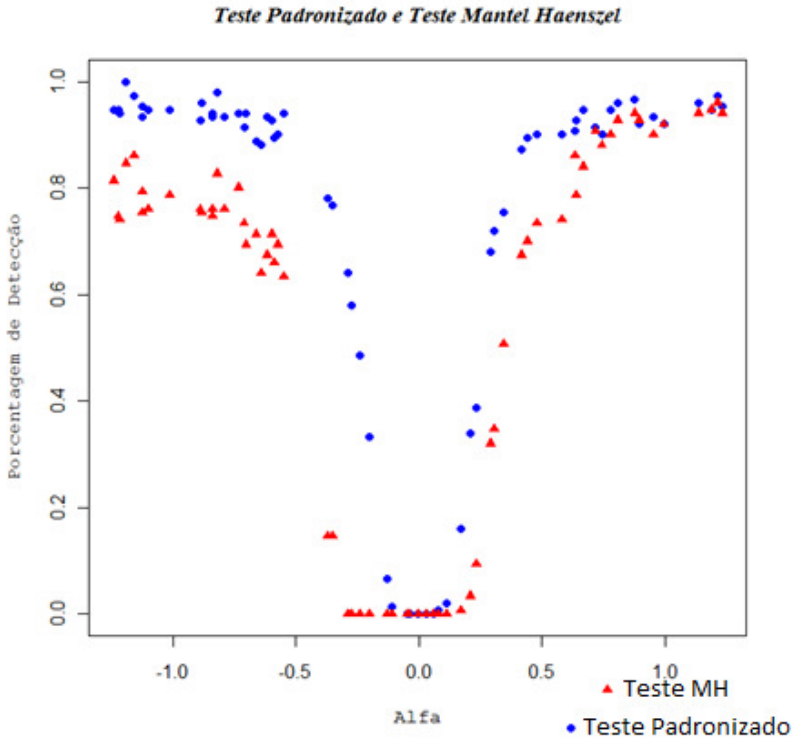
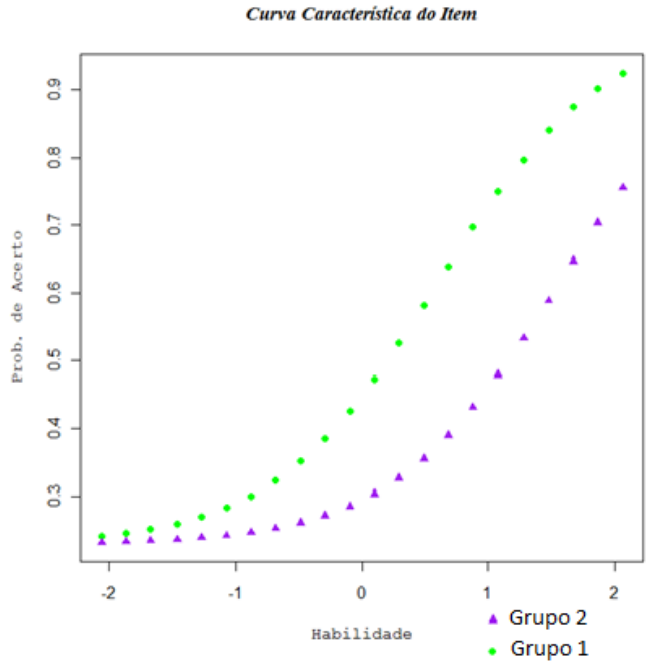


Gráfico 1- Curvas de poder estimadas para os testes.



A	0.9019831
B1	0.5969211
B2	1.553685
C	0.2275
ALFA	0.9567644
MHPvalue	0.03699583
DIFPAD	0.1409512

Gráfico 2- Curva Característica do Item 1

5. Observações Finais

Para os próximos estudos pretende-se medir também o DIF nos parâmetros A e C, aqui foi avaliado apenas o DIF no parâmetro B. Serão também implementados outros testes, como por exemplo, o Mantel Haenszel generalizado. Tais implementações serão apresentadas no 19º SINAPE.

6. Referências Bibliográficas

1. Osterlind, S.J, Everson, H. T. Differential Item Functioning.
2. SOARES, Tufi Machado; GAMERMAN, Dani and GONCALVES, Flávio Bambirra. Análise bayesiana do funcionamento diferencial do item, 2007.
3. Andriola , W. B. Descrição dos Principais Métodos para Detectar o Funcionamento Diferencial dos Itens (DIF), 2001.

7. Agradecimentos

Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG