Análise do Comportamento Diferencial dos Itens de Geografia: estudo da 4ª série avaliada no Proeb/Simave 2001

TUFI MACHADO SOARES[1]
Professor do Programa de Mestrado em Educação
Pesquisador do Centro de Políticas Públicas e Avaliação da Educação (CAEd)
Universidade Federal de Juiz de Fora – UFJF
tufi@caed.ufjf.br

SILENE FELIZARDO DE MENEZES GENOVEZ
Aluna do Programa de Mestrado em Educação
Universidade Federal de Juiz de Fora - UFJF
silene@powermail.com.br

AILTON FONSECA GALVÃO
Analista do Centro de Políticas Públicas e Avaliação da Educação (CAEd)
ailton@caed.ufjf.br

Resumo
Apresenta-se uma análise do comportamento diferencial (DIF) dos itens de geografia, aplicados aos alunos da 4ª série no Proeb-2001 nas diferentes regiões do Estado de Minas Gerais. *Grosso modo*, os resultados sugerem que itens relacionados a questões ambientais são mais fáceis para os alunos da região metropolitana de Belo Horizonte do que para os alunos do interior do Estado. Por outro lado, os itens que avaliam a relação entre o espaço urbano e o espaço rural são mais fáceis para os alunos do interior.
Palavras-chave: avaliação educacional, teoria da resposta ao item, análise do comportamento diferencial.

Resumen
Se presenta un análisis del comportamiento diferencial (DIF) de los ítens de geografía, aplicados a los alumnos del 4º grado en el Proeb-2001, en las distintas regiones del estado de Minas Gerais. En líneas generales, los resultados sugieren que ítens relacionados a cuestiones ambientales les resultan más fáciles a los alumnos del área metropolitana de Belo

[1] Os autores agradecem ao CAEd e à Secretaria Estadual de Educação pelo apoio e pela cessão dos itens apresentados.

Horizonte que a los alumnos del interior del Estado. Por otra parte, los ítens que evalúan la relación entre el espacio urbano y el espacio rural les resultan más fáciles a los alumnos del interior.
Palabras-clave: evaluación educacional, teoría de la respuesta al ítem, análisis del comportamiento diferencial.

Abstract
This article presents a differential (DIF) analysis for geography items in Proeb-2001. The groups in the analysis are composed of 4th grade students of different geographic regions in the state of Minas Gerais. The results suggest that the items associated with environmental issues are easier for students from the Belo Horizonte metropolitan area while items which compare urban and rural spaces are easier for students from the inland.
Key words: educational assessment, item response theory, DIF analysis.

1 INTRODUÇÃO

O Sistema Mineiro de Avaliação da Educação – Simave, criado por resolução da Secretaria de Estado da Educação em 2000, tem implementado o Programa de Avaliação da Rede Pública de Educação Básica – Proeb. Esse programa de avaliação em larga escala tem por objetivo produzir sistematicamente informações sobre o desempenho dos alunos e colocá-las à disposição do público. Os testes são aplicados a todos os alunos da 4ª e da 8ª série do ensino fundamental e do 3º ano do ensino médio da rede estadual e da rede municipal que aderiram ao sistema. O programa inclui dois outros instrumentos importantes para o processo de avaliação: questionário do aluno, para obter dados sobre o perfil socioeconômico e trajetória escolar dos estudantes e questionário dos professores e diretores da escola, para traçar o perfil dos profissionais da educação. A associação dos dados coletados nesses dois instrumentos possibilita a identificação dos chamados fatores contextuais associados ao desempenho. Inicialmente, o Proeb foi previsto para ciclos de avaliação de dois anos. Em 2000, foram avaliadas competências em Língua Portuguesa (leitura) e Matemática; em 2001, Ciências Humanas e Ciências da Natureza; em 2002, Língua Portuguesa voltou a ser avaliada; em 2003, Matemática, novamente.

O Proeb utiliza um teste de múltipla escolha e não procura saber o que cada aluno aprendeu individualmente, mas quais conteúdos a unidade escolar e o sistema educacional foram capazes de ensinar ao conjunto de seus alunos, podendo, conseqüentemente, avaliar se estes estão cumprindo a função de democratização e acesso ao conhecimento, e verificar o grau de desigualdade existente dentro do sistema educacional. Assim, o objetivo da avaliação não deve ser apenas o de constatar eventuais diferenças, mas de interpretá-las adequadamente para que se possa tomar decisões eficazes no processo de ensino/aprendizagem, além de subsidiar políticas educacionais que conduzam à democratização e à qualidade de ensino.

Os itens dos testes do Proeb foram produzidos com base na proposta curricular de Minas Gerais e nas matrizes utilizadas pelo Sistema Nacional de Avaliação da Educação Básica (Saeb). Cada item do teste está associado a uma competência específica. Para cada competência há um determinado número de itens. Uma vez elaborados, os itens são pré-testados, selecionando-se aqueles que demonstram oferecer mais informação sobre o aprendizado do aluno.

Atualmente, na área educacional, vem sendo utilizada a Teoria da Resposta ao Item – TRI que consiste em modelos para tratamento de itens a respeito de variáveis latentes, os quais relacionam a probabilidade de um aluno responder de forma correta e suas habilidades na área do

conhecimento avaliada, as quais não são observadas diretamente. A TRI permite a comparação entre grupos diferentes, desde que os modelos dos itens sejam todos conhecidos e estejam na mesma escala ou que haja itens comuns aos testes aplicados a esses grupos. O modelo mais utilizado é o modelo logístico, unidimensional, de 3 parâmetros, para itens de múltipla escolha e respostas dicotômicas (do tipo certo, representadas por $Y = 1$, e errado, representadas por $Y = 0$), dado por:

$$P(Y = 1 | \theta) = c + (1 - c) \frac{1}{1 + e^{-1.7a(\theta - b)}}$$

1) O parâmetro b é o parâmetro dificuldade do item – que representa o grau de dificuldade apresentado pelo item;
2) a é o parâmetro de discriminação do item – que se associa à capacidade do item de distinguir alunos com diferentes níveis de habilidade;
3) c é o parâmetro de acerto casual – que se associa à probabilidade de acerto ao acaso do item.

As medidas provenientes do Proeb permitem uma boa comparação dos resultados entre agregados de alunos. Assim, pode-se comparar o desempenho das diversas unidades escolares, como também o desempenho de alunos de diferentes regiões do Estado. Um pressuposto importante de qualquer modelo para avaliação educacional que garanta a comparabilidade dos resultados é que o item apresente o mesmo comportamento nos diversos grupos populacionais que estão sendo avaliados. Quando os modelos da TRI estão sendo utilizados, essa questão do comportamento se traduz na estabilidade dos parâmetros dos modelos dos itens para as diferentes populações. No entanto, embora em grau elevado o DIF possa prejudicar a comparabilidade dos resultados, quando moderado e localizado em poucos itens, o DIF além de, praticamente, não afetar a proficiência produzida pode, se devidamente analisado, trazer informações importantes sobre diferenças curriculares e diferenças socioculturais, por exemplo, entre as regiões.

Este trabalho teve o objetivo de analisar o comportamento diferencial dos itens de geografia aplicados no Proeb-2001, nas diferentes regiões do Estado de Minas Gerais reunidas nos Pólos Regionais, que constituem uma organização regional de Secretarias Regionais de Ensino às quais as escolas se subordinam.

2 CARACTERÍSTICAS DA ESCALA DE GEOGRAFIA.

Segundo o relatório técnico do Simave[2], o teste de geografia, do 1º ano do ciclo intermediário (4ª série) do ensino fundamental, procurou investigar a situação dos alunos em relação à compreensão das categorias básicas de leitura e interpretação do espaço geográfico, com prioridade para os mecanismos que envolvem as operações de orientação, localização e representação. A maior parte dos itens concentrou-se na avaliação das condições gerais de alfabetização cartográfica. Buscou-se avaliar o entendimento que o aluno tem do mundo, pela sua compreensão das relações de produção e transformação do espaço, bem como a compreensão dos objetos, independente de si mesmo (descentração), e sob os diferentes pontos de vista segundo os quais podem ser representados.

3 COMPORTAMENTO DIFERENCIAL (DIF): O QUE É E COMO PODE SER MEDIDO?

O objetivo de uma análise de DIF é verificar se um item tem ou não o mesmo comportamento para indivíduos pertencentes a dois grupos distintos, mas de mesma habilidade cognitiva. Em geral, deseja-se verificar se um item apresenta graus de dificuldade diferentes para subgrupos da população que têm o mesmo nível de conhecimento.

Grande parte das causas determinantes do DIF ainda é desconhecida. Estudos conduzidos pelo *Educational Testing Service – ETS*, nos Estados Unidos, apontam que o DIF pode ser causado, basicamente, por uma tricotomia de fatores: a **familiaridade** com o conteúdo do item, que também pode ser associada à exposição do tema ou a um fator cultural; o **interesse** pessoal naquele dado conteúdo e a **reação emocional negativa** provocada pelo conteúdo (Stricker, Emmerich, 1999).

Segundo Elliot *et al* (2002), no caso específico do Brasil, "tendo em vista a dimensão continental do país e as peculiaridades de cada uma de suas regiões que, certamente se refletem na vida social, econômica e cultural da população, incluindo nela a educação e suas formas de dinamização nas escolas, o comportamento de um item pode diferir porque, em seu enunciado, ilustrações e alternativas de respostas aparecem": temas regionais, mais familiares em determinadas regiões do que em outras; características lingüísticas, como termos, expressões e gírias locais usados em algumas regiões, mas não em todas; fatos ocorridos em

[2] Simave – Minas Gerais: avaliação da educação – Ciências Humanas/Ciências da Natureza, julho de 2002.

um estado/região e, portanto, neles mais conhecidos; nomes/palavras que associam a resposta certa do item a algum aspecto específico da região; temas provavelmente mais focalizados pelo ensino de uma região; temas que possivelmente não são igualmente explorados nos currículos das cinco regiões, por diferença de ênfase, que em essência se classificam segundo a familiaridade com o conteúdo.

Grosso modo, pode-se entender que um item apresente comportamento diferencial (com respeito à sua dificuldade) entre dois grupos específicos de indivíduos - por exemplo, entre negros e brancos, entre alunos do sexo masculino e feminino, etc – quando estes forem agrupados sistematicamente em grupos de mesma habilidade cognitiva (pareamento) e, mesmo assim, as probabilidades de acerto do item (caso de itens dicotômicos) forem significativamente diferentes para os grupos pareados. De fato, este é um tipo específico de comportamento diferencial que se associa ao grau de dificuldade do item. Outros tipos menos comuns podem ser também analisados, como o comportamento diferenciado com respeito à capacidade de discriminação do item e do acerto casual.

Para efeito de análise, quando se está comparando o desempenho de um item em dois grupos diferentes, um deles denomina-se grupo de Referência (R), e o outro é denominado grupo Focal (F). Normalmente, quando se têm vários grupos diferentes, pode-se escolher um deles como o grupo de referência, por alguma razão em particular, e realizar a análise comparativa do comportamento do item nos demais grupos em relação ao comportamento nesse grupo, mas também é possível que se deseje realizar a análise comparativa entre todos os grupos. Este estudo considerou o primeiro caso, que se mostrou bastante satisfatório.

4 IDENTIFICAÇÃO DOS ITENS DE GEOGRAFIA COM COMPORTAMENTO DIFERENCIAL

A identificação dos itens que apresentaram DIF foi realizada em duas etapas. Na primeira etapa, por meio do *software* BILOG-MG e do procedimento descrito no anexo, foram identificados os itens que apresentavam algum comportamento diferencial significativo com respeito à dificuldade, utilizando-se modelos da TRI. O grupo de referência adotado foi o correspondente ao pólo regional identificado como Capital, que agrega a região metropolitana de Belo Horizonte e cercanias. Os demais grupos considerados foram os pólos: Norte, Centro-sul, Triângulo e Zona da Mata, que correspondem às respectivas regiões geográficas do Estado de Minas Gerais. Dos 85 itens de geografia do teste foram identificados os 16

itens a seguir que, aparentemente, apresentaram algum comportamento diferenciado mais relevante. Para esses itens são exibidas na Tabela 1 as diferenças nos parâmetros de dificuldade dos modelos da TRI estimados para cada região, e o diagnóstico do comportamento diferencial encontrado:

Tabela 1: Itens que Apresentaram Comportamento Diferencial

Modelo para análise do comportamento diferencial por região

Item	Diferenças nos parâmetros de dificuldades				Diagnósticos
	1 – 5	2 – 5	3 –5	4 -5	
H04010MG	0.015 0.079	0.036 0.069*	0.702 0.123*	-0.771 0.068*	Item mais fácil para a Zona da Mata e mais difícil para o Triângulo em relação à capital.
H04038MG	-0.233 0.071*	-0.331 0.071*	-0.499 0.087*	-0.212 0.074*	Item mais fácil para todas as regiões do que para a capital.
H04013MG	0.119 0.072*	0.109 0.062*	-0.467 0.075*	0.687 0.084*	Item mais fácil para o Triângulo e mais difícil para a Zona da Mata em relação à capital.
H04079MG	-0.366 0.082*	-0.214 0.085*	0.067 0.103*	-0.151 0.089	Item mais fácil para o Centro-sul, Zona da Mata e principalmente para o Norte do que para a capital.
H04096MG	-0.170 0.048*	-0.185 0.046*	-0.148 0.059*	-0.264 0.050*	Item mais fácil para todas as regiões do que para a capital.
H04127MG	0.036 0.151*	-0.179 0.140*	-0.313 0.167*	-0.186 0.149*	Item mais fácil para todas as regiões, exceto a região Norte, do que para a capital.
H04194MG	0.044 0.121*	0.023 0.120*	0.315 0.132*	-0.149 0.124*	Item mais difícil para o Triangulo e mais fácil para a Zona da Mata do que para a capital.
H04089MG	-0.261 0.063*	-0.208 0.063*	-0.233 0.076*	-0.136 0.066*	Item mais fácil para todas as regiões quando comparadas à capital.
H04279MG	0.147 0.054*	0.185 0.055*	0.247 0.066*	0.209 0.057*	Item mais difícil para todas as regiões quando comparadas à capital.
H04111MG	-0.456 0.105*	-0.266 0.105*	-0.365 0.125*	-0.253 0.110*	Item mais fácil para todas as regiões quando comparadas à capital.
H06010MG	0.225 0.064*	0.167 0.063*	0.283 0.076*	0.111 0.066*	Item mais difícil para todas as regiões quando comparadas à capital.
H06038MG	0.183 0.071*	0.144 0.067*	0.252 0.086*	0.153 0.072*	Item mais difícil para todas as regiões quando comparadas à capital.
H06050MG	-0.232 0.096*	-0.201 0.092*	-0.093 0.108*	-0.170 0.097*	Item mais fácil para quase todas as regiões quando comparadas à capital.
H06026MG	0.325 0.081*	0.176 0.079*	0.234 0.093*	0.211 0.083*	Item mais difícil para todas as regiões quando comparadas à capital.
H04236MG	-0.216 0.125*	-0.219 0.112*	-0.432 0.134*	-0.442 0.114*	Item mais fácil para todas as regiões quando comparadas à capital.
H06031MG	-0.236 0.051*	-0.171 0.048*	-0.183 0.064*	-0.224 0.052*	Item mais fácil para todas as regiões quando comparadas à capital.

* Erro-padrão da diferença.

1. Norte; 2. Centro-Sul; 3. Triângulo; 4. Zona da Mata; 5. Capital.

Após serem identificados os itens acima, que apresentavam algum tipo de comportamento diferencial, foram calculadas as estatísticas clássicas descritas no anexo e produzidos gráficos comparativos das respostas atribuídas pelos 5 grupos. Essas estatísticas e os gráficos foram produzidos por meio do *software* SisAni (Sistema de Análise de Itens), desenvolvido pela equipe de Estatística do CAEd/UFJF), e teve por objetivo confirmar ou não o DIF para os itens indicados anteriormente e instrumentar a análise das possíveis causas do comportamento diferencial observado (Soares, Galvão, Genovez, 2004). Dentre essas estatísticas está a estatística alfa (delta) de Mantel Haenzel (ver o anexo para a definição dessa estatística) que permite analisar a intensidade do comportamento diferencial apresentada pelo item. O sistema SisAni permite que essas análises mais acuradas sejam produzidas. Para efeito de classificação dos itens quanto ao DIF apresentado utiliza-se, neste trabalho, a seguinte regra:

Quadro 1: Classificação do Grau do Comportamento Diferencial

Valores da estatística	Grau do Comportamento Diferencial
$Abs(AlfaD\ MH) \leq 0.5$	DIF insignificante
$0.5 < Abs(AlfaD\ MH) \leq 1.0$	DIF pequeno
$1.0 < Abs(AlfaD\ MH) \leq 1.5$	DIF intermediário
$1.5 < Abs(AlfaD\ MH)$	DIF alto

Esse critério será empregado nas análises descritas a seguir.

5 INTERPRETAÇÃO E ANÁLISE DO COMPORTAMENTO DIFERENCIAL DOS ITENS DE GEOGRAFIA DA 4ª SÉRIE

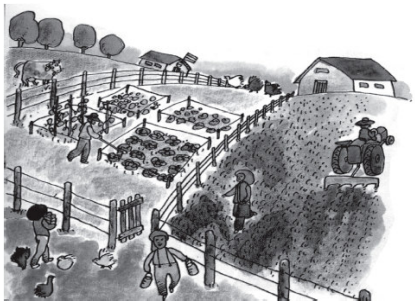
Dentre outras, na elaboração de um item para um teste de uma avaliação em larga escala, que utiliza modelos dicotômicos para produção da proficiência, deve-se levar em consideração a idade e a escolaridade do grupo no qual será aplicado o teste. O item deve ser objetivo e claro, para que não deixe dúvidas quanto à alternativa correta e não deve possuir dicas que indiquem a resposta. Finalmente, as questões não devem conter termos ou vocabulários que favoreçam mais um grupo em detrimento de outro. Normalmente, os itens que apresentam defeitos muito graves são

identificados e excluídos em pré-testes e análises estatísticas preliminares, antes de utilizá-los para a produção da proficiência do aluno; nesse caso se encontram os itens que apresentam comportamento diferencial muito elevado. Portanto, não se espera, em princípio, que haja itens com comportamento diferencial muito elevado. No entanto, ainda assim, alguns itens exibem algum grau de comportamento diferencial, como é o caso dos itens identificados na seção 4. Esse grau de comportamento diferencial não afeta o resultado da avaliação, pois praticamente não interfere no resultado da proficiência estimada, mas pode trazer alguma informação adicional que seja relevante para entender algumas das possíveis diferenças pedagógicas e/ou algumas possíveis diferenças devido às características regionais ainda não percebidas.

Inicialmente, observa-se que os itens H04038MG, H04096MG, H04127MG, H04089MG, H04111MG, H06050MG e H06031MG avaliam o conhecimento sobre a relação entre o espaço urbano e o espaço rural, discriminando os produtos do campo dos produtos da cidade. Em todos esses itens obteve-se o mesmo resultado: são itens mais fáceis para os alunos de todas as regiões, quando comparados aos alunos do pólo Capital. Provavelmente, porque, em geral, os alunos dessas regiões apresentam maior conhecimento e facilidade de acesso ao espaço rural do que os alunos da capital.

Ao se analisar, por exemplo, a imagem da questão H04038MG observam-se alguns elementos típicos da paisagem rural, como, por exemplo, os animais (galinhas e boi), plantações, etc.

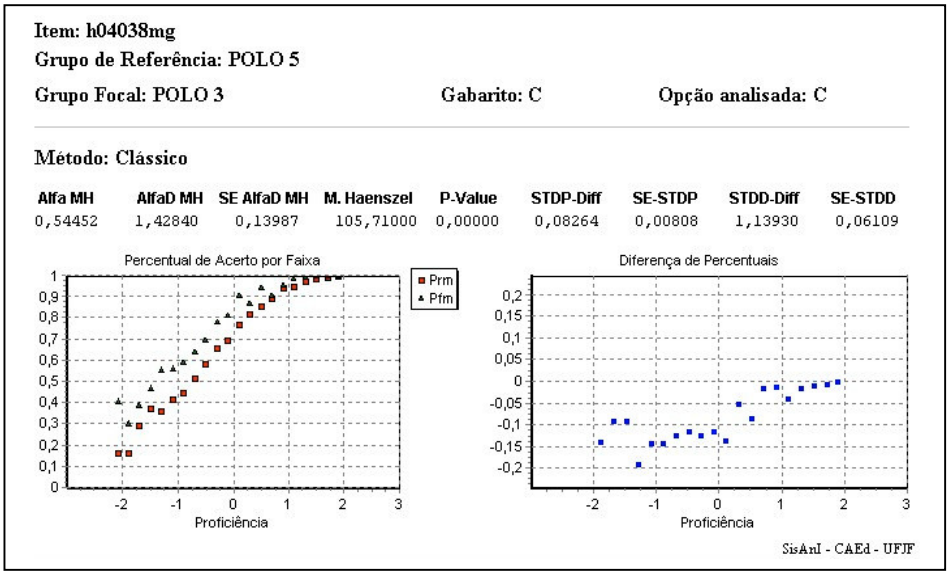
(PROEB-2001) Observe o desenho e responda às duas questões seguintes:



(H04038MG) O desenho mostra uma realidade que se relaciona a um espaço:

A) Urbano.
B) Industrial.
C) Rural.
D) Metropolitano.

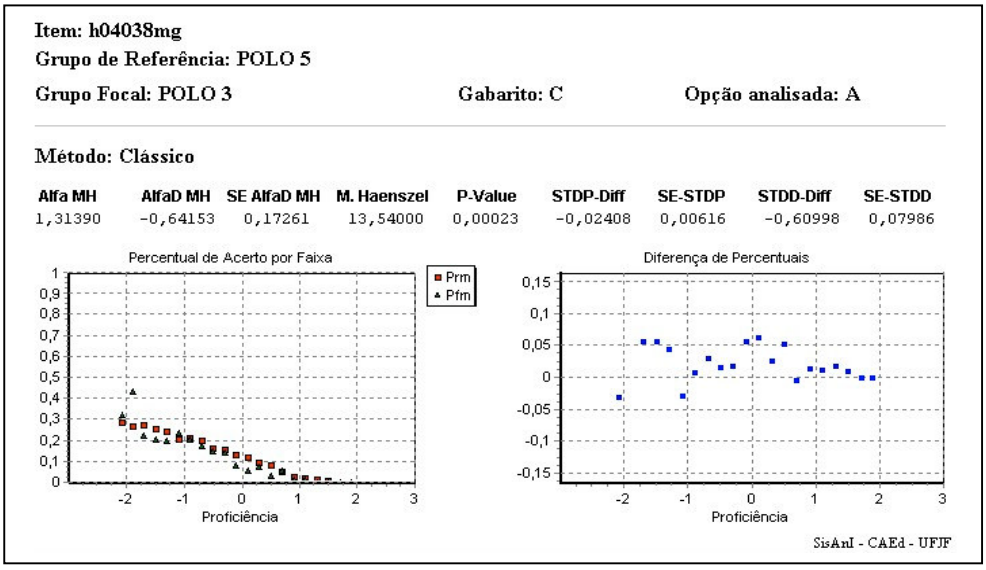
A seguir, apresentam-se os resultados obtidos para a comparação do comportamento do item na escolha da opção C – resposta correta, entre os grupos 3 (focal - pólo Triângulo) e 5 (referência – pólo capital) do item H04038MG. São incluídas as estatísticas clássicas (apresentadas no anexo) e os gráficos que mostram o comportamento ao longo da escala de proficiências. Os resultados para todas as regiões e as opções de respostas podem ser encontrados no Anexo 1 de Soares, Galvão e Genovez (2004).



Nota-se, inicialmente, que o item apresenta uma dificuldade maior para os alunos da capital – grupo de referência – confirmando o resultado já encontrado na comparação dos parâmetros de dificuldade estimados para os modelos de três parâmetros da TRI desse item, para ambos os grupos, e apresentados na Tabela 1. Observando-se o valor da estatística delta de Mantel Haenszel (alfa D MH = 1.42) pode-se classificar esse comportamento diferenciado como de grau intermediário e, pela análise gráfica, observa-se que ele é mais ou menos uniforme entre os níveis de proficiência –1 a 1, não apresentando, aparentemente, alteração na sua discriminação, nem no seu acerto casual. Como se pode notar pelos dois gráficos, a diferença entre os percentuais de acerto para os alunos desses dois grupos chega a alcançar 0,20.

Analisando as diferenças entre as respostas atribuídas à opção A, nota-se que os alunos do grupo de referência escolheram essa opção mais

freqüentemente do que os alunos do grupo focal e, de fato, essa característica pode ser também observada, e praticamente na mesma proporção, para as demais opções de respostas. Assim, o comportamento diferenciado, que se reflete também na freqüência de escolha das demais opções (além da opção correta), ocorre, praticamente, na mesma proporção para todas elas, não caracterizando, assim, que exista uma opção como a mais procurada pelos alunos do grupo de referência.



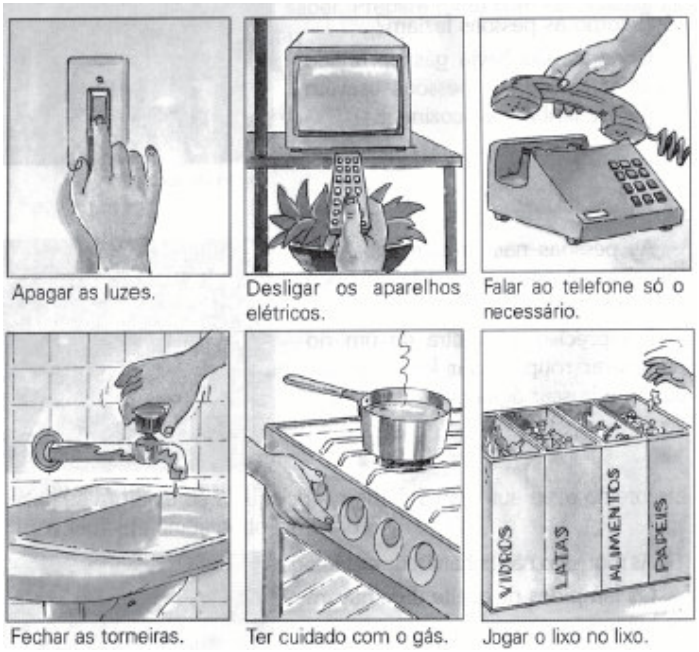
As mesmas conclusões são observadas para todos os demais pólos regionais. Embora não sejam apresentadas aqui, as análises dos itens H04038MG, H04096MG, H04127MG, H04089MG, H04111MG, H06050MG e H06031MG são praticamente as mesmas e como esses itens estão todos associados ao mesmo tipo de competência as conclusões podem ser empiricamente generalizadas.

Observa-se a necessidade, então, de se trabalhar melhor a diferenciação entre o espaço urbano e o rural, tanto do ponto de vista conceitual quanto o de identificação visual, para os alunos de áreas metropolitanas. Isso pode ser feito realizando excursões a sítios e fazendas, o que garantirá o contato direto do aluno com o meio ambiente, quando ele terá a oportunidade de fazer questionamentos e críticas com relação às diferenças das condições de vida urbana e rural por meio da observação, da comparação, do registro e da descrição do meio. Alternativamente, recomenda-se a utilização de vídeos, revistas e jornais com gravuras

adequados e devidamente avaliados pelo pedagogo. De qualquer forma, acredita-se que se deve procurar sempre a valorização da experiência do aluno, trabalhando-a de forma mais concreta nas séries iniciais, quando o aluno tem maior dificuldade em assimilar conceitos puramente abstratos. Claro que essas são apenas algumas idéias, outras interpretações e sugestões poderiam, naturalmente, ser apresentadas por especialistas no assunto.

Observa-se que os itens H04279MG, H06010MG e H06026MG também apresentam características comuns, pois trabalham assuntos relacionados com as questões ambientais tipicamente urbanas. Como a visibilidade desses problemas ambientais é maior nas grandes cidades, pois é onde se debatem mais esses temas, há uma facilidade um pouco maior por parte dos alunos do pólo capital em desenvolverem questões relacionadas a esse tema. Como exemplo, tome-se o item a seguir:

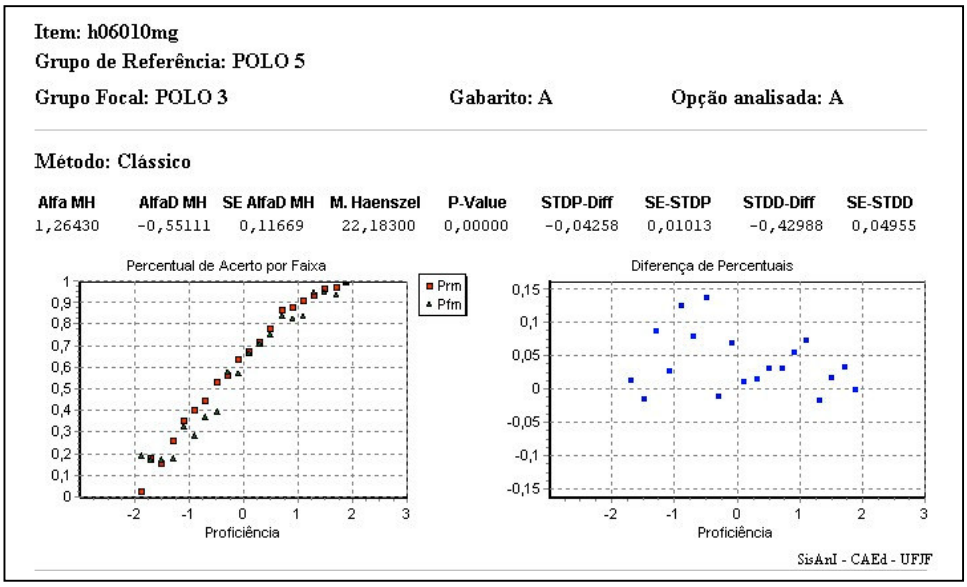
" (H06010MG) Observe as imagens:



Qual a conseqüência de medidas como as mostradas nas imagens?

A) Economia doméstica e preservação de recursos naturais.
B) Diminuição do conforto e destruição ambiental.
C) Aumento dos gastos com serviços e aumento do volume de lixo urbano.
D) Má qualidade de vida e destruição do espaço urbano."

Esse item apresentou, para todas as regiões, um percentual de acerto menor do que o observado para a região metropolitana; no entanto, as diferenças são bastante pequenas como pode ser verificado nos resultados seguintes:

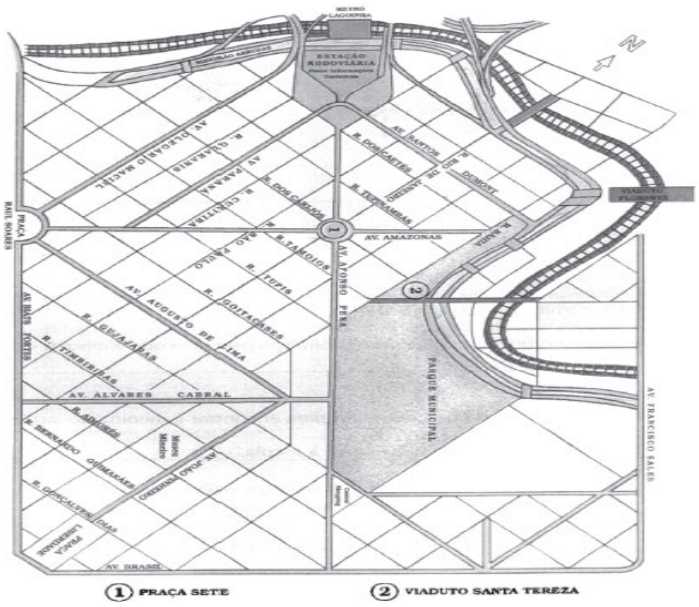


Portanto, há a necessidade de se trabalhar melhor as questões ambientais com os alunos, no interior, por meio de trabalhos práticos que estejam relacionados à reciclagem de lixo, ao racionamento de água e à energia, etc., procurando enfatizar sempre a relação homem-espaço, a forma como o homem produz a tecnologia, a utilização que se faz dela e as conseqüências deste uso. É importante apresentar aspectos da utilização dos recursos naturais renováveis e não-renováveis, a relação consumo humano e o meio ambiente. Sempre é interessante trabalhar as questões ligadas ao meio ambiente dentro de uma visão ampla, crítica e participativa, em que o aluno se veja como o sujeito principal, ao mesmo tempo responsável pelos problemas ambientais e vítima das conseqüências da degradação do meio ambiente (devido à sua ação e à de outros elementos sociais) e, por fim, como um agente transformador da relação homem/meio ambiente.

Finalmente, os 3 itens seguintes exibem comportamento diferencial relevante, porém não facilmente esclarecido. Duas dessas questões parecem ter sido mal formuladas o que pode ter provocado esse tipo de comportamento. São elas:

1)

"(H04013MG) Observe a planta abaixo. Ela representa o centro de Belo Horizonte, capital de Minas Gerais.



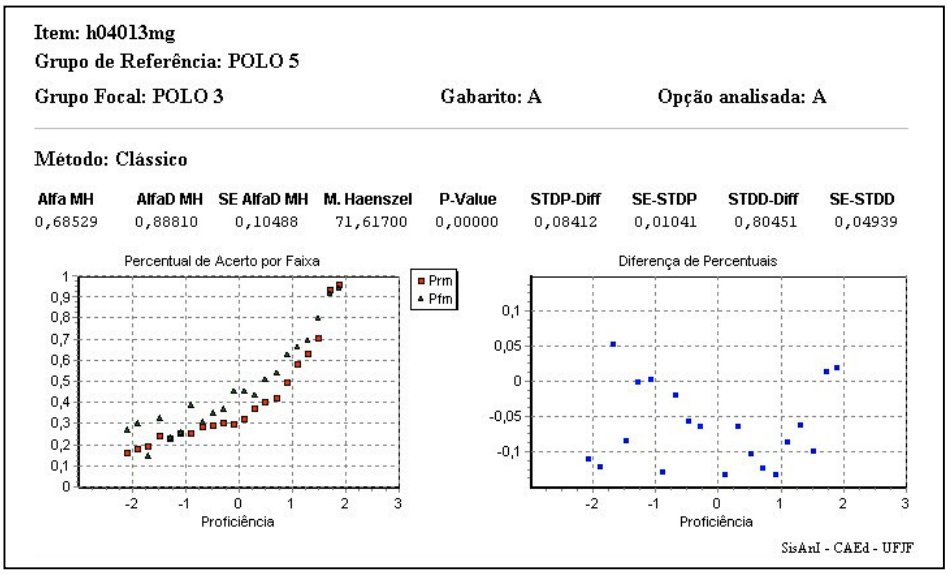
O centro é:
A) Um bairro.
B) Uma rua.
C) Um parque.
D) Uma estação rodoviária."

Esta é uma questão que pode ser resolvida sem a imagem apresentada. Por outro lado, a imagem causa confusão na resposta do aluno. Apesar de a pergunta estar se referindo ao centro, o ponto de destaque desta imagem é a estação rodoviária e, de fato, boa parte dos alunos optou pela opção D, como pode ser visto na tabela abaixo:

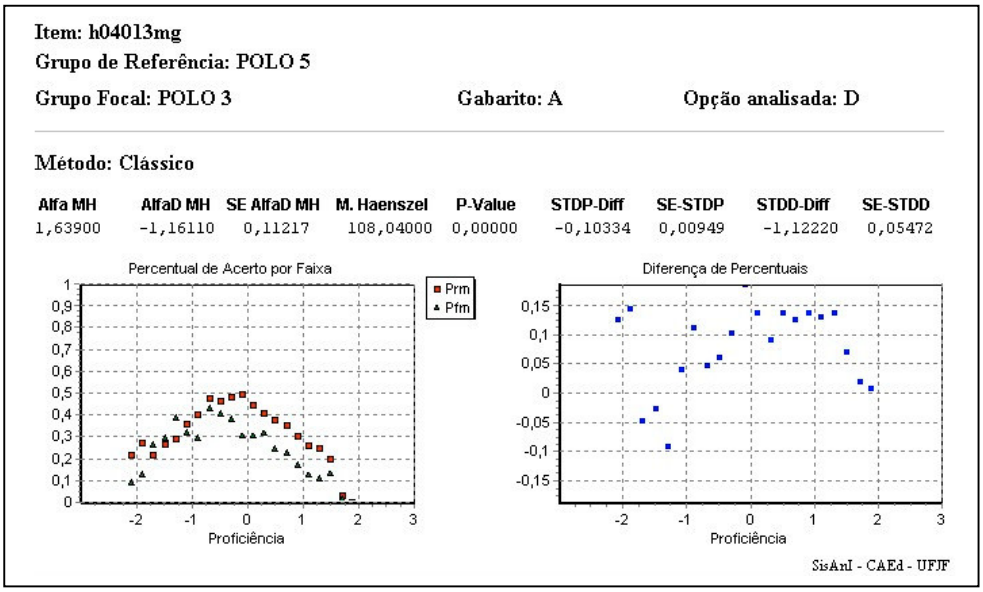
Tabela 2

Opção	Freqüência de respostas	Percentual
A	13831	36,6%
B	5881	15,6%
C	4363	11,6%
D	13319	35,3%
inválidas	425	0,9%

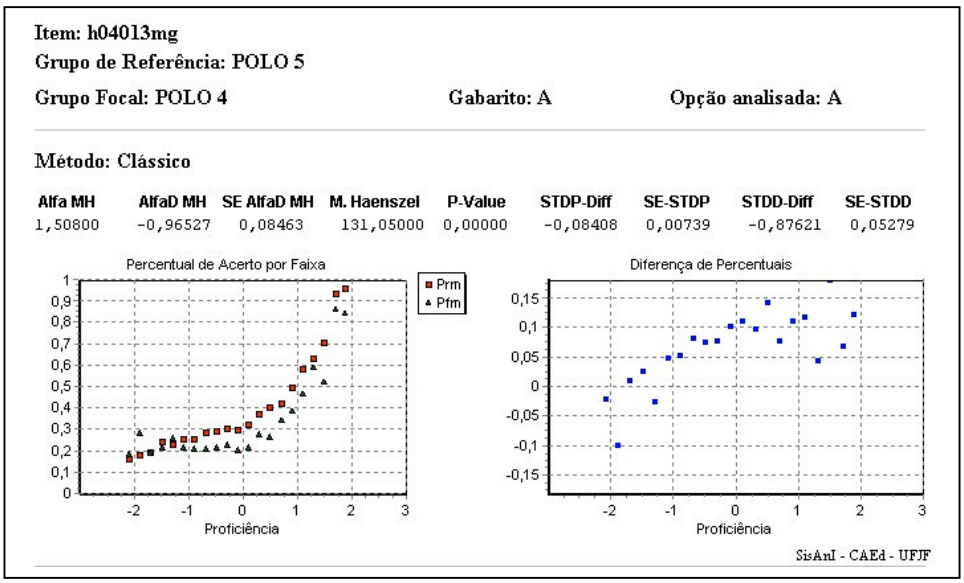
Observando-se o comportamento de cada região, nota-se que para os alunos do Triângulo o item foi um pouco mais fácil do que para os alunos do pólo capital:



Sendo que os alunos do pólo capital optaram com maior freqüência pela opção D do que os alunos do Triângulo:



Em contrapartida, o item foi um pouco mais difícil para os alunos da Zona da Mata:



Eles escolheram mais as opções B (uma rua) e C (um parque), talvez porque o centro da cidade de Juiz de Fora (a maior cidade da Zona da Mata) seja, muitas vezes, identificado por uma rua específica (a rua Halfeld) ou por um parque (o parque Halfeld).

2)

"(H04194MG) Observe a figura abaixo e responda:

Figura 1 Figura 2 Figura 3

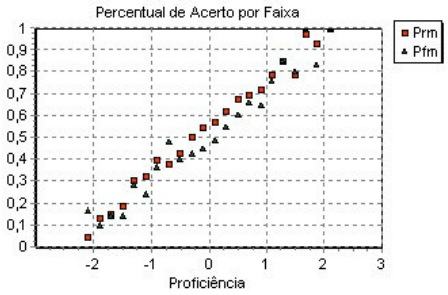
Que elementos aparecem nas 3 figuras?
A) Prédio e torre.
B) Prédio menor e várias casas.
C) Rio e mata.
D) Mar e montanhas."

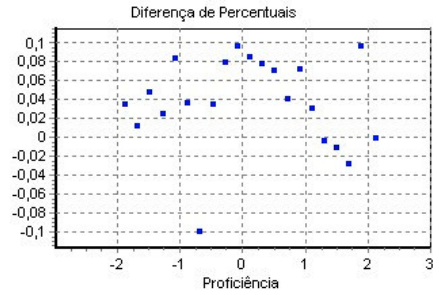
Nesse item a imagem não está muito nítida. A questão também não está muito clara, pois o que é prédio menor? A questão deixa dúvidas principalmente com relação às opções A e B como pode ser observado pelos resultados a seguir:

Item: h04194mg
Grupo de Referência: POLO 5
Grupo Focal: POLO 3 **Gabarito: A** **Opção analisada: A**

Método: Clássico

Alfa MH	AlfaD MH	SE AlfaD MH	M. Haenszel	P-Value	STDP-Diff	SE-STDP	STDD-Diff	SE-STDD
1,25050	-0,52532	0,10682	24,09700	0,00000	-0,04833	0,01039	-0,45934	0,04897



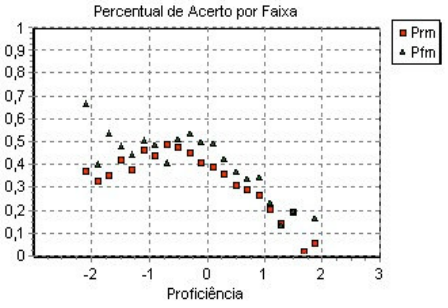


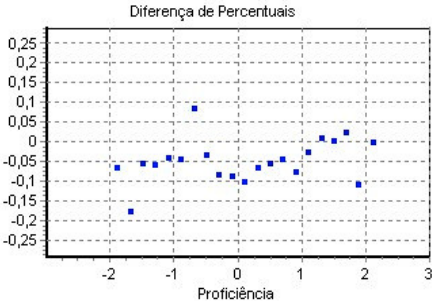
SisAnI - CAEd - UFJF

Item: h04194mg
Grupo de Referência: POLO 5
Grupo Focal: POLO 3 **Gabarito: A** **Opção analisada: B**

Método: Clássico

Alfa MH	AlfaD MH	SE AlfaD MH	M. Haenszel	P-Value	STDP-Diff	SE-STDP	STDD-Diff	SE-STDD
0,80032	0,52344	0,10447	24,95600	0,00000	0,04960	0,01027	0,49458	0,05001





SisAnI - CAEd - UFJF

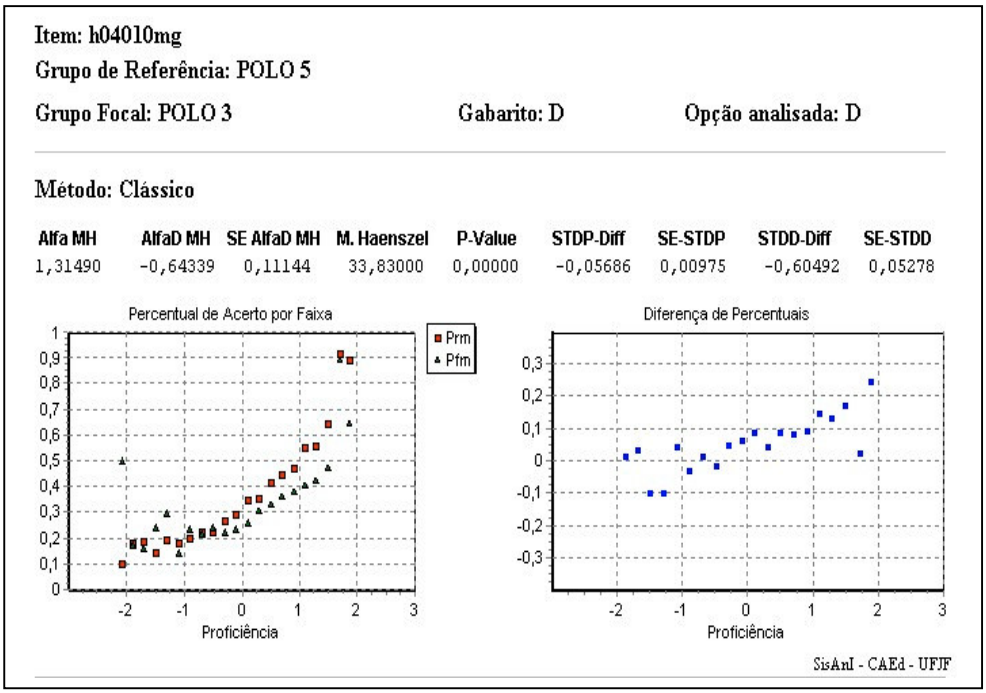
No entanto, as diferenças entre os comportamentos das respostas atribuídas aos itens são muito pequenas.

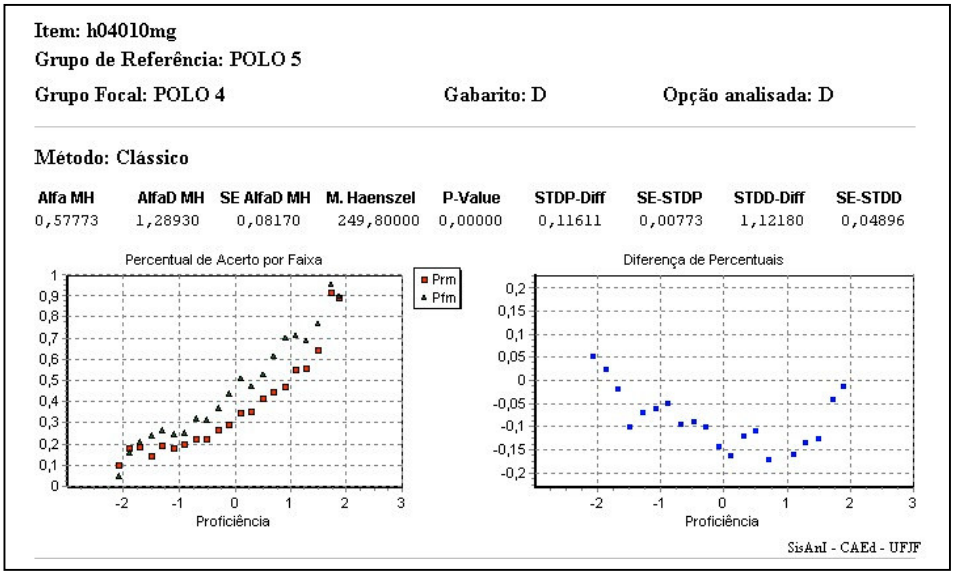
A última questão analisada trabalha o processo de transformação de um produto agrícola num produto industrial:

3)

"(H04010MG) Até chegar à nossa mesa, alguns produtos agrícolas passam por etapas de transformação. É o caso do angu, que é um alimento muito consumido em Minas Gerais. Que etapas são estas?"
A) Mandioca_fubá_angu.
B) Trigo_farinha_angu.
C) Fubá_milho_angu.
D) Milho_fubá_angu."

A competência avaliada é a relação do espaço urbano e do espaço rural. A alternativa correta corresponde à opção D. A questão parece estar bem construída, pois não há uma alternativa que influencie a resposta do aluno. Não se conseguiu ainda entender porque essa questão foi mais fácil para a Zona da Mata e mais difícil para o Triângulo do que para as demais regiões, como mostram os resultados abaixo:





As análises mostraram, ainda, que os alunos do Triângulo optaram mais freqüentemente pela opção A do que os de outras regiões.

Finalmente, na análise mais acurada dos itens H04079MG, H06038MG e H04236MG, utilizando-se as estatísticas e os métodos exemplificados acima, não se confirmou o comportamento diferenciado encontrado na análise baseada nos modelos da TRI. De fato, o comportamento diferenciado encontrado para esses itens é praticamente desprezível. Os resultados podem ser encontrados no Anexo 2 de Soares, Galvão e Genovez (2004).

6 CONCLUSÃO

A análise de comportamento diferencial mostrou-se bastante interessante pois apontou diferenças de competência, em geografia, dos alunos das diferentes regiões do Estado de Minas Gerais, especialmente com relação a itens que procuram avaliar as diferenças entre o espaço urbano e o espaço rural (que se mostraram desfavoráveis para os alunos da região metropolitana) e também as questões associadas ao meio ambiente (que se mostraram desfavoráveis aos alunos do interior, quando comparados aos da região metropolitana). Esse fato sugere que, para se alcançar eqüidade, o conteúdo desses itens precisa ser reforçado, adequadamente, nas regiões onde o item apresentou um comportamento

aquém do esperado. Análises dos itens da 8ª e da 3ª série também estão sendo realizadas e fazem parte da continuidade natural deste trabalho.

7 REFERÊNCIAS BIBLIOGRÁFICAS

BOCK, D. R.; ZIMOWSKI, M. F. Multiple Group IRT. In: LINDEN, W. J. V.; HAMBLETON, R. K (eds.). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag, 1995.

COLE, N. S. History and Development of DIF. In: HOLLAND, P. W.; WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ:Lawrence Erlbaum, 1993.

DORANS, N. J.; HOLLAND, P. W. DIF Detection and Description: Mantel-Haenszel and Standardization. In: HOLLAND, P. W.; WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 1993.

DORANS, N. J.; KULICK, E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, n. 23, p.355-368, 1986.

ELLIOT, L. G.; FONTANIVE, N. S.; ARRUDA, R. S.; KLEIN, R.; SOUZA, C. P.; SOARES, S. L. A. *SAEB 2001*: Relatório da Análise do Comportamento Diferencial dos Itens (DIF) entre Regiões. Rio de Janeiro: Fundação Carlos Chagas; Fundação Cesgranrio, 2002. (mimeo)

HOLLAND, P. W. On the study of differential item performance without IRT. *Proceding of the 27th Annual Conference of the Military Testing Association*. v. 1, p. 282-287. San Diego, 1985.

KLEIN, R.; ELLIOT, L. G.; FONTANIVE, N. S. *Saeb 99*: Relatório da Análise de comportamento diferencial dos itens entre regiões. Rio de Janeiro: Fundação Cesgranrio, 2000.

LONGFORD, N. T.; HOLLAND, P. W.; THAYER, D. T. Stability of the MH D-DIF Statistics Across Populations. In: HOLLAND, P. W.; WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 1993.

LORD, F. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980.

MANTEL, N.; HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, n. 22, p. 719-748, 1959.

MISLEV, R. J. Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, n. 11, p.3-31, 1986.

O'NEILL, K. A.; McPEEK, W. M. Item and test characteristics that are associated with differential item functioning. In: HOLLAND, P. W.; WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 1993.

PHILIPS, A.; HOLLAND, P. W. Estimation of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, n. 43, p. 425-431, 1987.

PROEB 2001 - Boletim Pedagógico. Ciências Humanas. *Competências e habilidades investigadas pelo SIMAVE para a 4ª e 8ª séries do Ensino Fundamental e 3ª série do Ensino Médio*. Secretaria do Estado da Educação. Minas Gerais, UFJF/CAED.

ROBINS, J.; BRESLOW, N.; GREENLAND, S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, n. 42, p. 311-323, 1986.

SIMAVE (2001). *Sistema Mineiro de Avaliação da Educação Pública*: uma construção coletiva. Secretaria de Estado da Educação, Minas Gerais, UFJF/LAME.

SOARES, T. M.; GALVÃO, A. F.; GENOVEZ, S. F. M. *Análise do Comportamento Diferencial dos Itens Utilizando o SisAni*. Juiz de Fora: CAEd/UFJF, 2004. (mimeo)

SOARES, T. M.; PEREIRA, D. R. M. Estudo de critérios de adequação para modelos da teoria da resposta ao item (TRI) aplicado ao caso do ensino fundamental da micro-região de Juiz de Fora em 1999. *Educação em Foco*, v. 6, n. 2, p. 91-108, 2002.

STRICKER, L. J.; EMMERICH, W. Possible Determinants of Differential Item Functioning: Familiarity, Interest and Emotional Reaction. *Journal of Educational Measurement*, v. 36, p. 347-366, 1999.

THISSEN, D.; STEINBERG, L.; WAINER, H. Detection of Differential Item Functioning Using the Parameters of Item Response Models. In: HOLLAND, P. W. WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 1993.

VALLE, R. C. Comportamento Diferencial do Item: uma apresentação. *Estudos em Avaliação Educacional*, n.25, p.3-21, jan./jun. 2002.

WAINER, H. Model-Based Standardized Measurement of an Item's Differential Impact. In: HOLLAND, P. W.; WAINER, H. (eds.). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 1993.

Recebido em: março 2005

Aprovado para publicação em: maio 2005

ANEXO

REVISÃO DOS MÉTODOS UTILIZADOS NA ANÁLISE DO COMPORTAMENTO DIFERENCIAL

1 Métodos Clássicos

Segundo Dorans e Holland (1993), a estatística de Mantel-Haenszel (M-H) foi proposta num contexto diferente por Mantel e Haenszel (1959), e adaptada por Holland (1985) e Holland e Thayer (1988) para uso na detecção do comportamento diferencial. Admitindo-se que cada grupo possa ser dividido em M subgrupos pareados de indivíduos com medidas de desempenhos similares (isto é, os indivíduos do m-ésimo subgrupo do grupo Focal apresentam desempenho similar aos indivíduos do m-ésimo subgrupo de referência), pode-se formar, então, M tabelas de contingência do tipo:

Tabela de Contingência 2x2 básica

Detecção do Comportamento diferencial

Grupo	Resultado do item		Total
	Certo	Errado	
Focal (F)	C_{Fm}	E_{Fm}	N_{Fm}
Referência (R)	C_{Rm}	E_{Rm}	N_{Rm}
Total	C_m	E_m	N_m

Onde, C_{Fm} é a freqüência observada de acertos para o grupo focal no m-ésimo subgrupo, C_{Rm} é a freqüência observada de acertos para o grupo de referência, E_{Fm} é a freqüência observada de erros para o grupo focal e, E_{Rm} é a freqüência observada de acertos para o grupo de referência. Em particular, se os valores observados nas tabelas de contingência forem obtidos a partir dos dados populacionais, então a razão de chances (*odds ratio*) teórica, ou populacional, entre o grupo referência e o grupo focal:

$$\alpha_m := \frac{[C_{Rm} / E_{Rm}]}{[C_{Fm} / E_{Fm}]}, \text{ para } m = 1, ..., M$$

Note-se que essa medida representa a discrepância observada em cada grupo de proficiências entre os desempenhos do grupo de referência e do grupo focal.

Admitindo-se que se houver comportamento diferencial este ocorra uniformemente nos diferentes grupos de proficiência, tal que $\alpha_m = \alpha,$ para $m = 1,...,M,$ uma medida global (uma estimativa no caso de amostras) de α é, então, dada por:

$$\alpha_{MH} = \frac{\left[\sum_m \frac{C_{Rm}E_{Fm}}{N_m}\right]}{\sum_m \frac{C_{Fm}E_{Rm}}{N_m}}$$

não havendo diferença para $\alpha_{MH} \cong 1$. Como é comum na teoria clássica se utilizar da estatística delta ($\Delta \coloneqq 13 - 4 \left\{\phi^{-1}(p)\right\}$ onde $\phi^{-1}(p)$ é o valor de distribuição normal para o qual a distribuição acumulada alcança *p*, o percentual de acerto do item) como uma medida da dificuldade do item, Holland e Thayer (1988) expressam α em termos das diferenças dos valores dos deltas e propõem a seguinte expressão alternativa para α:

$$\alpha_{MH}^{\Delta} = -\ 2.35\ \ \ell n(\alpha)$$

que, em certo sentido, padroniza o grau de comportamento diferenciado, segundo a dificuldade do item.

Se os termos correspondentes na tabela de contingência forem variáveis aleatórias dependentes da amostra, então $\widehat{\alpha}_{MH}^{\Delta}$, acima definido, é o estimador correspondente. Uma expressão para o erro padrão do estimador $\widehat{\alpha}_{MH}^{\Delta}$ foi desenvolvida por Robins, Breslow e Greenland (1986) e adaptada por Phillips e Holland (1987) (*cf.* Longford, Holland, Thayer, (1993), p.175):

$$VAR(\widehat{\alpha}_{MH}^{\Delta}) = \frac{1}{\sum_m \frac{C_{Rm}E_{Fm}}{N_m}} \sum_{m=1}^{M} N_m^{-2} (C_{Rm}E_{Fm} + C_{Fm}E_{Rm}\widehat{\alpha}_{MH}^{\Delta})[C_{Rm} + E_{Fm} + \widehat{\alpha}_{MH}^{\Delta}(C_{Fm} + E_{Rm})]$$

O teste de significância proposto consiste na comparação das seguintes hipóteses:

$$H_0: \quad \alpha_m = 1, \quad \text{para } m = 1,...,M.$$

$$H_1: \quad \alpha_m = \alpha \neq 1, \quad \text{para } \quad m = 1,...,M.$$

Sob a hipótese nula, acima, a estatística de Mantel e Haenszel,

$$M\text{-}H := \frac{\left[\left|\sum_m C_{Rm} - \sum_m \frac{N_{Rm} C_m}{N_m}\right| - 0.5\right]^2}{\sum_m Var(\hat{C}_{Rm})}$$

onde:

$$Var(\hat{C}_{Rm}) = \frac{\left[N_{Rm} C_m \; N_{Fm} \; E_m\right]}{\left[N_m^2 (N_m - 1)\right]}$$

se distribui aproximadamente como uma estatística χ^2, com um grau de liberdade.

Longford, Holland e Thayer (1993) apontavam que, até então, no *Educational Testing Service – ETS*, nos procedimentos para identificação de itens com comportamento diferencial no pré-teste ou na primeira administração, os itens eram classificados em três categorias, a partir das quais decisões específicas eram tomadas:

1) na primeira categoria o item classificado ou apresentava um valor não significativo (>0.05) para a estatística $\widehat{\alpha}_{MH}^{\Delta}$ ou o valor absoluto de α_{MH}^{Δ} (a estimativa correspondente) era menor que 1; nesse caso, a presença de DIF seria desconsiderada e o item poderia ser selecionado livremente;
2) na segunda categoria, o item classificado apresentava uma valor significativo para $\widehat{\alpha}_{MH}^{\Delta}$, mas $1 \leq \left|\alpha_{MH}^{\Delta}\right| \leq 1.5$. Nesse caso, se houver possibilidade, o item seria substituído por outro equivalente;

3) na terceira categoria, o item classificado apresentava um valor absoluto de α_{MH}^{Δ} maior que 1.5 e $\widehat{\alpha}_{MH}^{\Delta}$ é significativamente maior que 1.0. Nesse caso, o item só seria selecionado se fosse essencial às especificações.

Basicamente, entende-se que um item apresente DIF *uniforme*, quando este favorece uniformemente um grupo em relação a outro e apresente DIF *não-uniforme*, quando há uma interação entre o nível de habilidade e a performance no item, de modo que a direção do DIF muda ao longo da escala de habilidade. Naturalmente, a presença de DIF não-uniforme conduz a um comportamento diferenciado quanto à discriminação do item. Uma crítica ao método de Mantel-Haenszel (MH) é que ele não é sensível ao DIF não-uniforme. Este problema motivou a busca por técnicas de detecção do DIF que superassem essa limitação, como é o caso da regressão logística. No entanto, o método MH ainda é a metodologia mais utilizada para análise do DIF, inclusive pelo ETS, nos exames do *National Assessment of Educational Progress* (NAEP), e aqui no Brasil, na análise do Saeb.

Os chamados procedimentos baseados em padronização (Dorans, Kulick, 1986) são métodos flexíveis, que sob certas condições produzem resultados equivalentes ao método de Mantel-Haenszel na detecção do comportamento diferencial, porém eles admitem outras possibilidades métricas para medir a quantidade de DIF, além de fornecerem uma metodologia que pode ser aplicada na análise das causas prováveis do comportamento diferenciado em função dos percentuais de respostas atribuídos aos "distratores", no caso de teste de múltipla escolha. Basicamente, a análise por procedimentos de padronização parte da análise gráfica comparativa das respostas atribuídas pelo grupo focal e pelo grupo de referência, divididos segundo subgrupos pareados de alunos de acordo com alguma medida de proficiência. Uma medida global da diferença de desempenho, por parte de ambos os grupos, é a seguinte:

$$STD\,P-DIFF = \frac{\sum_{m=1}^{M} w_m (p_{Fm} - p_{Rm})}{\sum_{m=1}^{M} w_m},$$

onde p_{Fm} e, p_{Rm} são, respectivamente, o percentual de acerto do item em ambos os grupos, e w_m é o peso correspondente ao grupo (normalmente, o peso mais usado é o número de indivíduos no grupo focal, $w_m = N_{Fm}$).

Observando que $STD\,P-DIFF \in [-1, 1]$, no ETS o critério empregado para admitir inexistência de DIFF é um valor

$$STD\,P-DIFF \in [-0.05, 0.05],$$

valores entre

$$STD\,P-DIFF \in [-0.1, -0.05) \cup (0.05, 0.1]$$

indicariam uma presença moderada de DIFF e fora desses limites uma presença mais severa. Novamente, se considerarmos que os dados são amostrais, o estimador natural dessa estatística apresenta a seguinte expressão para o seu desvio-padrão (ibidem, p. 50):

$$SE(STD\,P-DIFF) = [\frac{P_F\,(1-P_F)}{N_F} + VAR(P_F^*)]^{0.5}, \text{ onde}$$

$$VAR(P_F^*) = \sum_{m=1}^{M} \frac{N_{Fm}^2\; p_{Rm}\,(1-p_{Rm})}{N_{Rm}\,N_F^2},$$

$$P_F := \frac{\sum_{m=1}^{M} N_{Fm}\;\; p_{Fm}}{\sum_{m=1}^{M} N_{Fm}}, \quad P_F^* := \frac{\sum_{m=1}^{M} N_{Fm}\;\; p_{Rm}}{\sum_{m=1}^{M} N_{Fm}}.$$

Uma proposta de Dorans e Holland (1993) é utilizar a medida $STD\,P-DIFF$, assim como a análise gráfica como a apresentada no exemplo acima, na análise das respostas comparativas de cada distrator de tal forma que as possíveis causas do comportamento diferencial observado pudessem ser investigadas.

2 Métodos Baseados em Modelos da TRI

Uma vez que o item tenha seu desempenho representado por um modelo estatístico é natural supor que diferença significativa do mesmo modelo para grupos diferentes pode constituir-se em medidas de um provável comportamento diferenciado do item.

Os primeiros métodos desse tipo parecem ser os de Lord (1980), que propõe um teste normal para verificar se há diferença entre os parâmetros de dificuldade do item, usando a estatística:

$$d := \frac{\hat{b}_F - \hat{b}_R}{\sqrt{\mathrm{var}(\hat{b}_F) + \mathrm{var}(\hat{b}_R)}}$$

ou para testar, simultaneamente, as diferenças entre os parâmetros de dificuldade e discriminação: $D^2 := v^T \Sigma^{-1} v$, onde $v := [\hat{b}_F - \hat{b}_R, \hat{a}_F - \hat{a}_R]$ e, Σ, é a matriz de covariância amostral das diferenças entre os estimadores dos dois parâmetros (assintoticamente D^2 é distribuído segundo uma distribuição $\chi^2(2)$). Lord sugere que o parâmetro de acerto casual seja fixado para ser o mesmo nos dois grupos. Outro ponto importante é que, naturalmente, ambas as estimativas dos parâmetros devem estar devidamente equalizadas o que é natural num processo de calibração simultânea.

Thissen, Steinberg e Wainer (1993) propõem um procedimento geral para detecção de um provável comportamento diferenciado que denominaram de método geral TRI-Razão de Verossimilhança ("general IRT-LR"). De fato, a idéia geral do procedimento é baseada no emprego de modelos para grupos múltiplos (Mislev, 1987; Bock, Zimowski, 1995). O teste de razão de verossimilhanças empregado é um conhecido procedimento para decisão sobre dois modelos aninhados. *Grosso modo*, o método propõe que se divida o conjunto de itens do teste em dois grupos, um grupo denominado de itens âncoras, para os quais o comportamento diferenciado não é significativo (pode não conter itens, se todos os itens do teste forem testados), e um segundo grupo, composto pelos itens para os quais se deseja testar o comportamento diferenciado (com pelo menos 1 item). Os autores propõem utilizar o método de máxima verossimilhança marginal que consiste na maximização do logaritmo da função de verossimilhança marginalizada:

$$P_m(\mathrm{P}\,;X_g) = \int_{\Theta} \prod_{i=1}^{N} P_i(X_{g,i};\mathrm{P})\; g(\theta;n_g)\; d\theta$$

onde n_g representa os parâmetros da distribuição das proficiências para o grupo g. Quando se utiliza o modelo de 3 parâmetros para os itens, esses autores salientam que, na análise do comportamento diferenciado para o parâmetro de dificuldade, os parâmetros de discriminação e acerto casual devem ser constantes para os grupos. E, na análise do comportamento diferenciado da discriminação, o parâmetro de acerto casual deve ser constante. Assim, uma análise do comportamento diferencial, segundo esses três parâmetros, deveria ser conduzida de tal forma que primeiro fosse analisado o comportamento diferencial do parâmetro **c**, depois o parâmetro **a,** e, finalmente, o parâmetro **b**. Definido o tipo de comportamento diferencial que se deseja testar e quais itens entram no teste, dois modelos são comparados por meio de um teste de razão de verossimilhanças: o modelo que não considera comportamento diferenciado por grupo e o modelo que considera esse comportamento. Note-se que os modelos globais para todos os itens, em ambas as situações, estão aninhados hierarquicamente e, sob condições apropriadas, as razões das verossimilhanças apresentam, assintoticamente, distribuição χ^2 com graus de liberdade igual à diferença entre o número de parâmetros dos dois modelos.

O *software* BILOG-MG apresenta uma implementação desse método para verificar a presença do comportamento diferencial quanto ao parâmetro de dificuldade. Não há maiores detalhes no manual, mas parece que foi implementada a possibilidade de ajustar (para n itens simultaneamente) um modelo para grupos múltiplos que impõe mesmo parâmetro de acerto casual e mesmo parâmetro de discriminação para os itens nos diferentes grupos, admitindo diferentes valores do parâmetro dificuldade para os itens segundo os quais se deseja testar o provável comportamento diferencial.

Recebido em: março 2005

Aprovado para publicação em: maio 2005