ESTIMATIVAS NÃO PARAMÉTRICAS DAS CURVAS CARACTERÍSTICAS DOS ITENS EM TESTES EDUCACIONAIS

MARCOS ANTONIO DA CUNHA SANTOS
JOSÉ FRANCISCO SOARES

RESUMO

A proficiência dos alunos submetidos a um teste para a medida de uma competência pode ser estimada através das técnicas estatísticas associadas à Teoria de Resposta ao Item (TRI). Em situações reais, o cálculo dos parâmetros clássicos, da tabela de distribuição dos itens, curvas características das diferentes escolhas e a identificação de comportamento diferencial dos itens são importantes para identificar os itens que podem influir na precisão dos resultados da TRI. O objetivo deste artigo é apresentar um conjunto de procedimentos implementados como uma macro do software livre R, *que produzem informações úteis na análise prévia do comportamento dos itens de uma forma simples e interativa. Além disso, apresentamos um método não paramétrico para verificação de comportamento diferencial de itens através da visualização das curvas de respostas. Esse método não é disponível em outros softwares e constitui em poderosa ferramenta descritiva dos itens.*

PALAVRAS-CHAVE TEORIA DE RESPOSTA AO ITEM • AVALIAÇÃO DA APRENDIZAGEM.

RESUMEN

El rendimiento de alumnos sometidos a un prueba para medir una competencia es usualmente estimada a través del modelo de la Teoría de Respuesta al Ítem, cuyo ajuste a los ítems define la calidad de la estimativa. Frente a esto, en situaciones reales de prueba, se realiza un análisis previo de los ítems con el objetivo de seleccionar aquellos que, por presentar un comportamiento empírico anómalo, deben ser excluidos del análisis. El objetivo de este artículo es presentar un conjunto de comandos y procedimientos implementados en un macro del software libre R, *que produce esas informaciones de forma simple e interactiva. Además, ese macro implementa una estimativa no paramétrica de las curvas características de cada opción de respuesta, información no disponible en otros softwares y que se constituye en una poderosa herramienta descriptiva de los ítems.*

PALABRAS CLAVE TEORÍA DE LA RESPUESTA AL ÍTEM • EVALUACIÓN DEL APRENDIZAJE.

ABSTRACT

The proficiency of students taking a test to measure a certain competence or ability is usually estimated using the Item Response Theory model; the adjustment to the items defines the quality of the estimate. Thus, in a real testing situation, a preliminary analysis of the items is carried out in order to select the items which, because they present an empirical anomalous behavior, should be excluded from the analysis. The aim of this paper is to present a set of procedures implemented as a macro of the free R software, which produce this preliminary analysis in a simple and interactive way. Furthermore, this macro implements a nonparametric estimate of the curves which are characteristic for each answer choice, the kind of information that is not available in other software and which is a powerful descriptive tool of the items.

KEYWORDS ITEM RESPONSE THEORY • LEARNING ASSESSMENT.

INTRODUÇÃO

A proficiência dos alunos submetidos a um teste para a medida de uma competência é usualmente estimada através de modelo da Teoria de Resposta ao Item (TRI), cujo ajuste aos itens define a qualidade da estimativa. Diante disso, em situações reais, é importante a realização de uma análise prévia dos itens. No âmbito do Saeb – Prova Brasil, os seguintes procedimentos têm sido utilizados: cálculo dos parâmetros clássicos de caracterização do item, a tabela de distribuição das respostas dos alunos entre as opções do item, as curvas características das diferentes opções, os respectivos autovalores da matriz de correlação entre os itens e uma medida do comportamento diferencial do item. Essas informações permitem identificar com segurança os itens do teste cujo comportamento empírico sugere sua exclusão da análise TRI.

O objetivo deste artigo é apresentar um conjunto de comandos e procedimentos, implementados como uma macro do *software* livre R, que produz essas informações de forma simples e interativa. Além disso, essa macro produz uma estimativa não paramétrica das curvas características de cada opção de

resposta, informação não disponível em outros *softwares* e que se constitui em poderosa ferramenta descritiva dos itens.

Este artigo está organizado da seguinte maneira: considerando que os indicadores do comportamento empírico aqui considerados são amplamente conhecidos, o artigo faz apenas uma apresentação sumária de suas propriedades, remetendo o leitor para a literatura na área. Em seguida apresenta o cálculo desses indicadores no ambiente do *software* R. A macro correspondente está disponibilizada a todos os interessados.

SÍNTESES DO COMPORTAMENTO EMPÍRICO DE ITENS

O comportamento empírico de um item é primeiramente caracterizado por dois indicadores, cujas características foram estudadas na teoria clássica de testes: os indicadores de dificuldade e de discriminação.

DIFICULDADE

A dificuldade do item é definida como a porcentagem de alunos que escolheram a opção correta. Um valor próximo de um no índice de dificuldade indica que o item é fácil, e próximo de zero indica que se trata de um item difícil. Em testes que pretendem diagnosticar o aprendizado dos alunos e não hierarquizar seus desempenhos, a dificuldade dos itens deve variar. Pasquali (2003) recomenda que a dificuldade dos itens seja dividida em cinco faixas: (0-20, 20-40, 40-60, 60-80 e 80-100). Recomenda, ainda, que 10% dos itens sejam distribuídos em cada uma das duas faixas extremas, 20%, em cada uma das faixas seguintes, e 40% na faixa média. Na prática é usual, além do cálculo do indicador de dificuldade, registrar também em uma tabela a porcentagem de alunos que se submeteram ao teste e que escolheram cada uma das opções de resposta.

DISCRIMINAÇÃO

O escore total dos alunos é definido como o número de itens que um dado aluno acerta. O comportamento empírico de um bom item é tal que a resposta dos alunos a esse item está associada positivamente ao escore total dos alunos no teste. Essa associação é usualmente medida através de um coeficiente de

correlação. Nesse caso o coeficiente de correlação adequado recebe o nome de bisserial. A correlação bisserial é uma estimativa da correlação de Pearson entre uma variável binária, como é o caso a resposta ao item (correta/incorreta) e o escore total.

Antes da popularização do coeficiente de correlação bisserial, usava-se uma medida de discriminação obtida pela diferença entre a dificuldade do item no grupo de alunos que estão classificados entre os 27% superiores e no grupo de alunos que constituem os 27% inferiores. Essa medida, usualmente denotada por D, produz resultados similares aos obtidos pelo coeficiente de correlação bisserial e, portanto, não é necessária.

Um bom item é aquele que discrimina os alunos de desempenho superior daqueles com desempenho inferior e apresenta, portanto, um valor positivo e alto do coeficiente de correlação bisserial, assim como do índice D. Itens com esses indicadores nulos ou negativos apontam para um item com comportamento anômalo e que não deve ser incluído em outras análises.

INDICADOR DE UNIDIMENSIONALIDADE

Para a estimação da proficiência de alunos, através dos modelos da Teoria de Resposta ao Item, é necessário que o teste considerado seja unidimensional (ANDRADE, TAVARES, VALLE, 2000, p. 16). Em outras palavras, é necessário que os itens que compõem o teste sejam indicadores de um único construto.

Há muitos métodos para caracterizar a unidimensionalidade de um teste. A análise preliminar dos itens trabalha com uma definição restrita do conceito de unidimensionalidade. Isso se justifica, pois, do ponto de vista pedagógico, é razoável supor que o bom desempenho em qualquer teste exige usualmente mais de uma competência. Por exemplo, a competência leitora é necessária para o entendimento das questões que, em princípio, medem o domínio de outras competências. Assim sendo, o que se busca, na prática, são evidências de que os itens utilizados no teste estão associados a um fator dominante, e não que exista um único fator. Em outras palavras, o problema prático importante é medir o grau de unidimensionalidade de um teste e não somente se um teste é ou não unidimensional.

REGRESSÃO ITEM-TESTE

Lord e Novick (1968) recomendam a análise da associação entre a porcentagem de acertos de cada item, dentre os alunos que acertaram uma dada porcentagem de itens no teste. Essa recomendação é aqui estendida para a consideração da porcentagem de escolha de cada uma das categorias de resposta. Assim, para cada item devem ser produzidas tantas curvas quantas forem as opções de respostas. A análise dessas curvas é uma ferramenta poderosa para a verificação do comportamento empírico do item. Assim sendo, sua disponibilização é essencial para uma análise inicial efetiva. Neste artigo essa técnica é ilustrada com a suavização dessas curvas, utilizando-se para isso técnicas não paramétricas.

COMPORTAMENTO DIFERENCIAL DE ITENS

Do ponto de vista psicométrico, "um item demonstra DIF (*Differential Item Functioning*) quando pessoas do mesmo nível de proficiência, mas pertencentes a diferentes grupos, não têm a mesma probabilidade de acertar o item" (HAMBLETON, SWAMINATHAN, ROGERS, 1991, p. 109-110).

A análise DIF tem sido aplicada para investigar a existência de viés em provas devido à presença de grupos de indivíduos com diferenças em relação a características raciais e étnicas, gênero, região geográfica, circunstâncias socioeconômicas e outras características que possam produzir um comportamento diferencial. A aplicação dessa análise se estende a qualquer preocupação em relação à validade do uso dos resultados de provas na presença de grupos com provável comportamento diferencial.

As análises DIF não utilizam necessariamente os parâmetros de item calibrados com TRI, como a técnica de Mantel-Haenzel, desenvolvida por Holland e Thayer (1988), e o método da regressão logística, desenvolvido por Swaminathan e Rogers (1990).

CÁLCULO DOS INDICADORES

A forma de cálculo desses indicadores reflete as opções consideradas na organização do teste. Há, essencialmente, duas maneiras. A primeira, que será denominada "formato vestibular", contempla a situação em que todos os alunos são submetidos aos mesmos itens, eventualmente com a organização de diferentes cadernos de prova, que diferem entre si apenas

pela ordem dos itens. A segunda, que será denominada "formato Saeb", contempla a situação em que os alunos resolvem itens diferentes, ainda que haja itens comuns. Nesse caso usa-se algum esquema BIB. Nessa situação o escore total dos alunos não são comparáveis entre si, já que as dificuldades dos itens são diferentes. Diante disso algumas adaptações devem ser feitas antes do cálculo de alguns dos indicadores.

MATRIZ DE DADOS

TABELA 1 - Exemplo de matriz de dados no formato vestibular

CNDCOD	ESCOLA	ITEM 1	ITEM 2	...	ITEM 15	SEXO	RAÇA	NSE	GNSE
35	...	0	0	...	1	2	4	1,185	5
43	...	0	0	...	0	2	1	-0,489	2
60	...	0	0	...	1	2	1	-0,699	2
86	...	1	0	...	0	2	1	-0,491	2
124	...	1	NA	...	0	2	1	-0,483	2
140	...	0	1	...	0	1	3	-0,589	2
159	...	1	1	...	NA	2	4	-0,961	1

A tabela 1 mostra um exemplo de matriz de dados no formato vestibular. Nas linhas encontram-se os dados de identificação, as respostas aos itens e o valor das variáveis de natureza socioeconômicas de cada candidato. Em destaque, ao centro da tabela, a matriz de dados somente dos itens, com a codificação 1 para acerto e 0 para resposta errada. Variáveis faltantes são possíveis no conjunto de dados e no modelo TRI. No ambiente de programação do *software R* os dados faltantes são registrados como "NA". A matriz da tabela 1 refere-se a um exemplo de itens corrigidos; no entanto, matrizes com itens não corrigidos (respostas "A", "B", "C" etc.) também podem ser utilizadas em uma análise prévia.

UMA MACRO DO SOFTWARE R PARA ANÁLISE PRÉVIA

Os cálculos deste estudo foram efetuados com o *software R*, gratuito e muito difundido na comunidade de pesquisa acadêmica em estatística e áreas afins. Esse *software* foi desenvolvido para utilização em análises estatísticas. É possível também a sua utilização através de programas desenvolvidos por usuários, para fins específicos. Esses programas adicionais ("macros"), uma vez elaborados, podem ser divulgados aos interessados

para uso no ambiente R através de simples arquivos texto (*scripts*) ou como um adendo ao *software* (*package*).

A ideia básica deste trabalho foi implementar no ambiente R ferramentas úteis para a fase de análise prévia dos dados de testes educacionais. Essa fase de análise tem se mostrado de grande importância para preceder análises mais sofisticadas, como é o caso do ajuste da matriz de resposta aos modelos TRI de um, dois ou três parâmetros.

A necessidade da fase de análise prévia pode ser justificada principalmente pelos seguintes motivos: na imensa maioria das aplicações práticas, a utilização de modelos de análise TRI é muito dependente da qualidade do "input". Problemas na matriz inicial de dados como, por exemplo, a existência de itens individuais que apresentam comportamento atípico em relação ao modelo proposto pode ocasionar problemas de convergência do algoritmo EM. Esse algoritmo é utilizado para ajustes de dados ao modelo TRI na maioria dos programas e *softwares* existentes. Problemas de convergência desse algoritmo são conhecidos e uma análise prévia rigorosa dos dados de entrada torna-se necessária para detecção de possíveis fontes de problemas. Além disso, a avaliação de presença de DIF, por exemplo, tem sido normalmente efetuada através de cálculos de indicadores que podem mascarar, sob determinadas condições, a real extensão do problema. Nos casos citados, uma análise prévia é um importante meio de auxiliar a obtenção de indicadores de qualidade, sob o ponto de vista da detecção de problemas reais.

Como exemplo de uma análise prévia, a figura 1 mostra, para uma matriz de dados de itens não corrigidos, o percentual de acertos para um item da prova. Nesse exemplo foram utilizados os dados de um teste do Saeb. A probabilidade de acerto foi calculada em uma janela móvel, isto é, dado um intervalo com pontuação inicial e final, é calculado o percentual de cada resposta dada nessa faixa de pontuação. Por exemplo, para a escolha da resposta A, é calculado o [número de escolhas "A"]/[número de candidatos] em cada faixa de pontuação. Esse tipo de gráfico fornece uma visualização imediata da evolução das respostas em função da faixa de pontuação dos candidatos. Na figura 2 encontra-se o mesmo tipo de informação, porém com a utilização de curvas suavizadas (ajuste por kernel).

FIGURA 1 – Proporção de escolha de cada opção em item público do Saeb

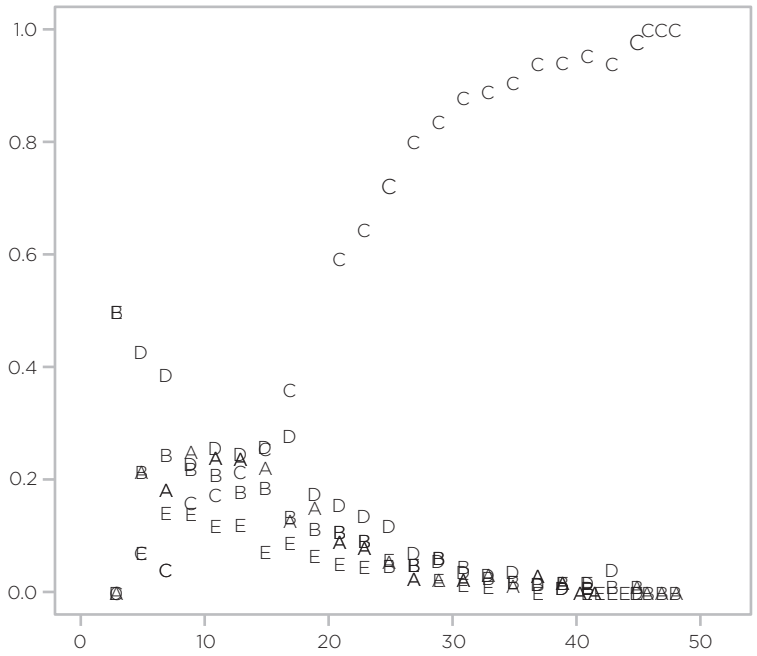
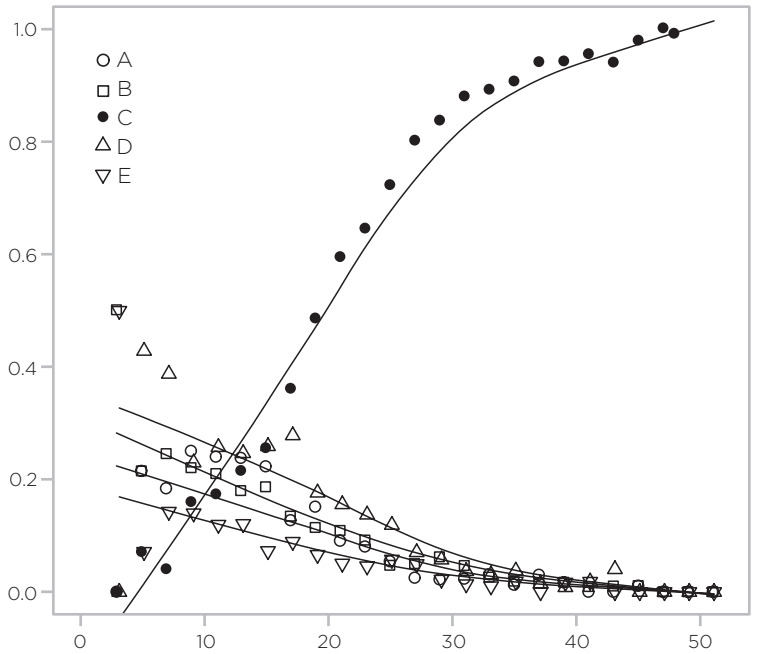


FIGURA 2 – Uma visualização alternativa para a proporção de escolhas de item público do SAEB com utilização de curvas suavizadas



Com o programa desenvolvido para o R neste trabalho, gráficos como os das figuras 1 e 2 podem ser gerados através de um único comando na janela de console:

```
> curvaResposta(item=22,B,passo=2,tipo=3)
```

Nesse exemplo, "B" é a matriz de dados; "passo" e "tipo" são parâmetros para o tipo de gráfico escolhido.

UM TESTE NÃO PARAMÉTRICO PARA DETECÇÃO DE DIF

Ainda no contexto de análise prévia, implementamos na macro para o R um teste não paramétrico para detecção de DIF, conhecido como *teste de permutação*. Testes de permutação são testes estatísticos efetuados sem a adoção de um modelo estatístico prévio para o comportamento dos dados. Esses testes são utilizados em várias áreas do conhecimento e exigem uma programação voltada para maior velocidade e desempenho durante os cálculos computacionais, razão pela qual não foram muito utilizados no passado, devido às limitações computacionais da época. Uma descrição dos aspectos teóricos dos testes de permutação pode ser encontrada em Phillip (1993) e uma apresentação de testes similares, com exemplos de aplicação em várias áreas do conhecimento, pode ser encontrada em Manly (1997).

Os testes não paramétricos são indicados quando não há informação suficiente que justifique a adoção de um modelo estatístico específico para os dados. Outra situação em que são indicados é em análises prévias, quando se deseja justamente verificar o comportamento dos dados disponíveis antes da adoção de um modelo estatístico mais elaborado. Por exemplo, antes de ser efetuada uma análise TRI para a adoção de uma escala, é importante verificar através de procedimentos estatísticos como testes não paramétricos, se há evidências de DIF ou de itens com comportamento atípico antes de uma segunda análise.

O teste para verificação de presença de DIF adotado neste trabalho é intuitivo e consiste basicamente em, dados os grupos de indivíduos A e B dos quais se deseja verificar DIF,

obter um "envelope de confiança" para comparação das curvas de probabilidade de acerto dos dois grupos. Esse procedimento é realizado da seguinte maneira: as curvas de probabilidade de acerto são calculadas para os grupos A e B, com o mesmo número de indivíduos. Um novo grupo A', do mesmo tamanho do grupo A, é formado escolhendo-se aleatoriamente os indivíduos entre os dois grupos. Uma nova curva de probabilidade de acerto para o grupo A' (formado por candidatos permutados) é calculada, procedimento que é repetido várias vezes (tipicamente centenas de vezes). Uma vez armazenadas todas as curvas de probabilidade de acerto, curvas de referência podem ser geradas adotando-se os percentis 0.975 e 0.025 para construção de intervalos empíricos de 95% de confiança em cada faixa de pontuação. As curvas originais (grupos A e B ou referencia e focal) devem então ser comparadas com o envelope de confiança obtido através das permutações.

O pressuposto básico do teste descrito no parágrafo anterior é que, *sob a hipótese nula* H_0 de similaridade entre os grupos, não há diferença significativa entre os grupos permutados. Portanto, sob H0, as curvas de probabilidade de acerto, obtidas nos grupos com indivíduos permutados, quando comparadas com as curvas originais dos grupos A e B não devem apresentar diferenças significativas. Se a curva obtida originalmente para o grupo B estiver fora do envelope de confiança, há evidência de comportamento diferencial do item no nível de confiança do teste.

A figura 3 mostra as curvas de proporção de acerto por faixa de pontuação para um item de matemática (item 32) no vestibular UFMG-2004[1], para dois grupos de 5.000 candidatos cada (grupos de referência e controle). Os grupos foram separados de acordo com diferentes graus de um indicador do nível socioeconômico do candidato. O primeiro grupo foi formado com todos os candidatos com maior indicador socioeconômico (grupo foco) e o segundo, formado por candidatos com menor indicador socioeconômico. Em tracejado na figura está o envelope de confiança de 95% obtido por teste de permutação. O envelope de confiança foi calculado a partir de 100 curvas obtidas para grupos de 5.000 candidatos cada, formados através de permutações entre os indivíduos dos dois grupos originais. A área dos círculos é proporcional ao

1 Os dados para esta análise foram cedidos ao GAME em 2005, sem a identificação dos alunos envolvidos.

número de indivíduos nos grupos foco e referência em cada faixa de pontuação. Pode-se observar que as curvas obtidas para os dois grupos originais estão dentro do envelope de confiança para quase todas as faixas de pontuação. Esse fato é um indicador de ausência de DIF significativo. Nesse exemplo, a pontuação total refere-se à soma dos pontos obtidos nas provas de língua estrangeira, química, matemática, física, história e biologia (90 itens).

FIGURA 3 – Curvas de proporção de acerto por faixa de pontuação, para os grupos de referência e controle, para um item de matemática no vestibular UFMG-2004. Em tracejado o envelope de confiança de 95%, obtido com 100 grupos de 5.000 indivíduos escolhidos aleatoriamente entre os grupos focal e referência

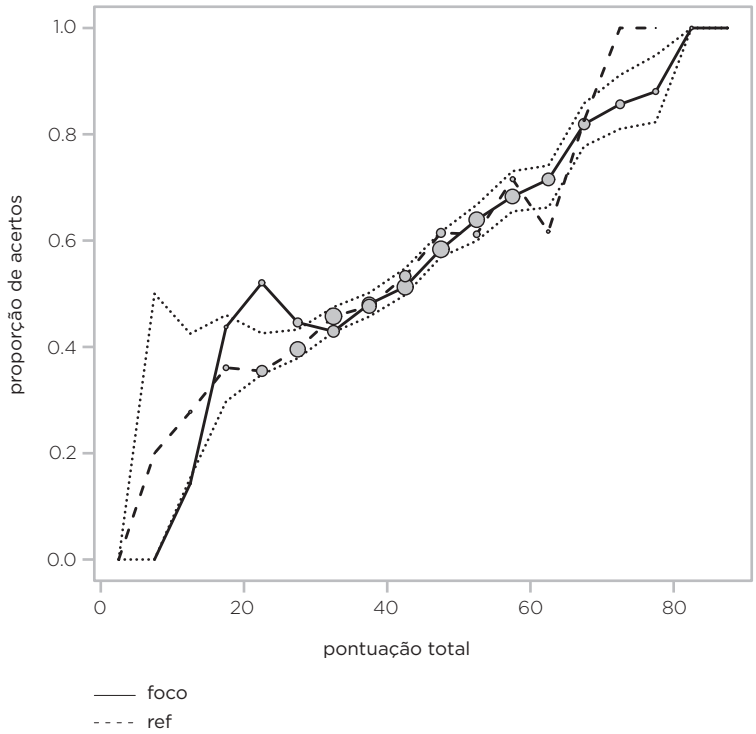
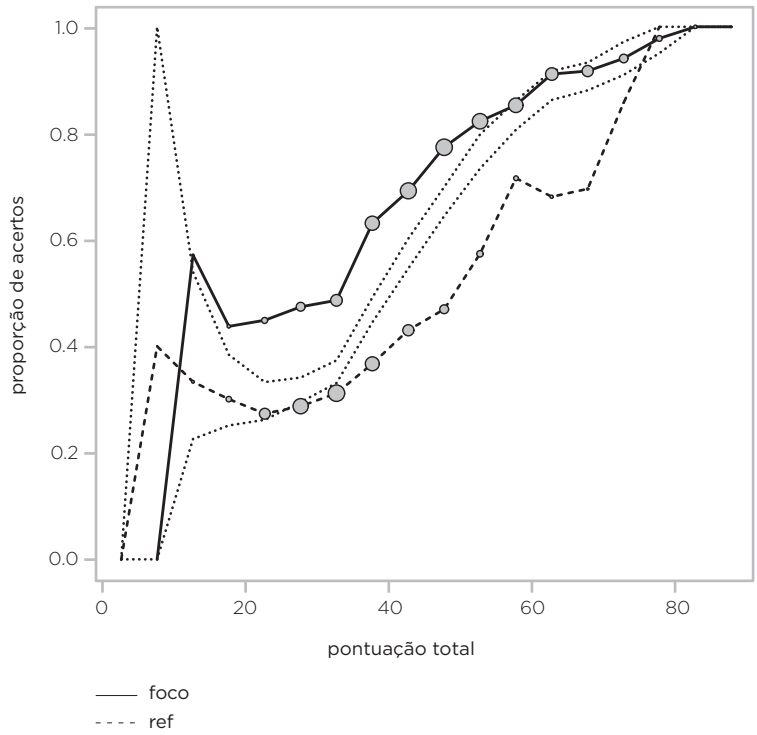


FIGURA 4 – Curvas de proporção de acerto por faixa de pontuação, para os grupos de referência e controle para um item de língua estrangeira no vestibular UFMG-2004. Em tracejado o envelope de confiança de 95%, obtido com 100 grupos de elementos permutados



A figura 4 mostra as curvas de proporção de acerto por faixa de pontuação, para os grupos de referência e controle, de um item de língua estrangeira no vestibular UFMG-2004 (item 08). Nesse caso, como no exemplo anterior, dois grupos de 5000 indivíduos cada foram analisados, com o primeiro grupo formado apenas por indivíduos com maior índice de nível socioeconômico, e o outro, com candidatos de menor nível socioeconômico. Em tracejado o envelope de confiança de 95%, obtido, como no exemplo anterior, através de 100 grupos permutados. O número do item é apenas para referência no banco de dados. É possível observar que os candidatos que obtiveram baixa pontuação na soma desses testes não apresentaram desempenho significativamente diferente nos dois grupos. Para candidatos em faixas de pontuação intermediária, no entanto, há uma diferença entre os dois grupos que é relevante do ponto de vista do teste estatístico, de acordo com os valores

obtidos para o envelope de confiança. A curva parcialmente fora do envelope indica que a diferença observada entre as duas curvas de acertos tem pouca probabilidade de ser atribuída ao acaso. Dessa forma, há evidência de DIF para os dois grupos em determinadas faixas de pontuação.

Esses procedimentos podem ser efetuados através de comandos simples através da macro R desenvolvida neste trabalho, e espera-se que essa macro venha a contribuir de forma eficaz para o monitoramento e detecção de DIF, durante pré-análises de dados de testes educacionais.

DISCUSSÃO

Neste artigo foi apresentado um procedimento para a análise descritiva dos itens efetuados com a utilização de programas desenvolvidos para o software R. Essa análise deve preceder ao ajuste de modelos TRI. Esses modelos têm grandes e fortes hipóteses e, assim sendo, seu uso sem análise prévia pode levar a situações como a não convergência ou convergência do algoritmo para ajuste do modelo TRI com valores pouco razoáveis. Dessa forma, a boa prática recomenda uma análise prévia cuidadosa do comportamento de cada item.

Importante destacar que não se tratou aqui de verificação da qualidade do ajuste, tema correlato, mas que deve ser considerado depois do ajuste de modelos TRI. Ou seja, embora com interseção, o trabalho aqui apresentado não pretende substituir as curvas descritivas dos resultados do ajuste, tais como as apresentadas nas sínteses de Klein (2003).

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item*: Conceitos e Aplicações. Sinape - Simpósio Nacional de Probabilidade e Estatística, 2000.

GOOD, Phillip. *Permutation Tests*. Springer Series in Statistics – Springer-Verlag, New York, Inc., 1993.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of Item Response Theory*. CA-Sage Publications, 1991.

HOLLAND, P. W.; THAYER, D. T. Differential item performance and the

Mantel-Haenszel procedure. In: WAINER, H.; BRAUM, H. I. (Org.). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum, 1988.

KLEIN, R. Utilização da Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (Saeb). *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 11, n. 40, p. 283-96, jul./set. 2003.

LORD, F. M.; NOVICK, M. R. *Statistical theories of mental test scores*. Reading MA: Addison-Welsley, 1968.

MANLY, Bryan F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. London: Chapman & Hall, 1997.

PASQUALI, L. *Psicometria*: teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2003.

R. FOUNDATION FOR STATISTICAL COM PUTING. *R development core team*: R: A language and environment for statistical computing. Vienna, 2011. Disponível em: http://www.R-project.org/.

SWAMINATHAN, H.; ROGERS, H. J. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, n. 27, p. 361-370, 1990.

MARCOS ANTONIO DA CUNHA SANTOS
Professor Doutor do Departamento de Estatística da Universidade Federal de Minas Gerais – UFMG
msantos@est.ufmg.br

JOSÉ FRANCISCO SOARES
Professor Doutor do Programa de Pós-graduação da Faculdade de Educação da Universidade Federal de Minas Gerais – UFMG
francisco-soares@ufmg.br

Recebido em: DEZEMBRO 2011
Aprovado para publicação em: MAIO 2012