ANÁLISE CLÁSSICA DE TESTES COM DIFERENTES GRAUS DE DIFICULDADE

ADRIANO FERRETI BORGATTO
DALTON FRANCISCO DE ANDRADE

RESUMO

AO objetivo deste trabalho é mostrar a aplicação da metodologia descrita por meio de macro do SAS®, usando as provas do Saeb 2005 aplicadas a estudantes da 3ª série do ensino médio. Além da análise dos dados do Saeb, foi utilizado um conjunto de dados simulado a partir de duas provas com graus de dificuldade diferentes. Observou-se, principalmente no estudo de simulação, que os resultados obtidos pelo escore podem gerar conclusões erradas a respeito do item.

PALAVRAS-CHAVE TEORIA DE RESPOSTA AO ITEM • *NORMIT* • ANÁLISE DE DADOS.

RESUMEN

El objetivo de este trabajo es mostrar la aplicación de la metodología descrita a través de un macro del SAS®, usando las pruebas del Saeb 2005 aplicadas a los estudiantes de tercer año de la enseñanza media. Además del análisis de esos datos, se utilizó un conjunto de datos simulados a través de dos pruebas con grados de dificultad diferentes. Se observó, principalmente en el estudio simulado, que los resultados obtenidos a través de la puntuación puede generar conclusiones erradas al respecto del ítem.

PALABRAS CLAVE TEORÍA DE LA RESPUESTA AL ÍTEM • *NORMIT* • ANÁLISIS DE DATOS.

ABSTRACT

The aim of this paper is to compare the performance of items using the measures normit and score in the Saeb 2005 tests administered to twelfth grade High School students. Besides the Saeb data, we also used a simulated dataset related to two tests with different degrees of difficulty. It was shown, mainly in the simulation study, that the results obtained with the score could lead to wrong conclusions about the quality of the item.

KEYWORDS ITEM RESPONSE THEORY • NORMIT • DATA ANALYSIS.

TEORIA CLÁSSICA DOS TESTES

A avaliação do desempenho dos estudantes depende, fundamentalmente, da qualidade dos itens das provas. Uma análise preliminar que complementa aquela realizada via a Teoria da Resposta ao Item (TRI) é a análise pela Teoria Clássica dos Testes (TCT), comumente chamada apenas de análise clássica.

A análise clássica dos itens de uma prova baseia-se em seus parâmetros descritivos, os quais auxiliam na interpretação da distribuição das respostas para cada alternativa. As propriedades psicométricas dos itens de uma prova correspondem aos seguintes parâmetros: índice de dificuldade (proporção de participantes que responderam ao item corretamente); índice de discriminação, que mede a capacidade do item de diferenciar os participantes de maior habilidade (27% dos respondentes com pontuações mais altas) daqueles de menor habilidade (27% dos respondentes com pontuações mais baixas), correspondendo à diferença entre a proporção de acertos do primeiro grupo e a do segundo grupo; e correlação bisserial entre a resposta numa dada categoria do item e a pontuação total na prova.

No índice de dificuldade, analisa-se o grau de dificuldade de cada item por meio da porcentagem de acerto. Ou seja, quanto menor a porcentagem de acerto maior será o grau de dificuldade.

Já o índice de dizscriminação analisa, para determinado item, as porcentagens de acertos dos grupos de estudantes com melhor e com pior desempenho. Espera-se que, para um item com boa qualidade, a porcentagem de acerto seja maior para o grupo com melhor desempenho, e quanto maior for a diferença entre as porcentagens de acertos dos dois grupos (com melhor e com pior desempenho), maior será a discriminação do item.

O coeficiente bisserial – uma medida de associação entre o desempenho no item e o desempenho na prova – estima a correlação entre a variável de desempenho no teste e uma variável latente (não observável) com distribuição normal que, por hipótese, representa a habilidade que determina o acerto ou erro do item.

O coeficiente bisserial é dado por:

$$r_{bis} = \frac{M^{+} - M^{-}}{S} \cdot \frac{p(1-p)}{h(p)},$$

sendo M^{+} a média da medida de desempenho para os alunos que acertaram o item, M^{-} a média da medida de desempenho no teste para os alunos que erraram o item, S o desvio-padrão da medida de desempenho no teste para todos os alunos, p o percentual de respostas e $h(p)$ o valor da densidade da distribuição normal com média 0 e variância 1 no ponto em que a área da curva à esquerda deste ponto é igual a p.

Todas as propriedades psicométricas descritas estão considerando o escore (número de acertos) do estudante. Entretanto, a utilização do escore para o cálculo dessas medidas só será válida se os estudantes foram submetidos à mesma prova ou provas paralelas (com o mesmo grau de dificuldade).

CONSTRUÇÃO DAS PROVAS DO SAEB

O Instituto Nacional de Estudos e Pesquisa (Inep) vem obtendo informações sobre o desempenho dos alunos dos ensinos

fundamental e médio desde 1995, por meio do Sistema Nacional de Avaliação da Educação Básica (Saeb). Esse sistema avalia o desempenho dos estudantes em Língua Portuguesa e Matemática, a partir da aplicação de testes educacionais a cada dois anos.

O desempenho dos estudantes é avaliado conjuntamente entre as séries pela Teoria da Resposta ao Item, por meio de uma escala única do Saeb com média 250 e desvio-padrão 50, em que a média 250 representa o desempenho médio dos alunos da 8ª série do ensino fundamental de 1997.

O Saeb adotou uma metodologia baseada na amostragem matricial de itens, que utiliza o esquema de montagem e aplicação de testes por Blocos Incompletos Balanceados (BIB). Sob essa estrutura, foram utilizados 13 blocos de itens, contendo 13 itens em cada bloco, distribuídos três a três em cada um dos 26 cadernos de prova, que são distribuídos aleatoriamente entre os alunos de uma mesma turma.

Considerando-se a análise com o BIB, o escore dos alunos apresenta a inconveniência de que os cadernos podem ter graus de dificuldade diferentes e, portanto, há problemas de comparabilidade do número de acertos entre estudantes submetidos a provas diferentes. Por exemplo, dois estudantes com o mesmo número de acertos que foram submetidos a provas diferentes poderiam ter desempenho diferente se tivessem realizado a mesma prova.

PROCEDIMENTOS E OBJETIVOS

Para obter o coeficiente bisserial, em vez de utilizar o escore como medida de desempenho, será utilizada a variável *normit*, a qual é obtida por meio de uma transformação não linear a partir dos escores examinados. Como o *normit* leva em consideração a dificuldade do caderno, os desempenhos dos estudantes submetidos a provas diferentes tornam-se comparáveis.

O *software* mais utilizado para a aplicação da Teoria Clássica dos Testes é o Iteman®. Entretanto, esse *software* considera somente o escore em suas análises, o que poderia trazer conclusões errôneas se usado no Saeb.

A fim de disponibilizar um *software* que utilize corretamente a TCT, considerando cadernos de provas com graus de dificuldade diferentes, os autores desenvolveram uma macro no SAS®.

O objetivo deste trabalho é mostrar a aplicação da metodologia descrita através de uma macro do SAS®, usando as provas do Saeb 2005 aplicadas a estudantes da 3ª série do ensino médio. Além da análise dos dados do Saeb, será utilizado um conjunto de dados simulado a partir de duas provas com graus de dificuldade diferentes.

INFORMAÇÕES SOBRE A MACRO DO SAS®

Para rodar a macro do SAS® é necessário entrar com as informações a seguir:

1. arquivo com a posição dos itens no caderno;
2. arquivo com o gabarito;
3. arquivo com as respostas.

Os arquivos são lidos no mesmo formato do arquivo de dados do Bilog-MG®. Entretanto, é necessário salvar o gabarito separadamente das respostas.

DESCRIÇÃO DOS COMANDOS

LOCALitem: Especifique o local e o nome do arquivo do posicionamento do item na prova.
LOCALresp: Especifique o local e o nome do arquivo de resposta.
LOCALgab: Especifique o local e o nome do arquivo de gabarito.
STARTCAD: Especifique a coluna em que começa o caderno.
ENDCAD: Especifique a coluna em que termina o caderno.
STARTITEM: Especifique a coluna em que os itens começam.
ITENSCAD: Especifique o número de itens que um caderno de prova possui.
NITENS: Especifique o número de itens que serão analisados no caderno de prova.
TOTITEM: Especifique o número total de itens a serem analisados.

5. RESULTADOS DA ANÁLISE CLÁSSICA DOS TESTES NO SAEB 2005

Para exemplificar o funcionamento da macro do SAS® na prova Saeb 2005, foram utilizados os cadernos de prova de Matemática da 3ª série do ensino médio.

Nos 26 cadernos de prova foram considerados 169 itens, com mais de 5.000 respostas em cada item.

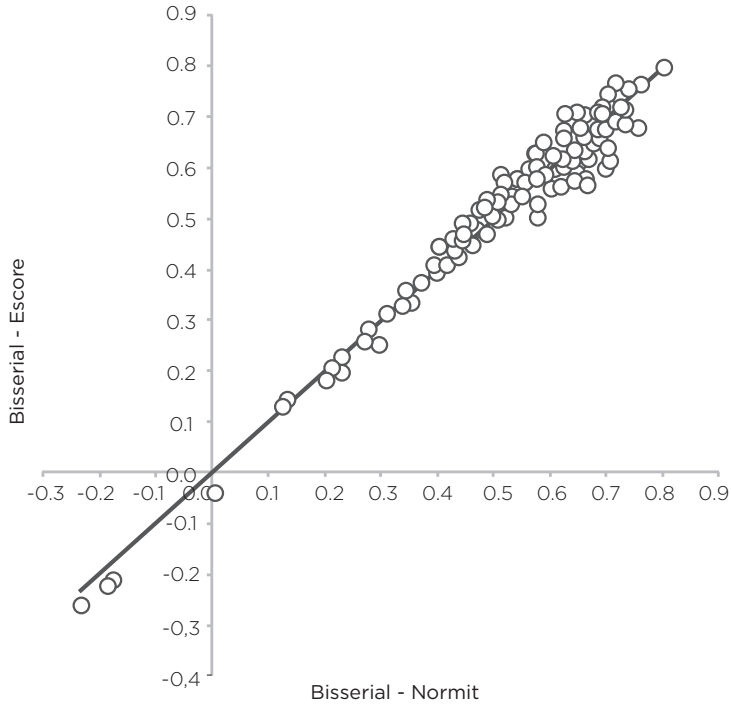
A tabela 1 mostra os resultados da análise clássica dos testes para os 20 primeiros itens obtidos na macro do SAS®. Nota-se que, apesar de a discriminação e o coeficiente bisserial dos itens não serem iguais, os valores obtidos pelo *normit* e pelo escore são bem parecidos.

TABELA 1 – Resultados da análise clássica dos testes

ITEM	GABARITO	DIFICULDADE	DISCRIMINAÇÃO		COEFICIENTE BISSERIAL	
			Normit	Escore	*Normit*	Escore
1	D	0,831	0,401	0,361	0,664	0,568
2	D	0,633	0,252	0,228	0,296	0,254
3	C	0,379	0,522	0,478	0,549	0,545
4	B	0,618	0,700	0,629	0,701	0,640
5	D	0,254	0,553	0,503	0,681	0,700
6	D	0,201	0,151	0,148	0,277	0,286
7	A	0,301	0,474	0,443	0,531	0,536
8	D	0,639	0,641	0,587	0,667	0,616
9	C	0,329	0,367	0,365	0,448	0,456
10	A	0,418	0,835	0,768	0,802	0,799
11	C	0,386	0,630	0,574	0,654	0,629
12	B	0,194	0,122	0,114	0,218	0,217
13	E	0,141	0,094	0,082	0,229	0,206
14	D	0,257	0,385	0,374	0,511	0,548
15	A	0,379	0,390	0,358	0,421	0,429
16	C	0,371	0,625	0,605	0,650	0,674
17	C	0,484	0,730	0,694	0,701	0,678
18	C	0,440	0,228	0,215	0,278	0,268
19	B	0,535	0,741	0,711	0,721	0,685
20	C	0,308	0,424	0,407	0,502	0,531

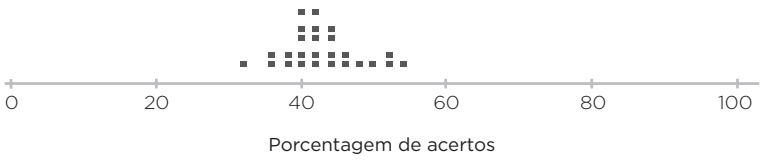
A fim de evidenciar a importância da utilização do *normit*, é apresentada, na figura 1, a comparação do resultado do coeficiente bisserial obtido por meio do *normit* e do escore.

FIGURA 1 – Gráfico de dispersão do coeficiente bisserial para o 3º ano do ensino médio de Matemática



A alta concordância entre os dois resultados mostra que os cadernos de prova foram elaborados com graus de dificuldade parecidos, o que pode ser evidenciado na figura 2. Verifica-se que a porcentagem média de acertos nos 26 cadernos é bastante parecida, sendo que a maior diferença foi encontrada entre o caderno 3, com 31,1%, e o caderno 10, com 54,3%, mas em geral os cadernos apresentam grau de dificuldade parecido. Caso algum dos cadernos de provas tivesse grau de dificuldade muito diferente dos demais, os itens comuns aplicados a esse caderno de prova teriam coeficientes bisseriais bem distintos quando utilizado o *normit*.

FIGURA 2 – Porcentagem média de acertos nos 26 cadernos

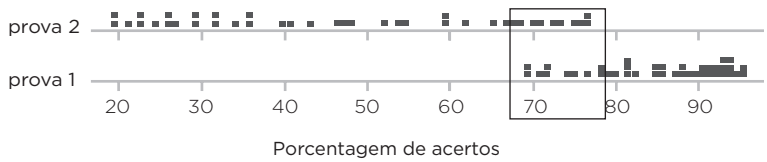


SIMULAÇÃO

A simulação apresentada aqui tem como objetivo comparar resultados de duas provas com 40 itens em cada uma, sendo dez itens comuns.

Para a simulação, foram geradas 1.000 respostas aos 40 itens da prova 1 e mais 1.000 respostas aos 40 itens da prova 2. Os dez itens mais difíceis da prova 1 também foram utilizados na prova 2. Entretanto, esses itens são os mais fáceis na prova 2, tornando evidente a diferença do grau de dificuldade entre as duas provas. A porcentagem média de acertos foi de 84,9%, na prova 1, e de 47,8%, na prova 2. Os 1.000 respondentes de cada uma das duas provas representam amostras aleatórias de uma mesma distribuição de proficiência. A figura 3 traz a porcentagem de acertos em cada uma das provas. O retângulo assinalado na figura mostra a porcentagem de acertos nos dez itens comuns.

FIGURA 3 – Gráfico de pontos para o índice de dificuldade (porcentagem de acertos)

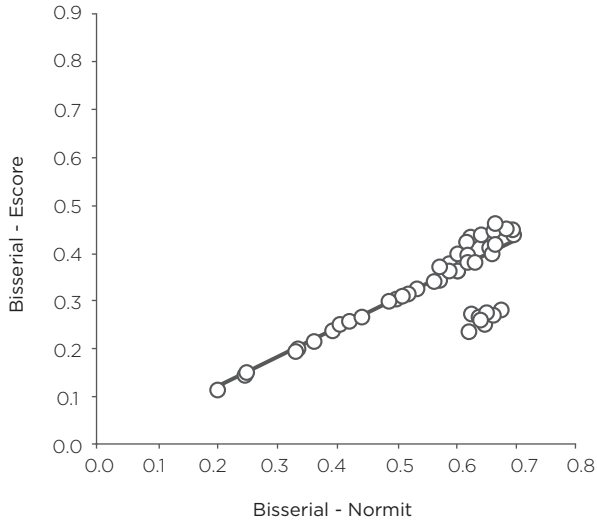


A fim de evidenciar a importância da utilização do *normit* em provas com graus de dificuldade diferentes, a figura 4 apresenta a comparação dos resultados dos coeficientes bisseriais obtidos por meio do *normit* e do escore.

Nota-se que existem pontos que não evidenciam uma relação linear entre os coeficientes bisseriais obtidos pelo *normit* e pelo escore. Esses pontos referem-se aos dez itens comuns às duas provas. Os coeficientes bisseriais calculados pelo *normit* são superiores aos do escore.

Com isso, pode-se inferir que a utilização do escore em itens que foram utilizados em provas com graus de dificuldade diferentes poderá eliminar itens de boa qualidade do banco de itens.

FIGURA 4 – Gráfico de dispersão do coeficiente bisserial para duas provas com grau de dificuldade diferente



CONCLUSÃO

Neste artigo pretendeu-se mostrar a importância em utilizar o *normit* considerando provas com graus de dificuldade diferentes.

A partir dos dados do Saeb 2005 e do estudo de simulação, observou-se que os resultados obtidos pelo escore podem gerar conclusões erradas a respeito do item.

A macro do SAS® desenvolvida facilita a análise clássica dos itens com estrutura BIB e gera resultados mais fidedignos quando as provas têm graus de dificuldade diferentes

REFERÊNCIAS BIBLIOGRÁFICAS

ASSESSMENT SYSTEM CORPORATION. *User's manual for the ITEMAN-Conventional Item Analysis Program*. 2. ed. Windows 3.x/95/NT version. St. Paul, MN: Author, 1998.

FLETCHER, Phillip R. A. Teoria da resposta ao item: medidas invariantes do desempenho escolar. *Ensaio*: avaliação e políticas públicas em educação, v.1, n. 2, p. 21-28, jan./mar. 1994.

FUNDAÇÃO CESGRANRIO. *Relatório da análise clássica do teste Saeb* 2003. jul. 2004.

PASQUALI, L. Provão (ENC) de Psicologia 2000 e 2001: análise dos parâmetros psicométricos. In: PRIMI, R. (Org.). *Temas em avaliação psicológica*. Campinas: Instituto Brasileiro de Avaliação Psicológica, 2002. p. 152-178.

VENDRAMINI, M.M.V.; SILVA, M.C.; CANALE, M. Análise de itens de uma prova de raciocínio estatístico. *Psicologia em estudo*. Maringá, v. 9, n. 3, p. 487-498, set./dez. 2004.

ADRIANO FERRETI BORGATTO
Professor adjunto do departamento de Informática e Estatística do Centro Tecnológico da Universidade Federal de Santa Catarina (UFSC)
borgatto@inf.ufsc.br

DALTON FRANCISCO DE ANDRADE
Professor voluntário junto ao Programa de Pós-graduação do Departamento de Engenharia de Produção da Universidade Federal de Santa Catarina (UFSC)
dandrade@inf.ufsc.br

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