MODELO NOMINAL DA TEORIA DE RESPOSTA AO ITEM: UMA ALTERNATIVA

Igor Reszka Pinheiro[1] *- Universidade Federal de Santa Catarina, Florianópolis, Brasil*
Flávio Rodrigues Costa - Universidade Federal de Santa Catarina, Florianópolis, Brasil
Roberto Moraes Cruz - Universidade Federal de Santa Catarina, Florianópolis, Brasil

RESUMO
Este artigo apresenta o Modelo de Resposta Nominal de Bock como uma alternativa para a parametrização dos itens de múltipla escolha, bem como para obtenção de estimativas mais precisas de traços latentes. Utilizando os dados dos 64.118 sujeitos de nove estados brasileiros e do Distrito Federal, que compõem a amostra normativa do Teste de Raciocínio Lógico Numérico de Costa, comparou-se o ajuste do Modelo de Resposta Nominal com os Modelos Logísticos de 1, 2 e 3 Parâmetros, e efetuou-se a análise das questões, dos distratores e da precisão com os dados politômicos na íntegra e dicotomizados. Conclui-se que, apesar das respostas nominais obtidas empiricamente não terem se encaixado adequadamente aos valores esperados pelo seu modelo, tal modalidade da Teoria de Resposta ao Item ainda ofereceu indícios dos motivos que levaram à má formulação de algumas questões, questionou o entendimento da própria habilidade avaliada e melhorou a precisão da medida.
Palavras-chave: Modelo de resposta nominal; Teoria de resposta ao item; Avaliação educacional.

ITEM RESPONSE THEORY NOMINAL MODEL: AN ALTERNATIVE

ABSTRACT
This paper presents Bock's Nominal Response Model as an alternative to multiple-choice items parameterization as well as to obtain more precise latent trait estimates. Using data from 64,118 subjects gathered in nine Brazilian states and Federal District, which composes the normative sample from the Costa's Logical Reasoning Test, the adjustment of Nominal Response Model was compared with 1, 2 and 3 parameters Logistic Models instances. Also, it was performed the analysis of item issues, distractors and accuracy with full information polytomous items and its dichotomized version. It was concluded that, despite empirically obtained nominal responses do not fit properly to the values expected by its model, this Item Response Theory type of data still leads to some of the reasons of poor question formulation, instigates the understanding of the evaluated ability itself, and improves the precision of the measure.
Keywords: Nominal response model; Item response theory; Educational assessment.

INTRODUÇÃO

Cada vez mais a psicometria cumpre o seu papel na avaliação educacional, fato que se constata pela utilização da Teoria de Resposta ao Item (TRI) na parametrização e na correção do Novo ENEM, do SAEB e de vários outros testes de amplitude nacional. A medida psicológica, nesse contexto, oferece notas de desempenho passíveis de comparação através dos anos, as quais, além de consistirem em um critério meritocrático para admissões, embasam políticas públicas e fornecem informações para a pesquisa, o que favorece professores, alunos, pais, escolas e, até mesmo, o Estado (Bock, 1986).

Sabe-se, porém, que apesar da maioria dessas avaliações empregarem questões de múltipla-escolha mutuamente exclusivas na apresentação dos seus itens, é comum a redução das respostas em padrões de *acerto* ou *erro* para a utilização de modelos dicotômicos da TRI (Vendramini, 2002). Tais modelos conferem aos itens de uma prova parâmetros de dificuldade (Modelo Logístico de 1 Parâmetro – ML1P), dificuldade e discriminação (Modelo Logístico de 2 Parâmetros – ML2P) ou dificuldade, discriminação e acerto casual (Modelo Logístico de 3 Parâmetros – ML3P), dependendo da complexidade da estimação que se deseja.

Como se deduz das equações exibidas na Figura 1, o parâmetro de dificuldade (*b*) presente em todos os modelos indica o ponto no eixo das abscissas em que a probabilidade de acerto do item equivale a 50%. Nesta representação gráfica, por exemplo, o item teria exatamente a mesma possibilidade de acerto ou erro para sujeitos com traço latente (θ) uma unidade de desvio padrão abaixo da média em uma distribuição normal. Já o parâmetro de discriminação (*a*), este presente apenas no ML2P e no ML3P, corresponde à inclinação da

[1] *Contato*:
E-mail: pinheiro_ir@yahoo.com.br.
O primeiro autor deste artigo recebeu auxílio financeiro da CAPES para a produção deste artigo.

tangente que toca a curva de probabilidades, indicando o grau de precisão com que um item é capaz de diferenciar dois valores do traço latente. Por fim, o acerto casual (*c*), que figura somente no ML3P, acena a expectativa de acertos independentemente do traço latente, uma vez que pessoas sem qualquer vestígio da qualidade mensurada ainda responderia corretamente ao item. Seja qual for a opção de estimação, contudo, percebe-se que em todos os modelos dicotômicos a única probabilidade complementar ao acerto é o erro, o que anula a contribuição dos diferentes distratores presentes em testes de múltipla-escolha.

ML1P

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}}$$

ML2P

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}$$

ML3P

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}$$

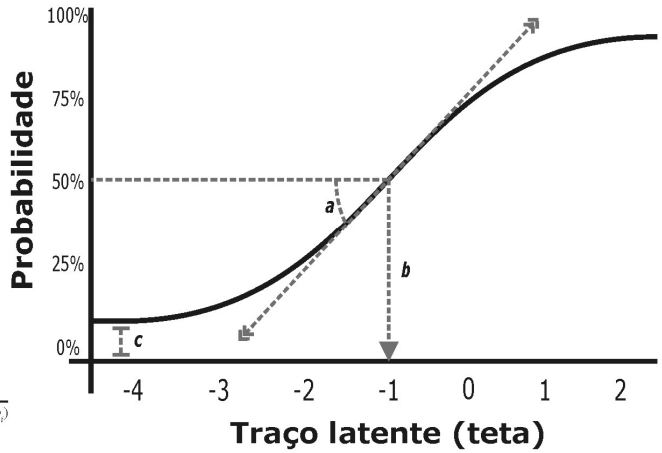

Figura 1. ML1P, ML2P e ML3P com Representação Gráfica [Fonte: Autor]

É intuitiva, entretanto, a noção de que diferentes erros são cometidos por diferentes motivos, mesmo em escalas unidimensionais. Nesse caso, o conhecimento parcial dos sujeitos se revelaria com maior ou menor intensidade dentre as alternativas incorretas de um item, o que, sempre que ponderado, acarretaria em uma maior precisão da medida como um todo (De Ayala, 1989). Esse é o raciocínio que fundamenta o Modelo de Resposta Nominal (MRN), uma alternativa para itens politômicos estimados conforme a TRI.

O MRN de Bock (1972) propõe que cada alternativa de um item seja modelada no nível nominal da medida, consistindo de uma probabilidade cumulativa à escolha correta. Nisso, a equação exibida na Figura 2 apresenta a probabilidade com que um indivíduo *j* selecione cada alternativa de resposta *k* (dentre as m_i opções) de um determinado item *i*, como uma função dos parâmetros de dificuldade[2] (c) e discriminação (a) associados ao traço latente do sujeito. No exemplo de item apresentado, portanto, pessoas com traços latentes até -3 desvios padrão estariam mais propensas a assinalarem a alternativa mais distante daquela considerada correta; pessoas com valores entre -3 e -1,5, mais propensas à marcação da segunda alternativa mais distante; pessoas na faixa de -1,5 e 0, teriam tendência à segunda opção mais próxima de estar correta e, por fim; pessoas acima do ponto 0 em desvios padrão, teriam maior probabilidade de acertar o item. Apesar de, matematicamente, o MRN, desse modo, consistir de uma particularidade do ML2P, percebe-se, nitidamente, que existem várias probabilidades complementares à alternativa correta, o que aumenta a curva de informação do item, possibilita a análise dos diferentes distratores e, ainda, fornece indícios de má formulação de uma questão.

Primeiramente, o ganho na curva de informação se reflete positivamente na precisão das estimativas de traço latente, já que o maior número de parâmetros que acompanham as alternativas de resposta estreita os infinitos pontos ótimos de discriminação. De Ayala (1989), apoiando essa idéia, relata casos em o MRN fornece até duas vezes mais informação que o ML2P aos examinados abaixo da média de sua respectiva população. Esse mesmo autor, em estudo utilizando simulações, ainda afirma que, enquanto o ML3P tende superestimar indivíduos de habilidade extremamente baixa e subestimar indivíduos de habilidade elevada, o MRN somente pouco subestima esses últimos.

Já no que se refere à análise dos distratores, Revuelta (2004) expõe o fato de que existem

[2] A simbologia empregada nos modelos logísticos e no MRN de fato difere. Neste último, o parâmetro de dificuldade se expressa pela letra *c*

inúmeras razões para a preferência por uma alternativa errada em testes de múltipla escolha além do mero *chute*. Quando bem formulados, os distratores presentes nos itens nominais costumam indicar diferentes linhas de raciocínio, o que indica vieses de pensamento, vícios de linguagem, limiares cognitivos e, mesmo, algum conhecimento específico de uma parcela da população. Tal característica do MRN, mais que auxiliar na detecção do funcionamento diferencial de um item, permite a criação de estratégias de ensino capazes de inibir os distratores mais usuais, e o conhecimento precoce de gargalos em alguma etapa da aprendizagem.

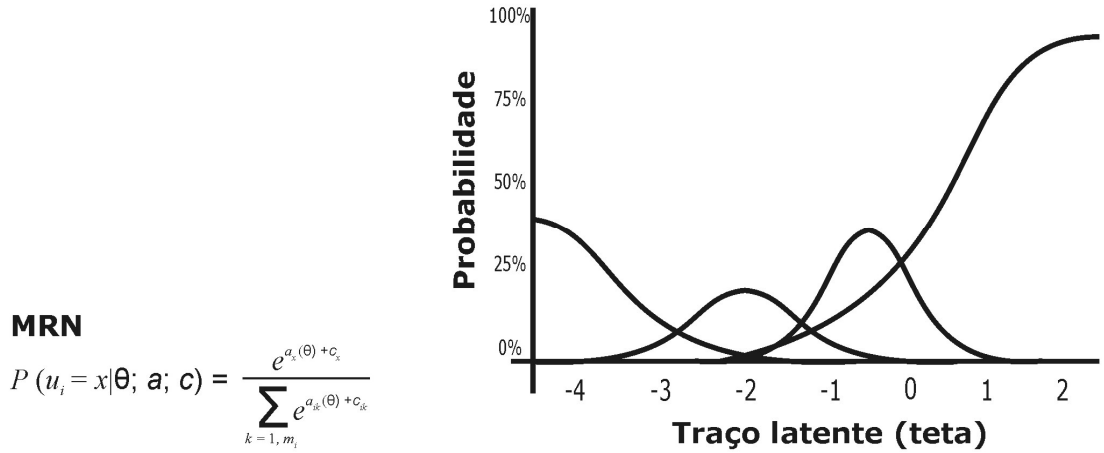

Figura 2. MRN com Representação Gráfica de Item com 4 Alternativas de Respostas [Fonte: Autor]

Por fim, o processamento das respostas por meio do MRN possui uma peculiaridade, pouco usual nos modelos logísticos, que fornece evidências empíricas de má formulação de uma questão: não é necessária a entrada do gabarito (Bock, 1972). É coerente supor que, superadas as limitações de estimativa, a alternativa indicada como correta pelos sujeitos com maior traço latente seja a correta, enquanto as demais alternativas consistam dos distratores. Uma inversão dessa lógica assinala erro de gabarito, problemas de digitação das respostas ou má formulação da questão, já que o traço latente avaliado não avança de forma paralela com as respostas emitidas. Não raro itens dessa natureza são rotulados como *pegadinhas*, ou mesmo findam por serem anulados.

De nada adianta, contudo, todo o aparato teórico do MRN se na prática os resultados de um teste de múltipla escolha qualquer não se ajustarem ao modelo, o que é pouco discutido pelos pesquisadores (De Ayala, 1989, 1999; DeMars, 2003; Revuelta, 2004, 2005; Wollack, Bolt, Cohen, & Lee, 2002; dentre outros), dada a preferência ou necessidade pelos estudos de simulação. O ajuste dos dados ao modelo refere-se ao grau com que o comportamento esperado se equivale ao comportamento obtido empiricamente (Mair, Reise, & Bentler, 2008), o que pode ser acessado por meio de gráficos ou testes estatísticos. Nisso, o ajuste do modelo para a TRI acena, simultaneamente, a sua fidedignidade e a sua validade, pois, quanto menor o valor do erro percebido, mais precisas serão as estimações, bem como as previsões e as extrapolações ao longo do tempo.

Este artigo, por conseguinte, tem como propósito comparar o ajuste dos três tipos de modelos logísticos com o MRN, assim como analisar as questões, os distratores e a precisão com base em seus respectivos indicadores. Acredita-se que, em se tratando dos testes de múltipla escolha, as diferentes categorias de respostas empíricas se ajustarão melhor ao MRN do que as suas correlatas dicotomizadas aos modelos logísticos. Ademais, há a hipótese de que enquanto os modelos logísticos de dois e três parâmetros indicarão seus itens de baixa qualidade genericamente por meio de baixos valores de discriminação e elevados erros padrão, o MRN será capaz de indicar precisamente quando houver má formulação ou vieses específicos nas respostas.

MÉTODO

Participantes

Com o consentimento do criador do teste, utilizou-se os dados dos 64.118 candidatos a cargos públicos, originalmente coletados em 1994 para a normatização do manual. Não existe o registro de dados precisos quanto à distribuição etária ou de

gênero da amostra, toda com formação superior concluída. Os inscritos realizaram as provas em nove estados brasileiros (Bahia, Ceará, Minas Gerais, Pará, Pernambuco, Paraná, Rio de Janeiro, Rio Grande do Sul e São Paulo) e no Distrito Federal, de acordo com o seu domicílio.

Instrumento

Nesta pesquisa utilizou-se o Teste de Raciocínio Lógico-numérico de Costa (1998), o qual é composto por vinte questões de múltipla escolha sendo uma alternativa, e apenas uma, a resposta correta de cada item. Existem sempre cinco opções de resposta, as quais indicam a letra (A, B, C, D ou E) que deverá ser assinalada no gabarito. Sugeriu-se aos aplicadores a utilização de um tempo máximo igual a sessenta minutos, esse podendo ser um pouco estendido caso haja a necessidade do preenchimento de uma folha de respostas externa ao caderno de exercícios. As instruções necessárias para a compreensão do teste encontravam-se na capa do caderno, local em que também foram preenchidos os dados pessoais. Após o início do teste, o aplicador não forneceu mais nenhuma orientação.

Procedimento

Efetuou-se o cálculo dos parâmetros de cada item do Teste de Raciocínio Lógico-numérico, por meio do ML1P, do ML2P, do ML3P e do MRN programados no software MULTILOG. As respostas das vinte questões de múltipla escolha foram consideradas dicotômicas para os três primeiros modelos e nominais para o último. Optou-se por um critério de convergência de 0,0001, o qual foi satisfeito após 39 iterações para o ML1P, 33 iterações para o ML2P, 64 iterações para o ML3P e 34 iterações para o MRN.

Após uma primeira parametrização, utilizada na discussão sobre os itens deficientes, três questões foram retiradas do cálculo e uma nova geração dos parâmetros foi realizada para a obtenção dos traços latentes em todos os quatro modelos. Com base nos dados obtidos pela segunda parametrização, então, foram estimados os valores de todos os 64.118 sujeitos em cada um dos quatro modelos, dentre os quais foram pinçados os 3.000 candidatos com melhor traço latente (excluídos os escores perfeitos), os 3.000 candidatos com pior traço latente (excluídos os escores nulos), os 3.000 candidatos com traços latentes medianos (1.500 acima e 1.500 abaixo da mediana) e 3.000 candidatos aleatórios para o teste de ajuste. Tal procedimento foi realizado com o auxílio do aplicativo MODFIT (Drasgow, Levine, Tsin, Williams, & Mead, 1995), e consistiu da geração dos gráficos de ajuste e do teste de Qui-quadrado ajustado sobre graus de liberdade.

Por fim, normalizou-se os valores de traço latente obtidos por cada um dos quatro modelos (aplicando-lhes média 0 e desvio padrão igual a 1) para fins de comparação. Os resultados obtidos foram submetidos à inspeção gráfica e à análise do criador do teste para discussão. A seguir apresentam-se apenas os resultados mais ilustrativos de cada uma das análises por motivo de economia de espaço, estando os dados completos disponíveis para consulta junto com o primeiro autor do artigo. Por não se tratar de um procedimento linear, havendo vários momentos de retorno à etapa anterior, optou-se pela apresentação conjunta dos resultados e da discussão, pois, do contrário, a inteligibilidade dos dados ficaria comprometida.

RESULTADOS E DISCUSSÃO

Diferentemente do MRN, no qual, por falta de gabarito, todos os itens são calculados independentemente de sua coerência interna, nos modelos logísticos de um, dois e três parâmetros somente são processados os itens que possuem correlação bisserial positiva. Na Tabela 01, portanto, a qual exibe os parâmetros dos três tipos de modelos logísticos em sua primeira calibração, não há valores para o item número dez. Além disso, o ML2P acusa déficit nos itens oito, onze, treze e dezenove, os quais possuem baixa capacidade de discriminação (a < 0,7), mesmo defeito apontado pelo ML3P para o item seis. Ademais, o mesmo item número treze, para o ML2P, também ostenta um parâmetro de dificuldade desproporcional, acompanhado de um elevado erro padrão (6,31), assim como o item oito possui mais de 50% de chance de acerto casual, quando calculado pelo ML3P. Conforme apontado pela literatura (Conde, & Laros, 2007), a principal hipótese levantada quando ocorrem esses tipos de problema nos diferentes modelos dicotômicos é a falta de unidimensionalidade.

A primeira calibração do MRN, por sua vez, não encontrou nenhum valor de erro padrão anormalmente elevado, mas, denunciou falhas nos itens dez e treze, os quais contradisseram o gabarito. Percebe-se na Figura 03, nitidamente, que as alternativas de resposta mais indicadas pelos sujeitos de elevados traços latentes foram *C* e *C*, enquanto inferia-se que o acerto indicaria as alternativas *D* e *A*, respectivamente. Esses resultados empíricos, primeiramente, corroboram com a correlação

bisserial negativa do item número dez durante a parametrização inicial dos modelos dicotômicos, já que a lógica da questão aparentemente foi invertida, fato indicado pelo total antagonismo entre traço latente e probabilidade de acerto. Quanto ao item treze, no qual a inversão é apenas parcial e desfaz-se em longíssimas habilidades (curva decrescente da alternativa *C* e ascendente da alternativa *A*), sabe-se que este está sujeito a alguma variável interveniente, dado o comportamento anormal na faixa de -3 a 3 desvios padrão.

Tabela 1. Parâmetros dos Itens na Primeira Calibração

	ML1P	ML2P		ML3P		
Item	B	a	b	a	b	c
01	-1,15	1,07	-0,98	1,06	-0,98	0,00
02	-1,55	1,24	-1,19	1,23	-1,19	0,00
03	-1,34	1,86	-0,83	1,85	-0,83	0,00
04	-1,88	1,52	-1,28	1,51	-1,28	0,00
05	-3,16	0,78	-3,39	0,76	-3,47	0,00
06	-0,06	0,70	-0,08	0,69*	-0,08	0,00
07	-0,53	1,08	-0,45	1,53	0,05	0,21
08	-1,25	0,63*	-1,59	1,35	0,34	0,51*
09	0,34	0,80	0,35	1,17	0,76	0,16
10	-	-	-	-	-	-
11	0,47	0,59*	0,63	1,00	0,90	0,05
12	0,79	0,86	0,78	1,38	0,42	0,11
13	3,18	0,19*	12,52*	0,76	-0,54	0,09
14	0,20	1,07	0,16	1,62	-0,61	0,08
15	-0,71	0,71	-0,82	1,80	0,42	0,27
16	-1,09	1,52	-0,75	1,02	-0,37	0,02
17	-0,25	0,97	-0,23	2,29	-0,05	0,32
18	-0,47	1,01	-0,42	1,98	-0,62	0,20
19	-0,21	0,15*	-1,04	2,40	-0,54	0,23
20	-0,96	1,30	-0,72	1,90	-1,58	0,09
Média	-0,51	0,95	0,03	1,44	-0,49	0,12
Desv.Pad.	1,29	0,43	3,16	0,51	1,00	0,14

* Valores fora dos padrões aceitáveis por Andrade, Tavares e Valle (2000)

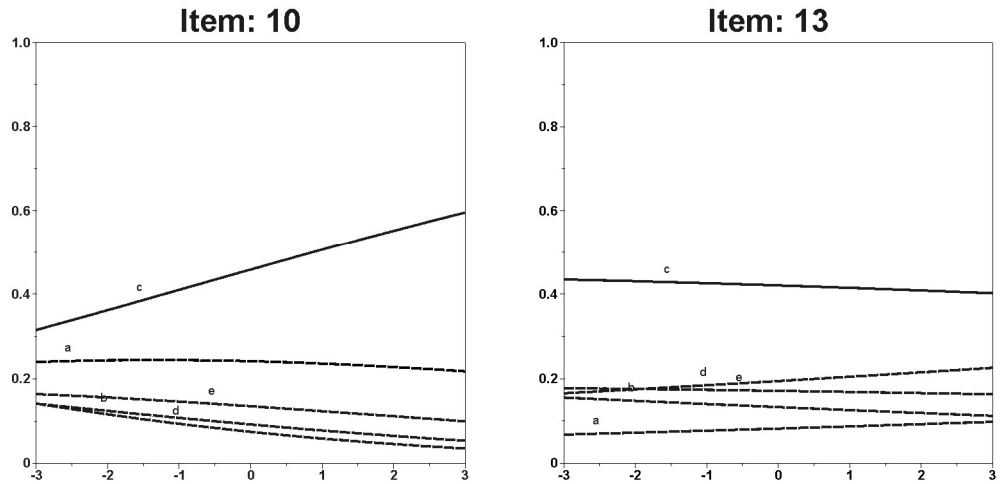

Figura 3. Curvas Características dos Itens 10 e 13 [Fonte: MULTILOG]

Em posterior consulta ao autor do teste, foi relatado que a questão dez, realmente, apresentou uma discutível validade aparente, dada a incompreensão causada pela apresentação de suas

alternativas. O único comentário em relação ao item treze referiu-se à sua extrema dificuldade, o que é claramente exibido pelo ML2P. Ademais, foi indicada a presença de uma *pegadinha* no item número oito, assinalado por um acerto casual de 51% no ML3P, e a possibilidade de um raciocínio paralelo para encontrar uma resposta correta alternativa do item 19, este de baixa discriminação para o ML2P. Enquanto os valores de referência dos modelos logísticos de 2 e 3 parâmetros, de forma genérica, indicaram com grande precisão todos os itens deficitários, o MRN foi facilmente capaz de discernir as instâncias de erro de digitação (item 10) e falta de unidimensionalidade (item 13), traços só detectados nos primeiros dois modelos se realizada uma análise dos resíduos.

Passando, então, para a análise dos distratores, percebe-se em todo o teste um comportamento demasiadamente homogêneo, o que deflagra um padrão de respostas que tende mais à conduta dicotômica, mesmo no MRN. Tomando o item número três como exemplo (Figura 04), visualiza-se os distratores amontoados em uma mesma região da curva característica desse item, enquanto a alternativa correta se sobressai distintamente. Isso anula qualquer ganho de informação almejado através do MRN, uma vez que os distratores não estão segregando adequadamente os diferentes tipos de erros. Há duas hipóteses para serem testadas em futuras pesquisas: 1) todos os distratores indicam a mesma classe de erro, o que não diferencia níveis de traço latente desiguais; e 2) as questões são de natureza heurística e não algorítmica, o que possibilita apenas acertos totais. Enquanto a primeira hipótese sugere investigações quanto aos processos de criação de itens, a segunda requer uma melhor conceituação daquilo que é considerado como raciocínio lógico-numérico.

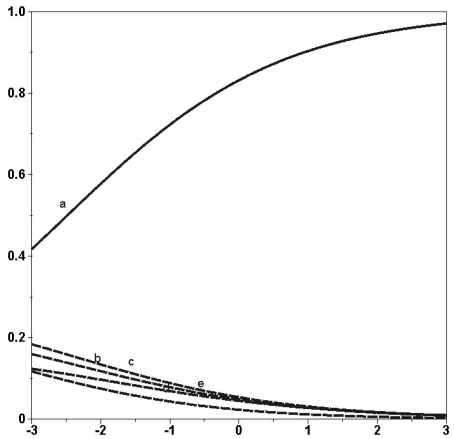

Figura 4. Curva Característica do Item 3 [Fonte: MULTILOG]

Continuando as análises, procedeu-se o cálculo do ajuste dos modelos, o qual pode ser visualizado graficamente nas Figuras 05, 06, 07 e 08, que correspondem ao item vinte nos modelos logísticos de 1, 2 e 3 parâmetros e ao MRN, respectivamente. Em todas as imagens há a comparação dos dados esperados (curvas logísticas) com os dados obtidos empiricamente (distribuição com intervalo de confiança). Quanto melhor for a sobreposição das duas seqüências de dados, mais ajustado é o modelo.

O ajuste do ML1P, visivelmente (Figura 05), está próximo da adequação somente para a amostra aleatória de 3.000 sujeitos dentre a população de 64.118 pessoas estudada. Há uma queda brusca da qualidade do ajuste quando a distribuição se enviesa para o topo dos traços latentes, e outra maior ainda quando se analisam as piores notas. Os valores medianos, apesar de melhores ajustados que as outras duas amostras extremas, também exibem um encaixe bem aquém do desejado para a utilização consistente dessa medida em seu universo restrito.

De modo semelhante se comportam os gráficos de ajuste do ML2P (Figura 06). A distribuição aleatória exibe uma sobreposição bastante adequada, os 3.000 sujeitos com melhores traços latentes estão, na maioria dos casos, fora dos limites estipulados, as respostas das 3.000 pessoas de pior desempenho estão ainda mais desajustadas e, por fim, os valores medianos apresentam uma insuficiente melhora. É interessante notar que, apesar de sutil, houve melhora no ajuste do item vinte, se comparado o ML2P e o ML1P, especialmente nas amostras aleatória e inferior.

Por sua vez, o ajuste do ML3P também exibe uma melhora em relação às suas versões mais simplificadas, estando este bastante adequado na distribuição de 3.000 sujeitos aleatórios (Figura 07). A inserção do parâmetro de acerto casual aparenta conferir uma pequena melhora na sobreposição das notas mais elevadas em troca de um, também pequeno, desencaixe dos resultados mais baixos. O ajuste dos traços latentes medianos não ostenta qualquer diferença significativa.

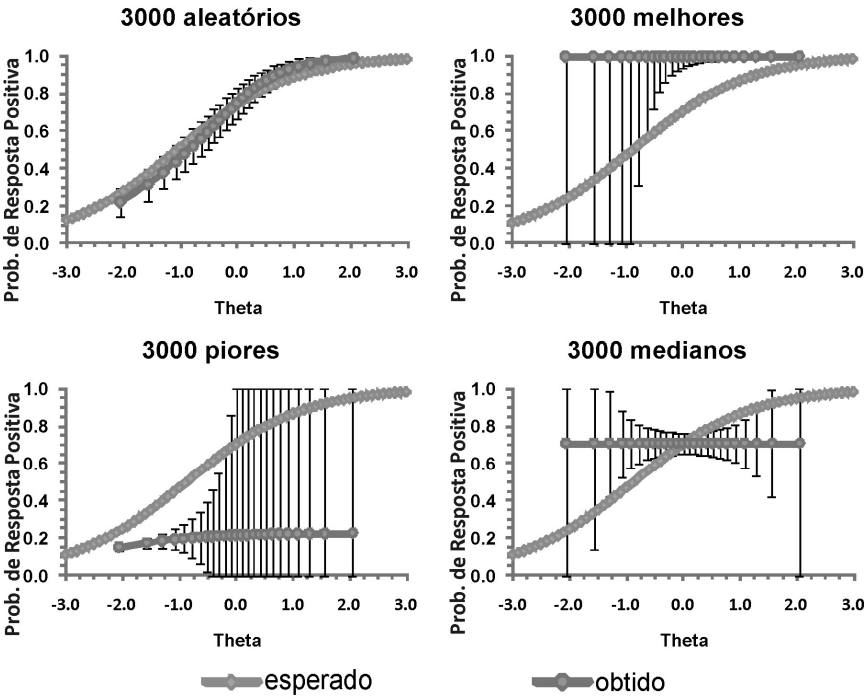

Figura 5. Gráficos de Ajuste do Item 20 no ML1P [Fonte: MODFIT]

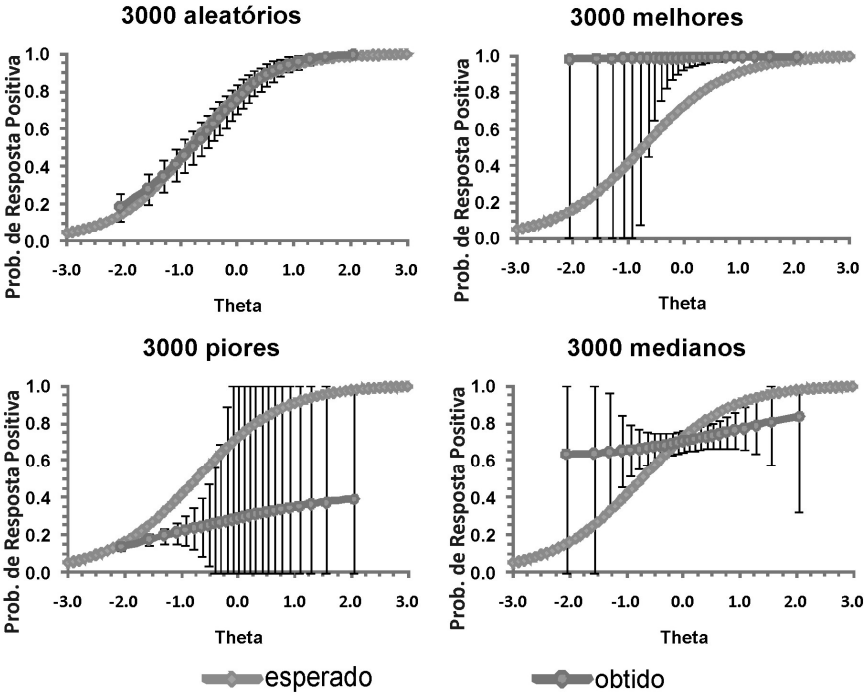

Figura 6. Gráficos de Ajuste do Item 20 no ML2P [Fonte: MODFIT]

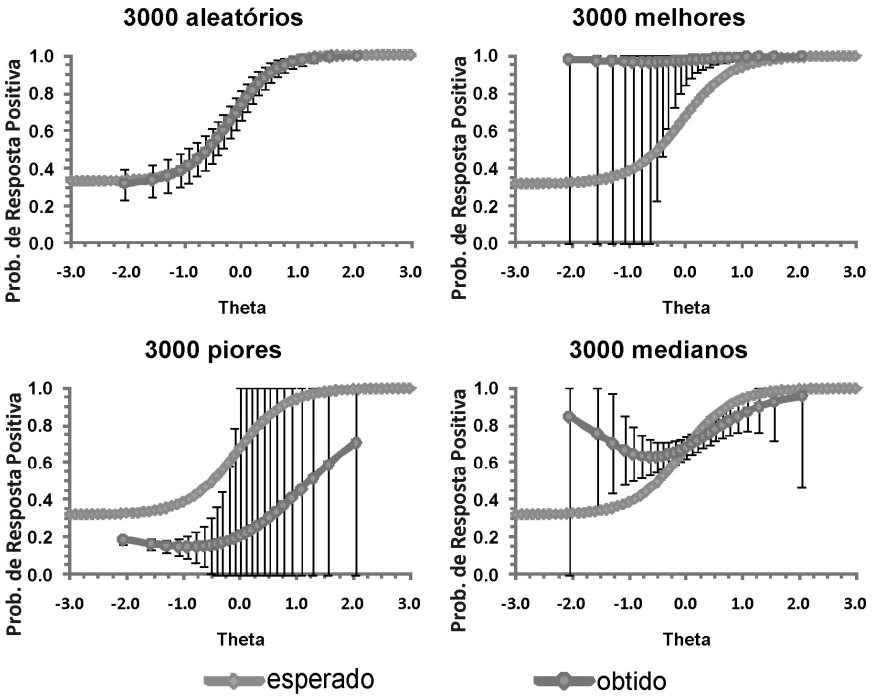

Figura 7. Gráficos de Ajuste do Item 20 no ML3P [Fonte: MODFIT]

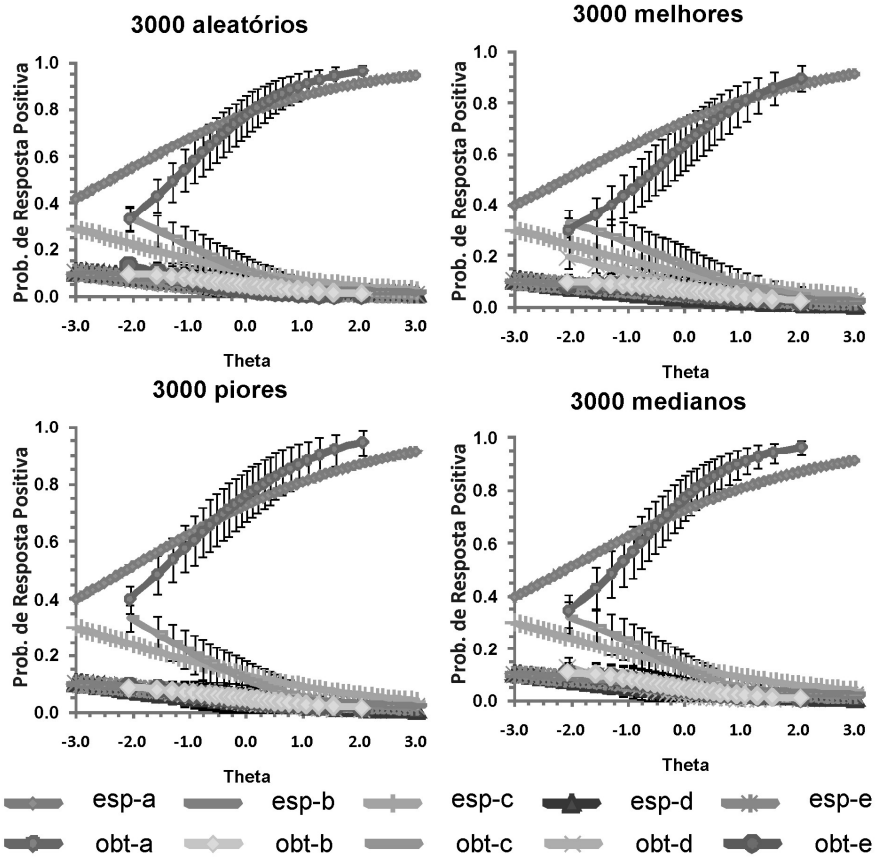

Figura 8. Gráficos de Ajuste do Item 20 no MRN [Fonte: MODFIT]

Finalmente, a quantidade de informação fornecida pelos gráficos de ajuste do item vinte ao MRN (Figura 08) torna difícil a sua comparação com os três modelos logísticos. Enquanto, contudo, o encaixe de cada alternativa em todas as quatro distribuições aparenta deficiência, as diferenças percebidas entre essas quatro modalidades são consideravelmente sutis se comparadas com as exibidas pelos modelos dicotômicos. Se, por um lado, então, no MRN houve perda na qualidade geral do ajuste, este avalizado pela distribuição de 3.000 sujeitos aleatórios, por outro, houve um ganho relativo em todas as demais amostras assimétricas, sempre que comparado com os modelos logísticos de 1, 2 ou 3 parâmetros.

A Tabela 2 resume essas comparações ao apresentar os valores do teste de Qui-quadrado ajustado sobre graus de liberdade, esses interpretados como satisfatórios quando iguais ou menores que três (Hernández, Espejo, & González-Romá, 2006). Nota-se que somente as distribuições aleatórias do ML2P e do ML3P podem ser consideradas adequadas, sendo o último o melhor deles. Para os três modelos logísticos, há, de fato, uma queda brusca na qualidade do ajuste quando utilizadas amostras assimétricas, o que não se constata no MRN. A ocorrência de uma melhora na sobreposição dos dados esperados e obtidos pelos sujeitos medianos do MRN merece investigações futuras. A pior instância, e também o maior contraste, de encaixe nos três modelos dicotômicos refere-se aos mais baixos traços latentes, o que se explica pela perda de informação ocorrida na generalização dos erros.

Tabela 2. Ajuste dos Modelos (Qui-quadrado Ajustado/Graus de Liberdade)

	3000 aleatórios	3000 melhores	3000 piores	3000 medianos
1 Parâmetro	4,95	666,92	1703,03	61,80
2 Parâmetros	2,98*	660,42	1481,57	64,98
3 Parâmetros	2,16*	660,13	1500,88	66,41
Nominal	13,74	21,75	24,18	7,37

* Modelos ajustados conforme valores de referência (Hernández, Espejo, & González-Romá, 2006)

Tal perda de informação, todavia, é perceptível em todas as ocasiões em que diferentes alternativas erradas são computadas por igual. Uma simples análise da precisão dos modelos, logo, foi realizada comparando diferentes desempenhos resultantes de padrões de respostas muito similares. A Tabela 03, como exemplo ilustrativo, exibe as respostas dos sujeitos número 707 e 715, ambos pinçados das 64.118 pessoas avaliadas através dos quatro diferentes modelos. O padrão de respostas é exatamente igual nos três modelos logísticos, havendo dezesseis respostas corretas e uma incorreta (item número 12), e bastante semelhante no MRN, no qual, dezesseis respostas condizem com a alternativa mais correta e uma difere quanto a dois distratores distintos. Nota-se que, apesar de melhores ajustados, os modelos dicotômicos não são capazes de distinguir os dois padrões de respostas, enquanto o MRN os discrimina e indica traços latentes mais precisos (sutilmente menores erros padrão).

Tabela 3. Precisão dos Modelos (Teta em Escala 0,1)

	Sujeito	Respostas																	Teta	Err.Pad.
1 Par	707	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,68	0,554
	715	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,68	0,554
2 Par	707	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,68	0,554
	715	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,68	0,554
3 Par	707	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,75	0,393
	715	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1,75	0,393
Nom	707	4	3	1	4	4	5	2	5	5	5	5	2	3	1	2	5	1	1,39	0,387
	715	4	3	1	4	4	5	2	5	5	5	3	2	3	1	2	5	1	0,86	0,385

Compreende-se as diferentes estimativas de traços latentes dos sujeitos 707 e 715 pelo MRN ao observar a curva característica do seu item doze (Figura 09). Além da alternativa correta segundo o gabarito, *2*, todas as demais opções também distinguem os níveis de conhecimento, atribuindo alguma habilidade parcial mesmo aos distratores. Enquanto, entretanto, o indivíduo 707 indicou a alternativa *5*, distrator mais próximo da resposta correta na faixa de -3 a 3 desvios padrão, o testando

715 optou pela resposta *3*, segundo distrator mais distante da medida de raciocínio lógico-numérico. Isso significa que, em testes de admissão com questões de múltipla escolha, como foi o caso da amostra normativa utilizada por esta pesquisa e na maioria das avaliações educacionais para ingresso em instituições de nível superior, o MRN é menos sujeito aos critérios de desempate alheios ao traço latente medido (como a idade, na maioria dos casos), já que seria necessário um padrão de respostas realmente idêntico, inclusive dentre os distratores, para uma mesma estimação de habilidade. Sugere-se, porém, parcimônia na interpretação de qualquer pequena vantagem em decisões eliminatórias.

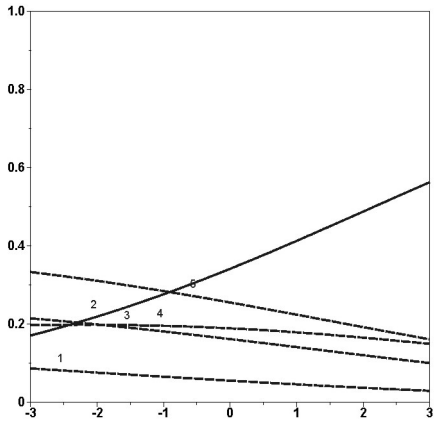

Figura 9. Curva Característica do Item 12 [Fonte: MULTILOG]

Acredita-se que a construção de itens pensados previamente para a estimação de habilidades por meio do MRN aproximaria as curvas características empiricamente obtidas (Figuras 03, 04 e 09, por exemplo), das curvas ideais de discriminação (aproximadamente a Figura 02), o que, por conseguinte, melhoraria ainda mais a precisão da escala, assim como o seu ajuste. É certo que a maior quantidade de informação (variabilidade nas respostas) exigida por este e pelos demais modelos politômicos facilmente é suprimida pelas amplas avaliações educacionais de amplitude nacional. Além disso, o rápido e incessante avanço das pesquisas que utilizam o modelo nominal na TRI promete suprir, muito em breve, psicólogos e educadores com níveis de referência para uma eficaz análise da qualidade de seus itens.

CONSIDERAÇÕES FINAIS

Este artigo revisou as principais qualidades do Modelo de Resposta Nominal da Teoria de Resposta ao Item, e colocou-as à prova em uma comparação com os Modelos Logísticos de 1, 2 e 3 Parâmetros. No que se refere à apreciação dos itens, enquanto os modelos dicotômicos requerem uma análise dos resíduos para inferir algo além de falta de unidimensionalidade, o que leva à exclusão de itens, o MRN, mesmo insipiente, demonstra uma fácil capacidade de indicar as principais causas de problemas, possibilitando a sua correção. Tais problemas, contudo, podem ser acarretados pela própria adaptação de um teste pensado para correção dicotômica ao MRN. Os testes de ajuste sugerem que não é prudente realizar essas adaptações, restringindo a sua utilização às avaliações construídas por itens com distratores carregados de informação.

Restam muitas dúvidas a serem respondidas nos assuntos tangentes ao MRN, o que se espera sanar incitando a curiosidade dos demais leitores pesquisadores. O que leva à aglomeração dos distratores de uma curva característica de item? Por que uma distribuição de habilidades medianas possui melhor ajuste ao MRN que uma amostra aleatória? Quais são os níveis de referência ideais para cada parâmetro de alternativa em um teste múltipla escolha? Especula-se que, antes mesmo da formalização das respostas para essas questões, o MRN da TRI já terá se consolidado como uma alternativa para as avaliações educacionais do Brasil.

REFERÊNCIAS

Andrade, D., Tavares, H., & Valle, R. (2000). *Teoria da resposta ao item: Conceitos e aplicações*. São Paulo: Associação Brasileira de Estatística.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two

or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. (1986). Designing the National Assessment of Educational Progress to serve a wider community of users: a position paper. *Chicago: Study Group on the National Assessment of Student Achivement*. Retirado em 05/12/2009, no World Wide Web: http://www.eric.ed.gov/PDFS/ED279664.pdf .

Conde, F., & Laros, J. (2007). Unidimensionalidade e a propriedade de invariância das estimativas da habilidade pela TRI. *Avaliação Psicológica, 6*(2), 205-215.

Costa, F. (1998). *Teste de raciocínio lógico-numérico: Manual*. São Paulo: Vetor.

De Ayala, R. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Phychological Measurement, 49*, 789-805.

De Ayala, R. (1999). Item parameter recovery for the Nominal Response Model. *Applied Psychological Measurement, 23*(1), 3-19.

DeMars, C. (2003). Sample size and the recovery of Nominal Response Model item parameters. *Applied Psychological Measurement, 27*(4), 275-288.

Drasgow, F., Levine, M., Tsin, S., Williams, B., & Mead, A. (1995). Fitting polytomous IRT models to multiple choice tests. *Applied Psychological Measurement, 19*, 143-165.

Hernández, A., Espejo, B., & González-Romá, V. (2006). The functioning of central categories Middle Level and Sometimes in graded response scales: does the label matter? *Psicothema, 18*(2), 300-306.

Mair, P., Reise, S., & Bentler, P. (2008). IRT goodness-of-fit using approaches from logistic regression. *UC Los Angeles: Department of Statistics, UCLA*. Retirado em 05/12/2009, no World Wide Web: http://escholarship,org/uc/item/1m46j62q.

Revuelta, J. (2004). Analysis of distractors difficulty in multiple-choice items. *Psychometrika, 69*(2), 217-234.

Revuelta, J. (2005). An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika, 70*(2), 305-324.

Vendramini, C. (2002). Aplicação da Teoria de Resposta ao Item na avaliação educacional. Em R. Primi (Org.), *Temas em avaliação psicológica* (pp. 116-127). Campinas: Instituto Brasileiro de Avaliação Psicológica.

Wollack, J., Bolt, D., Cohen, A., & Lee, Y. (2002). Recovery of item parameters in the Nominal Response Model: a comparison of Marginal Maximum Likelihood estimation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339-352.

Recebido em dezembro de 2009
Reformulado em maio de 2010
Aceito em julho de 2010

SOBRE OS AUTORES:

Igor Reszka Pinheiro: designer, especialista em Práticas Pedagógicas Interdisciplinares, doutorando do Programa de Pós-graduação em Psicologia da Universidade Federal de Santa Catarina.

Flávio Rodrigues Costa: psicólogo, mestre em Psicologia Social e da Personalidade, doutorando do programa de Pós-graduação em Psicologia da Universidade Federal de Santa Catarina.

Roberto Moraes Cruz: professor e pesquisador dos Programas de Pós-Graduação em Psicologia e em Engenharia de Produção da Universidade Federal de Santa Catarina. Graduado em engenharia civil pela Universidade Católica de Salvador e em psicologia pela Universidade Federal da Bahia, mestre em educação pela Universidade Federal da Bahia e doutor em engenharia de produção pela Universidade Federal de Santa Catarina.