Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Aplicação da teoria de resposta ao item - modelos unidimensionais aninhados - em um exame multidisciplinar com alunos de graduação

Edy Célia Coelho [1]
Paulo Justiniano Ribeiro Junior [2]
Wagner Hugo Bonat [3]

1 Introdução

As avaliações aplicadas à educação, psicologia, qualidade de vida ou gestão da qualidade em indústrias e empresas, vêm ocupando posição de destaque quando busca-se melhoria da qualidade. Segundo Piton (2012), a procura por melhoria da qualidade na execução dos processos de avaliação tem se tornado cada vez mais urgente nas aplicações nas avaliações educacionais.

A avaliação é um importante instrumento de medição, e ensinar é buscar, ao longo dos anos letivos, credenciar o aluno a absorver cada vez mais conhecimentos. Marcar uma opção em uma prova pode ser um ato de plena consciência do assunto indicado no enunciado, mas, também, pode ser apenas mais um item assinalado na expectativa de não deixar escapar alguma possível chance de acerto.

Para se tentar chegar mais próximo de uma conclusão real do conhecimento de um indivíduo avaliado, a Teoria da Resposta ao Item (TRI) é definida como modelos matemáticos construídos para representar a probabilidade de um indivíduo responder corretamente a um item em determinado teste.

A TRI vem sendo utilizada nacionalmente, principalmente por órgãos de avaliação institucional como o MEC, no âmbito do INEP. Quanto mais rápido o meio educacional tiver contato com esta teoria, a disseminação de seu uso poderá trazer uma benéfica contribuição à sociedade, por tratar-se de um elemento diferencial para melhor conhecer e retratar o conhecimento.

Para avaliar o aprendizado e o desenvolvimento de competências de acordo com o perfil dos estudantes, a Pontifícia Universidade Católica do Paraná (PUCPR) promove, todos os anos, em seus diversos cursos em todos os *campi*, o exame multidisciplinar. A avaliação parte da ideia de que as competências profissionais são construídas ao longo do curso, bem como de que em cada série estas são específicas, embora interajam e dependam umas das outras. Neste contexto, nota-se a busca por avanços nos processos de avaliação educacional, que constitui uma das mais importantes vertentes para a qualidade do ensino.

[1]PPMNEG-UFPR/SEED-PR. e-mail: *edyceliacoelho@gmail.com*
[2]LEG - UFPR.
[3]LEG - UFPR.

A todo momento, em várias situações e de várias formas, os conhecimentos dos indivíduos acerca de determinados assuntos são postos à prova. Alunos são classificados em disputas por uma vaga em instituições de ensino, candidatos concorrem a oportunidades de emprego na área pública, entre outros. Mas há um ponto crucial a ser levado em consideração na escolha ou aprovação de um indivíduo nestes exemplos de seleções: será que a competência que ele realmente adquiriu corresponde ao conhecimento que as questões da prova aparentemente avaliam? Será que a habilidade em resolver a prova realmente avalia o conhecimento que o candidato deve ter?

Estas questões são de relevância para o processo de avaliação educacional, porém, não podem ser resumidas somente a conceitos formais e que envolvam atribuições de notas, obrigatórias à decisão de avanço ou retenção em determinadas disciplinas. Trata-se de verificar um melhor entendimento dos mecanismos geradores de dificuldades na resolução das questões das provas, além de gerar a avaliação por competência buscando o desenvolvimento educacional. Os modelos da TRI podem contribuir neste processo de avaliação, permitindo uma análise mais consistente com a realidade.

Neste contexto, o objetivo deste artigo é descrever a aplicação da Teoria de Resposta ao Item, em um exame multidisciplinar aplicado pela Pontifícia Universidade Católica do Paraná. A ideia é incorporar os procedimentos da TRI em um sistema de avaliação educacional contínua, permitindo, desta forma, uma melhor compreensão dos mecanismos geradores de dificuldades no aprendizado, bem como um acompanhamento do desempenho dos cursos ao longo do tempo, contribuindo, assim, para a melhoria do Ensino Superior como um todo dentro da instituição.

O presente trabalho encontra-se dividido em cinco seções: esta primeira busca introduzir o problema de análise despertando uso dos métodos da TRI; a segunda apresenta a importância da TRI no contexto institucional e aborda os modelos da TRI utilizados nesta pesquisa; a terceira apresenta os principais resultados e discussões e a última apresenta as conclusões e recomendações para trabalhos futuros.

2 Material e métodos

Ao buscar fundamentos na literatura para aplicar a TRI, observa-se que não há muitas obras com foco em avaliação educacional. De acordo com Moreira Junior (2010), a teoria vem sendo utilizada desde 1995, e passou a ganhar popularidade devido às provas do ENEM para a seleção de candidatos em algumas universidades do país, em 2010. Alguns dos autores que abordaram a TRI na área de avaliação relatando as potencialidades da teoria na validade de testes são: Nojosa (2001), Vendramini et al. (2005), Andriola (2008 e 2009), Pasquali (2007, 2009 e 2011), Andrade et al. (2000) e Quaresma et al. (2012).

A maioria das avaliações atuais leva em consideração não apenas uma habilidade latente, mas várias. Em termos de provas multidisciplinares, é razoável pensar que múltiplas habilidades estejam sendo avaliadas, portanto, utilizadas pelos respondentes em cada uma das questões.

Porém, a metodologia clássica de TRI assume que uma habilidade geral é predominante, e é esta que se pretende avaliar.

O objetivo da PUCPR é avaliar o desenvolvimento de competências a partir da aplicação de um exame multidisciplinar. As questões que compõem o exame são elaboradas pelos docentes que atuam em cada curso e validadas pela sua coordenação e por seu Núcleo Docente Estruturante. As questões são então encaminhadas à Diretoria de Graduação e Avaliação Institucional, na Pró Reitoria Acadêmica. Ali é feita uma conferência na editoração das questões e, em seguida, a composição das provas. Após a aplicação, os gabaritos são retornados à Diretoria de Graduação e Avaliação Institucional para a correção e posterior divulgação dos resultados e emissão de relatórios. Cada coordenação de curso utiliza-se dos relatórios para organizar a atividade devolutiva aos alunos, abordando especialmente aquelas questões cujo índice de acerto tenha sido muito baixo, e reflete sobre a elaboração daquelas que tiveram índice de acerto muito alto.

Para o presente resumo, foi selecionada do banco de dados, disponível pela instituição, uma amostra com o maior numero de alunos do segundo semestre de 2011, de um único curso de graduação com trezentos e quarenta e três (343) alunos que realizaram a avaliação multidisciplinar. Dentre eles, foram escolhidos somente os do segundo período que realizaram a mesma prova, totalizando noventa e quatro (94) discentes da graduação. Portanto a questão nesse caso passa a ser, quão pequena é a amostra e os 30 itens que compões o exame, que foram elaborados pelos professor sem serem pré-testados. Por esse motivo, os dados em estudo é desafiador, mas é a realidade das avaliações educacionais internas. O fato é que se o modelo TRI, ajustar-se com bom comportamento, mostrado resultados interessantes e coerentes no conjunto de dados proposto é possível analisa-lo e também verificar os problemas numéricos e computacionais, por permitir inspecionar mais facilmente respondente e item.

O conjunto de dados foi processado pela instituição deixando em anonimato o nome do curso devido à ética profissional, até obter resultados considerados finais para a pesquisa. As notas foram processadas pelo método tradicional de sumarização de certo e errado pela Teoria Clássica dos Testes (TCT). Inicialmente, os resultados do conjunto de dados do exame multidisciplinar foram descritivamente analisados, contemplando uma única habilidade - a geral, que visa à formação do acadêmico.

O exame multidisciplinar para o curso e período selecionados, é composto por seis blocos (seis disciplinas) e, sem dúvida, todos fazem parte de uma habilidade geral, que é a formação do profissional. A prova é composta de seis habilidades, e compreende as seguintes disciplinas: Geometria Analítica e Álgebra Linear B (GAA); Cálculo Diferencial e Integral B (CDI); Física Geral e Experimental B (FGE); Química Geral e Inorgânica (QGI); Estatística (EST); e Introdução Experimental à Química (IEQ). Os resultados para as devidas comparações foram averiguados observando-se as seis habilidades, além da habilidade geral (HG). Cada disciplina abordou cinco (5) itens.

A princípio, os modelos TRI supõem que há uma única habilidade latente sendo medida.

Contudo, a prova multidisciplinar é composta de seis ”sub-habilidades”, as quais fazem parte de uma habilidade geral. Pensando nisso, a análise foi conduzida de duas formas: a primeira considerou que cada disciplina é uma habilidade latente e sua prova foi analisada separada das demais. A segunda, que a prova como um todo mede uma habilidade e todas as questões foram avaliadas conjuntamente. O objetivo é verificar se as abordagens apresentam diferenças relevantes e qual é a abordagem mais adequada para a presente situação.

Os modelos unidimensionais da TRI descrevem a relação entre as respostas observadas ao item e um traço latente, usualmente simbolizado por θ, que pode ser visto com a habilidade do individuo. Tais modelo são apropriados para dados nos quais um único fator comum está sendo avaliado pelos itens. Com a aplicação da TRI dentro da instituição, busca-se um melhor entendimento das provas que levem à compreensão de como o conhecimento por disciplina está sendo avaliado, bem como fornecer diagnósticos e subsídios para a implementação ou manutenção das metodologias educacionais.

Os modelos básicos da TRI podem ser vistos como modelos de efeitos aleatórios, com a particularidade que, na sua forma básica, não há parâmetros de variância para estimar explicitamente. Considere o caso de um teste formado por i questões, em que j indivíduos são avaliados. O modelo logístico de três parâmetros postula que a probabilidade de um indivíduo qualquer responder corretamente a cada uma das questões (Pij) envolve, para cada um dos itens: (i) a habilidade latente do individuo θ_j; (ii) a dificuldade β_i; (iii) a discriminância α_i; e (iv) a probabilidade de acerto casual c_i. A equação do modelo logístico de três parâmetros é:

$$P_{ij} = P(Y_{ij} = 1|\theta_j) = c_i + (1 - c_i) \cdot \frac{1}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \tag{1}$$

Pode-se identificar facilmente dois casos particulares. O primeiro, quando a probabilidade de acerto casual é desprezível, ou seja, $c_i = 0$. O segundo, quando a discriminância é igual para todas as questões, ou seja, $\alpha_i = \alpha$. No caso de $\alpha = 1$, tem-se o conhecido modelo de Rasch, detalhes em (RASCH, 1960; BIRNBAUM, 1968 e BOCK & LIEBERMAN, 1970).

Denote por B(n,p) a distribuição de probabilidade Binomial com parâmetros n e p e também, N(0,1) a densidade da distribuição Normal com média igual a zero e variância igual a 1 (Normal padrão). O modelo completo descrito de forma hierárquica é: $Y_{ij}|\theta_j \sim B(n = 1, P_{ij})$,
$\theta_j \sim N(0,1)$.

Para fazer inferência sobre os parâmetros deste modelo, é necessário a obtenção da verossimilhança marginal, obtida após a integração dos efeitos aleatórios, neste caso, as habilidades latentes θ_j. O integrando desta verossimilhança é o produto de uma binomial por uma gaussiana padrão, e não tem solução analítica. Desta forma, é necessário usar métodos para integração numérica. Detalhes da inferência sobre os parâmetros deste modelo podem ser consultados em (BONAT et al, 2012). Mais detalhes sobre os fundamentos da TRI e os modelos matemáticos utilizados podem ser encontrados em (VAN DER LINDEN e HAMBLETON, 1997; BAKER,

2001 e RECKASE, 2009).

Neste estudo para a estimação dos parâmetros dos itens utilizou-se o Software R, através do pacote ltm (modelos de variáveis latentes para dados dicotômicos). Os parâmetros foram estimados pela abordagem de Máxima Verossimilhança Marginal (RIZOPOULOS, 2011).

O modelo apresentado em (1) é bastante geral e supõe que cada item é descrito por 3 parâmetros, dificuldade, discriminância e acerto casual. Isso pode não ser adequado para situações por exemplo, onde a probabilidade de acerto casual seja desprezível, ou que existam diversas questões com discriminância não diferentes. Para levar estas possibilidades em consideração, considerou-se neste artigo um total de nove combinações do modelo geral, cada uma com certa particularidade, a fim de com a comparação destes modelos chegar ao que melhor descreve a realidade da prova em questão. A Figura 1 apresenta o conjunto de modelos construídos de acordo com a suposição para cada parte do conjunto de parâmetros.

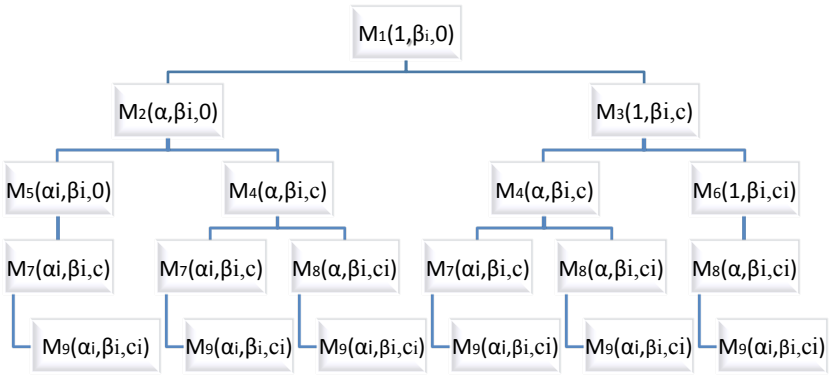

Figura 1: Diferentes modelos da teoria de resposta ao item seguindo a hierarquia

A Figura 1 mostra que partindo do modelo M1, o mais simples considerado, pode-se ir incluindo parâmetros até chegar ao modelo mais complexo, o M9. Este caminho passa por diversos modelos que são encaixados possibilitando o uso do teste de razão de verossimilhança.

O objetivo é estimar os nove modelos e compará-los a fim de encontrar o que melhor descreve a prova em análise. Conforme Reckase (2009), esta comparação pode tanto ser feita pelo teste de razão de verossimilhanças quando os modelos são encaixados, quanto pelo Critério de Akaike (AIC) ou pelo Critério Bayesiano (BIC). As três possibilidades foram contempladas na análise.

3 Resultados e discussões

Como em toda análise estatística de dados, é interessante começar com análises descritivas simples, a fim de tomar afinidade com os dados. A Tabela 1 apresenta a proporção de acertos para cada uma das 30 questões que compõem o exame multidisciplinar para o curso e período em estudo. Para uma melhor compreensão, numeraram-se as questões dentro das habilidades de 1 a 5, e na prova, de 1 a 30.Essa codificação será utilizada em toda a análise. De forma

geral, o que chama atenção na Tabela 1 é que em todas as questões a proporção de acertos está abaixo de 50%. A disciplina de GAA apresenta duas das questões com mais baixa proporção de acertos: as questões Q12(10%) e Q15(13%). Por outro lado, a disciplina GAA também apresenta a questão com maior proporção de acertos Q11(47%).

Tabela 1: Proporção de acertos por questões

Habilidade Geral	H. Q	% acerto	Habilidade Geral	H. Q	% acerto
QGI			CDI		
Q1	Q 1	44 %	Q16	Q 1	30%
Q2	Q 2	20%	Q17	Q 2	27%
Q3	Q 3	18%	Q18	Q 3	37%
Q4	Q 4	17%	Q19	Q 4	40%
Q5	Q 5	30%	Q20	Q 5	24%
IEQ			FGE		
Q6	Q1	18%	Q21	Q 1	15%
Q7	Q 2	39%	Q22	Q 2	23%
Q8	Q 3	24%	Q23	Q 3	39%
Q9	Q 4	49%	Q24	Q 4	31%
Q10	Q 5	39%	Q25	Q 5	15%
GAA			EST		
Q11	Q1	47%	Q26	Q 1	27%
Q12	Q 2	10%	Q27	Q 2	37%
Q13	Q 3	16%	Q28	Q 3	26%
Q14	Q 4	27%	Q29	Q 4	35%
Q15	Q 5	13%	Q30	Q 5	38%

FONTE: O autor (2012)
NOTA:Habilidade por questão (HQ) e Questão (Q).

Essa disciplina é a que apresenta um maior distanciamento entre a proporção de acertos das questões. Nas demais disciplinas, os acertos foram homogêneos. Para verificar a consistência interna do instrumento de avaliação e a correlação entre os itens avaliados, a Tabela 2 apresenta o *alfa de Cronbach* e a correlação ponto bisserial, para cada item, levando em consideração todas as questões e as habilidades específicas.

O coeficiente *alfa de Cronbach* (AC) é uma medida da consistência interna do instrumento de avaliação. Seu valor varia de 0 a 1, sendo que valores próximos de 1 indicam alta consistência. Foi avaliado para cada habilidade com todas as questões e retirando-se uma a uma. Por exemplo, para a habilidade QGI, o AC apresentou o valor de 0,60, quando avaliado com todas as questões desta habilidade. Por outro lado, apresentou o valor de 0,57 quando foi retirada a questão Q3.

Quando trata-se as habilidade de forma independente, verifica-se de forma geral que a consistência das provas é baixa, variando de 0,60 para QGI até 0,31 para GAA. Destaca-se ainda na Tabela 2 as questões Q1 de GAA e Q4 de GAA, como as que têm maior influência no AC.

Com relação à correlação ponto bisserial (PB), ela é análoga ao coeficiente de correlação de Pearson, porém, adequada para verificar a correlação entre variáveis categóricas. Essa correlação é avaliada dentro do contexto de TRI para verificar a concordância da questão com a prova como um todo. Considere o escore em um teste como o total de acertos do indivíduo. A correlação PB mede a correlação entre a resposta em um determinado item com o escore total da prova. Segundo Pasquali (2011), a idéia é que se a questão apresenta boa aderência ao instrumento de medida, ela deve apresentar uma correlação PB acima de 0,3. O escore total

pode ser calculado considerando o item ao qual se está testando, ou sem o item o que é de mais interesse. Por isso, a Tabela 2 apresenta as colunas (incluir) e (excluir).

Tabela 2: Correlação ponto bisserial e alfa de Cronbach's

Habilidades Questão (Q)	Ponto Bisserial com Total incluir	excluir	Cronbach's alpha *Todas as questões e Excluindo a questão	Habilidade Geral	Ponto Bisserial com Total incluir	excluir	Cronbach's alpha *Todas as questões e Excluindo a questão
QGI			0,60*				0,82*
Q 1	0,61	0,29	0,59	Q1	0,59	0,53	0,81
Q 2	0,61	0,36	0,55	Q2	0,32	0,25	0,82
Q 3	0,56	0,31	0,57	Q3	0,42	0,36	0,82
Q 4	0,65	0,43	0,52	Q4	0,44	0,38	0,82
Q 5	0,68	0,42	0,52	Q5	0,47	0,40	0,82
IEQ			0,49*				
Q 1	0,40	0,12	0,51	Q6	0,20	0,13	0,83
Q 2	0,57	0,25	0,45	Q7	0,41	0,33	0,82
Q 3	0,59	0,32	0,40	Q8	0,41	0,34	0,82
Q 4	0,64	0,32	0,39	Q9	0,62	0,56	0,81
Q 5	0,62	0,32	0,40	Q10	0,52	0,45	0,82
GAA			0,31*				
Q1	0,68	0,26	0,14	Q11	0,59	0,53	0,81
Q 2	0,24	-0,04	0,39	Q12	0,23	0,18	0,82
Q 3	0,40	0,05	0,35	Q13	0,20	0,14	0,83
Q 4	0,66	0,29	0,12	Q14	0,36	0,29	0,82
Q 5	0,48	0,18	0,25	Q15	0,24	0,18	0,82
CDI			0,54*				
Q 1	0,53	0,23	0,53	Q16	0,23	0,15	0,83
Q 2	0,62	0,35	0,46	Q17	0,39	0,31	0,82
Q 3	0,61	0,32	0,48	Q18	0,41	0,33	0,82
Q 4	0,61	0,30	0,49	Q19	0,54	0,47	0,81
Q 5	0,58	0,32	0,48	Q20	0,36	0,28	0,82
FGE			0,34*				
Q 1	0,53	0,24	0,24	Q21	0,21	0,15	0,83
Q 2	0,54	0,18	0,28	Q22	0,30	0,23	0,82
Q 3	0,52	0,09	0,38	Q23	0,48	0,40	0,82
Q 4	0,58	0,20	0,26	Q24	0,36	0,29	0,82
Q 5	0,45	0,14	0,32	Q25	0,35	0,29	0,82
EST			0,43*				
Q 1	0,58	0,28	0,34	Q26	0,41	0,34	0,82
Q 2	0,58	0,24	0,36	Q27	0,54	0,47	0,81
Q 3	0,41	0,08	0,47	Q28	0,33	0,25	0,82
Q 4	0,56	0,22	0,38	Q29	0,46	0,38	0,82
Q 5	0,62	0,29	0,32	Q30	0,46	0,38	0,82

FONTE: O autor (2012)

NOTA: No Cronbach alpha para verificar a confiabilidade dos itens com a notação (*) significa, que todas as questões são consideradas e sem o asterisco a questão é excluída

De forma geral, verifica-se que avaliando a prova toda, as duas medidas descritivas tendem a indicar melhores resultados, em relação a quando avaliamos a prova por disciplinas. A disciplina de GAA foi a que chamou mais atenção, apresentando diversas questões com baixa aderência e pouca consistência interna.

É importante analisar o que aconteceria se todos os alunos estivessem respondendo ao acaso às questões de uma prova de múltipla escolha. Para verificar qual seria esse comportamento, a Tabela 3 apresenta a frequência de acertos observada e o que é esperado sob a hipótese de que

os alunos estão respondendo aleatoriamente a todas as questões.

Na Tabela 3, observa-se que a maior frequência de acertos concentrou-se entre 7 e 14 questões. Dos 94 alunos que fizeram a prova, esperava-se que ninguém zerasse, sob a suposição de acerto casual. Porém, observou-se 19 alunos que não tiveram acerto em nenhum item da prova.

Tabela 3: Frequência de acertos do valor observado e o número esperado de acerto casual da habilidade geral

Acertos	Frequência	N. Esperado casual	Acertos	Frequência	N. Esperado casual
0	19	0,12	11	4	1,5
1	0	0,9	12	7	0,6
2	0	3,1	13	7	0,2
3	0	7,3	14	7	0,06
4	2	12,4	15	3	0,02
5	3	16,2	16	6	0
6	3	16,9	17	1	0
7	8	14,5	18	2	0
8	9	10,4	19	0	0
9	6	6,3	à	...	...
10	7	3,3	30	0	0

FONTE: O autor (2012)
NOTA: * Entre 19 à 30 a frequências de acertos e o número esperado de acerto casual foi zero.

Este é um resultado bastante relevante, pois pode indicar pelo menos duas situações que merecem atenção. A primeira é que um grupo de alunos optou por não fazer a prova e marcou as questões que sabiam estar erradas. A segunda é que eles não objetivam responder à prova de forma aleatória, mas, pelo contrário, tentam respondê-la corretamente, porém, devido a sua baixa habilidade, erram. De uma forma ou de outra, este é um resultado que indica que melhorias são necessárias, seja nas questões, seja no ensino ou mesmo na conscientização dos alunos quanto a responder à prova de forma responsável.

Observa-se também que entre 4 e 8 acertos o valor observado foi sempre menor do que a esperança sob acerto casual. Nota-se que ninguém acertou da primeira à terceira (1 a 3) questão; em seguida, 75 alunos acertaram entre 4 e 18 questões e, dentre a décima nona e trigésima (19 a 30) questão, não houve acerto, ou seja, ninguém gabaritou a prova.

Do ponto de vista da análise estatística, o conjunto de dados é bastante desafiador para a aplicação de modelos da TRI. A base é relativamente pequena ,apenas 94 respondentes, e ainda apresenta uma quantidade considerável de respondentes sem nenhum acerto. Por outro lado, com o conjunto pequeno, é por vezes conveniente para verificar problemas numéricos e computacionais por permitir inspecionar mais facilmente caso a caso, respondente ou item. Da mesma forma que a parte descritiva, o ajuste dos modelos TRI será conduzido por disciplina e com a prova como um todo.

Portanto, para cada disciplina e para a prova toda foram ajustados 9 modelos, e para cada um destes, foram calculados a log-verossimilhança marginal, o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC). Os modelos ajustados conforme a suposição

para os parâmetros de discriminância, dificuldade, e acerto ao acaso (α,β,c) foram: M_1 (1,β_i,0); M_2 (α,β_i,0); M_3 (1,β_i,c); M_4 (α,β_i,c); M_5 (α_i,β_i,0); M_6 (1,β_i,c_i); M_7 (α_i,β_i,c); M_8 (α,β_i,c_i); M_9 (α_i,β_i,c_i).

Para a comparação, de forma geral, busca-se o modelo com menor número de parâmetros (df), por ser de mais simples interpretação e estimação. Como recomendação geral, quanto menores o AIC e o BIC, melhor será o ajuste. A Tabela 4 apresenta as medidas de comparação para todos os modelos ajustados para as habilidades específicas e a geral.

Tabela 4: Habilidades dos alunos com os diferentes modelos da TRI

HABILIDADES	MODELOS (α, β, c)								
DOS	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
ALUNOS	$(1,\beta_i,0)$	$(\alpha,\beta_i,0)$	$(1,\beta_i,c)$	(α,β_i,c)	$(\alpha_i,\beta_i,0)$	$(1,\beta_i,c_i)$	(α_i,β_i,c)	(α,β_i,c_i)	(α_i,β_i,c_i)
GERAL (94 alunos)									
Parâmetros	df (30)	df (31)	df (31)	df (32)	df (60)	df (60)	df (61)	df (61)	df (90)
logLik	-1464,79	**-1461,20**	-1464,79	-1461,20	-1443,22	-1463,82	-1443,22	-1453,27	-1443,22
AIC	2989,57	**2984,41**	2989,57	2984,41	3006,44	3047,64	3006,44	3028,55	3066,44
BIC	3063,87	**3063,25**	3065,87	3063,25	3159,04	3200,24	3159,04	3183,69	3295,34
6 HABILIDADES (94 alunos)									
Parâmetros	df (5)	df (6)	df (6)	df (7)	df (10)	df (10)	df (11)	df (11)	df (15)
QGI									
logLik	-240,51	**-238,44**	-240,51	-238,44	-237,36	-240,51	-237,36	-237,69	-239,94
AIC	491,01	**488,89**	491,01	488,89	494,72	501,02	494,72	497,38	509,88
BIC	503,73	**504,15**	503,73	504,15	520,15	526,46	520,15	525,35	548,03
IEQ									
logLik	**-278,36**	-278,33	-278,36	-278,33	-276,66	-276,65	-276,66	-276,90	-276,66
AIC	**566,72**	568,66	566,72	568,66	573,33	575,3	573,33	575,81	583,33
BIC	**579,44**	583,92	579,44	583,92	598,76	600,73	598,76	603,79	621,48
GAA									
logLik	-223,32	-223,14	-223,32	**-220,91**	-218,92	-220,61	-218,02	-219,45	-218,02
AIC	456,64	458,28	456,64	**453,82**	456,04	461,22	456,04	460,91	466,03
BIC	469,36	473,54	469,36	**469,08**	481,48	486,66	488,88	469,36	504,18
CDI									
logLik	**-276,87**	-276,53	-276,87	-276,53	-275,26	-275,44	-275,26	-276,37	-275,26
AIC	**563,74**	565,06	563,74	565,06	570,51	570,89	570,51	574,74	580,51
BIC	**576,46**	580,32	576,46	580,32	595,95	596,33	595,95	602,72	618,66
FGE									
logLik	**-248,02**	-247,70	-248,02	-247,70	-245,52	-247,02	-245,53	-246,12	-245,39
AIC	506,04	507,4	506,04	507,4	511,03	514,03	511,06	514,24	520,78
BIC	**518,76**	522,66	518,76	522,66	536,47	539,47	536,5	542,21	558,93
EST									
logLik	**-286,70**	-286,60	-286,70	-286,67	-284,32	-285,53	-284,32	-285,32	-284,38
AIC	**583,4**	585,21	583,4	585,34	588,64	591,07	588,64	592,64	598,77
BIC	**596,12**	600,47	596,12	600,6	614,07	616,51	614,07	620,62	636,92

FONTE: O autor (2013)

NOTA: Os valores em negrito representa as escolhas do modelos conforme a logLik e os criterios estabelecidos do AIC e BIC.

Para escolha do modelo, considerou-se como exemplo a habilidade geral. Analisando os valores da logLik conforme o número de parâmetros (df), verifica-se que com o aumento da complexidade do modelo (mais parâmetros sendo estimados), o valor da logLik também cresce, o que indica que o procedimento computacional está coerente.

Observa-se que os modelos M_5 (α_i,β_i,0); M_7 (α_i,β_i,c); M_9 (α_i,β_i,c_i) foram os que obtiveram os melhores ajustes, ressaltando que os valores foram iguais a $-1443,22$. O motivo de os modelos possuírem o mesmo valor é porque o valor do acerto casual (c) foi para ambos estimado igual a zero.

Considerando os critérios AIC e BIC, observa-se que os modelos com menor AIC e BIC são o M_2 e o M_4. Porém,este último apresenta a mesma logLik que o M_2, pois o valor do parâmetro c foi estimado em zero, recaindo assim ao M_2. Desta forma, opta-se por escolher o modelo

M_2, pois ele apresenta um ajuste não diferente do modelo mais complexo M_9, porém, com 59 parâmetros a menos.

Os mesmos critérios foram aplicados para a escolha dos modelos das habilidades específicas. Os modelos escolhidos foram: para a habilidade QGI o M_2 (α,β_i,0), para IEQ, CDI, FGE e EST o M_1 (1,β_i,0), e para a GAA optou-se pelo M_4 (α,β_i,c). Tais modelos estão marcados em negrito na Tabela 4.

O resultado do ajuste dos modelos podem ser vistos através das Curvas Características do Itens (CCI), as quais apresentam toda a informação relevante proveniente do modelo. A CCI relaciona através de um gráfico a habilidade dos indivíduos, com a probabilidade de estes responderem corretamente a cada um dos itens. Ou seja, é um gráfico onde no eixo X está a habilidade, e no eixo Y a probabilidade de acerto, dado uma determinada habilidade.

A Figura 2 apresenta a CCI para o modelo ajustado com os trinta itens que compõem o exame multidisciplinar,considerando o modelo M_2.

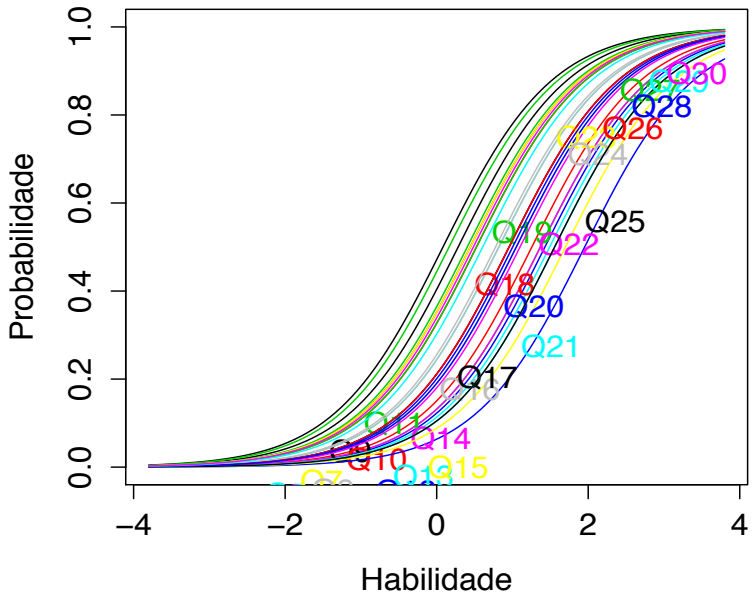

Figura 2: Curva característica do item para as trinta questões que compõe a prova multidisciplinar, modelo M2.

Na Figura 2, observa-se que a cauda da curva começa em zero para todas as questões conforme o M_2 (α,β_i,0). Também se verifica que a inclinação das curvas, que corresponde ao parâmetro de discriminância, é igual e foi estimado em 1,378. A curva representada pela Q12 que se aproxima do eixo da habilidade é o item mais difícil perante os outros, é o que requer maior habilidade para ser respondido corretamente. Por sua vez, a curva que está mais distante do eixo das habilidades é a Q_9, logo, é o item que no eixo das abscissas possui o menor valor, ou seja, é a mais fácil dentre as questões, por deter a menor habilidade.

A Figura 3 retrata as curvas características das trinta questões respondidas, pelos 94 respondentes, considerando 6 habilidades, ajustando o modelo conforme as disciplinas.

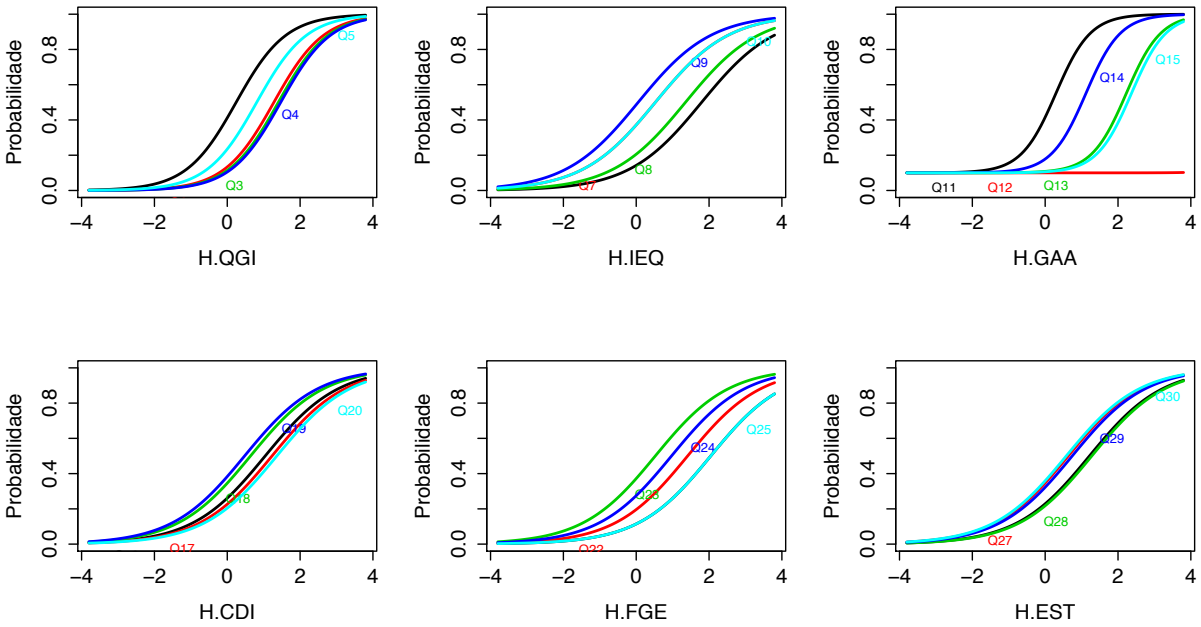

Figura 3: Curva característica do item para QGI -M_2 (α,β_i,0) , IEQ - M_1 (1,β_i,0), GAA - M_4 (α,β_i,c), CDI, FGE e EST - M_1 (1,β_i,0).

Analisando as CCI por disciplinas, verifica-se que para a maioria não existe uma alteração relevante em relação a sua forma, já que os modelos escolhidos foram praticamente os mesmos. Porém, a disciplina de GAA apresenta uma questão que mesmo um aluno com alta habilidade tem baixa probabilidade de acerto, e esta probabilidade é praticamente toda atribuída ao acerto casual.

A questão Q12 foi indicada como a mais difícil quando se avaliou a prova como um todo, apresentou baixa correlação PB, baixo AC e uma proporção de acerto de apenas 10%. Todas as medidas destacaram esta questão como fora do padrão das demais. Além disso, a disciplina de GAA foi sempre destacada como tendo itens não coerentes e com baixa proporção de acerto. Isso pode indicar que a prova desta disciplina pode precisar de uma reformulação como é visto pela CCI da Figura 3 .

De forma geral, os modelos permitiram indicar a dificuldade relativa de cada um dos itens, também permitiram inferir que a discriminância dos itens é praticamente a mesma, além de mostrar que o acerto casual é pouco significativo para a maioria dos itens. Para finalizar a análise, é interessante comparar os escores obtidos por alguns dos alunos, pelo TCT (número de acertos) e os escores obtidos com o modelo geral e para cada uma das habilidades específicas. A comparação não deve ser feita por valores, mas sim através da classificação dos alunos. O objetivo é verificar usando uma ou outra abordagem, por exemplo, para classificação dos alunos quanto à possibilidade de ganhar uma bolsa de estudos, os resultados seriam ou não drasticamente diferentes.

O procedimento consistiu em estimar a habilidade para todos os alunos por cada um dos modelos ajustados e, com base nestes, estabelecer um *ranking* dos alunos e compará-lo com a

classificação obtida pela correção tradicional da prova. A Tabela 5 apresenta esta análise.

Tabela 5: Comparação entre os escores obtidos pela TCT e TRI

A	A	TCT		TRI																					
l	c	Escore	P	H. Geral	P	6 Habilidade (94)																			
		(0 a 10)	o	(94) M_2	o	A	QGI	P	A	IEQ	P	A	GAA	P	A	CDI	P	A	FGE	P	A	EST	P
						c	M_2	o	c	M_1	o	c	M_4	o	c	M_1	o	c	M_1	o	c	M_1	o
1	12	4	7	0,54	7	1	0,00	5	2	0,20	4	1	-0,40	15	3	0,74	2	3	0,97	2	2	0,25	4
2	0	0	0	-1,56	0	0	-0,66	0	0	-0,85	0	0	-0,45	0	0	-0,78	0	0	-0,64	0	0	-0,79	0
9	15	5	4	0,85	4	5	1,85	1	2	0,20	4	1	0,36	9	2	0,28	3	2	0,49	3	3	0,71	3
12	16	5,3	3	0,94	3	2	0,51	4	5	1,63	1	2	-0,42	16	1	-0,21	4	4	1,44	1	2	0,25	4
13	15	5	4	0,85	4	3	0,95	3	4	1,15	2	1	0,36	9	4	1,19	1	1	-0,04	4	2	0,25	4
18	8	2,7	11	0,10	11	2	0,51	4	2	0,20	4	0	-0,45	0	3	0,74	2	0	-0,64	0	1	-0,24	5
26	18	6	1	1,14	1	2	0,51	4	5	1,63	1	2	1,00	4	4	1,19	1	2	0,49	3	3	0,71	3
27	17	5,7	2	1,04	2	2	0,51	4	2	0,20	4	2	1,00	4	4	1,19	1	2	0,49	3	5	1,63	1
33	18	6	1	1,14	1	5	1,85	1	4	1,15	2	3	1,48	1	4	1,19	1	0	-0,64	0	2	0,25	4
71	7	2,3	12	-0,02	12	2	0,51	4	1	-0,29	5	0	-0,45	0	1	-0,21	4	1	-0,04	4	2	0,25	4

Fonte: O autor (2012)
Nota: Alunos (Al);Acertos (Ac) por habilidades; Posição (Po) posicionamento de ranck pelo escore obtido dos 94 respondentes.

Para melhorar explorar os dados apresentados na Tabela 5, considere o aluno 1. Este teve 12 acertos, ficou com um escore de 4, uma vez que respondeu corretamente a 40% da prova, e foi o sétimo melhor aluno na classificação final. Quando avaliado pela TRI, seu escore ficou em 0,54, pontuação que o levou novamente ao sétimo melhor desempenho.

Para a construção da Tabela 5, observou-se que nas habilidades estimadas pelos modelos M1 e M2, onde somente o parâmetro de dificuldade foi estimado, a habilidade estimada levou a um *rankeamento* igual ao do TCT. Diferenças ocorreram somente quando as habilidades foram estimadas pelo modelo M4, que inclui o parâmetro de acerto ao acaso - comum a todas as questões - e a discriminância - estimada para cada questão. Somente este modelo, entre os três escolhidos, possibilitou obter escores diferentes entre a TRI e a TCT. Como exemplos, os alunos 26 e 33, que tiveram mudanças em suas posições.

4 Conclusões

Este resumo apresentou o emprego dos modelos de TRI a um exame multidisciplinar aplicado pela PUCPR a alunos de graduação. Os resultados mesmo com uma amostra pequena, mostraram que, em geral, a prova apresenta boa coerência interna, bem como boa aderência ao instrumento de medida proposto.

A principal contribuição da TRI é apresentar de forma sistemática a análise não apenas dos alunos, mas também da prova, dando uma visão crítica de sua construção e de sua capacidade em aferir o conhecimento dos alunos. Apresenta informações relevantes quanto à dificuldade de cada item, se os itens têm diferentes níveis de discriminância e se a probabilidade de acerto casual deve ou não ser levada em consideração.

Com os resultados, a instituição pode contar com mais um instrumento para a busca da melhoria da qualidade do ensino oferecido aos seus alunos. É também um aporte que pode auxiliar os professores a identificar as áreas das disciplinas que geram mais dificuldades e que

necessitam de mais atenção, para garantir o desenvolvimento de competências pelos alunos.

Deixa-se aqui, como proposta para futuras pesquisas, a avaliação deste exame considerando modelos da TRI que sejam capazes de contemplar que mais de uma habilidade latente está sendo avaliada, bem como que considerem que várias habilidades podem influenciar a resposta de cada item.

Referências

[1] ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C.**Teoria da Resposta ao Item: conceitos e aplicações**. São Paulo: ABE Associação Brasileira de Estatística, 2000.

[2] ANDRIOLA, W. B.**Uso da Teoria de Resposta Ao Item (TRI) para analisar a equidade do processo de avaliação do aprendizado discente.** Revista Iberoamericana de Avaliação Educacional, v. 1, p. 171-189, 2008

[3] BAKER, F. B.**The basics of Item Response Theory.**2. ed. USA: ERIC Clearinghouse on Assessment and Evaluation, 2001.

[4] BAKER, F. B.; KIM, S.**Item Response Theory: parameter estimation techniques.** 2. ed. revised and expanded. New York: Marcel Dekker, 2004.

[5] BIRNBAUM, A.**Some latent trait models and their use in inferring and examinee?s ability.** In: Loed FM, Lord MR. Novick, statistical theories of mental test scores. Reading: Addison Wesley; p.17-20, 1968.

[6] BOCK, R. D., & LIEBERMAN, M. **Fitting a response model to n dichotomously scored items.** Psychometrika, 35, 179?197, 1970.

[7] BRASIL. MINISTERIO DA EDUCAÇÃO. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS (INEP).**Mapa da educação superior.** Brasília, PF: MEC; INEP, 2004. 85p.

[8] BRASIL. MEC/INEP.**Relatório IDEB.** Disponível por: http://portalideb.inep.gov.br/. Acesso em: 22 mar. 2012.

[9] BOOMSMA A.; VAN DUIJN, M. A. J.; SNIJDERS, T. A. B.**Essays on item response theory**. Lecture Notes in Statistics (Springer-Verlag), n. 157. New York: Springer, 2000.

[10] BONAT, H. W. et al.**Métodos Computacionais em Inferência Estatística.** ABE-Associação Brasileira de Estatística, SINAPE, 2012.

[11] CHALMERS Phili.**R. mirt: A Multidimensional Item Response Theory Package for the R Environment.** Vol. 48, Issue 6, May 2012.

[12] QUARESMA, S.E.; DIAS, S. T. C.; SARTORIO,D.S.**Avaliação da aprendizagem e das provas do centro de formação interdisciplinar/UFOPA via Teoria da Resposta ao Item.** UFOPA. Disponível em: http://www.sbec.org.br/evt2012/trab16.pdf.Acesso em: 25 jul. 2013.

[13] FRANCISCO, R.**Aplicação da Teoria da Resposta ao Item (TRI) no Exame Nacional de Cursos (ENC) da Unicentro.** 2005. 144 f. Dissertação (Mestrado em Ciências) - Pós- Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2005.

[14] LORD F.M.**Applications of item response theory to practical testing problems.** Hillsdale: Erlbaum; 1980.

[15] NOJOSA, Ronald T.**Modelos Multidimensionais para a Teoria de Resposta ao Item.** Pernambuco, UFPE, Tese de Mestrado, 2001.

[16] PITON-GONÇALVES, J.**Desafios e Perspectivas da Implementação Computacional de Testes Adaptativos Multidimensionais para Avaliações Educacionais.** Tese de Doutorado, - Universidade de São Paulo, 2012

[17] R. DEVELOPMENT CORE TEAM.**R: A language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing, 2009. Disponível em: http://www.lsw.uni-heidelberg.de/users/christlieb/teaching/UKStaSS10/R-refman.pdf. Acesso em: 25 jul. 2013.

[18] RASCH G.**Probabilistic models for some intelligence and attainment tests.** Copenhagen: Danish Institute for Educational Research and St. Paul; 1960.

[19] RECKASE, M. D.**Multidimensional Item Response Theory: Statistical for social and behavioral sciences.** Springer Science Business Media: LLC, 2009.

[20] RIZOPOULOS, D.**ltm:An r packages for latent variable modelling and item response theory analyses.** R package. Disponível em: http://cran.R- project.org/package=ltm

[21] VAN DER LINDEN, W. J.; HAMBLETON, R. K.**Handbook of Modern Item Response Theory.** New York: SpringerVerlag, 1997.

[22] VENDRAMINI, C. M. M.; DIAS, A. S.**Teoria de Resposta ao Item na análise de uma prova de estatística em universitários. Psico-USF, v. 10, n. 2, p. 201-210, jul./dez. 2005.**