APLICAÇÃO DA TEORIA DE RESPOSTA AO ITEM EM UM EXAME MULTIDISCIPLINAR

EDY CÉLIA COELHO *

PAULO JUSTINIANO RIBEIRO JUNIOR **

WAGNER HUGO BONAT ***

PAULO SERGIO MACUCHEN NOGAS ****

* Doutoranda em Métodos Numéricos em Engenharia pela UFPR. Professora da Pontifícia Universidade Católica e da Secretária Estadual Educação do Paraná, Curitiba, Brasil. E-mail: edyceliacoelho@gmail.com. Rua: Cel. Pretextato P. F.T. Ribas, 300; Cep 80310260, Ctba/PR.
** PhD em Estatística pela Lancaster University e Professor do Departamento de Estatística da UFPR, Curitiba, Brasil. E-mail: paulojus@leg.ufpr.br
*** Mestre em Métodos Numéricos em Engenharia pela UFPR e Professor do Departamento de Estatística da UFPR, Curitiba, Brasil. E-mail: wbonat@gmail.com
**** Doutor, Professor e Coordenador de Avaliação Institucional e da Graduação da Pontifícia Universidade Católica do Paraná, Curitiba, Brasil. E-mail:paulo.nogas@pucpr.br

Resumo: A Teoria da Resposta ao Item (TRI) na avaliação educacional, busca fornecer desempenho do educando, visando suas habilidades e competências, como alternativa para avaliar alunos, considerando o desenvolvimento do conhecimento, nos itens e não no teste. Este estudo fornece uma visão geral dos modelos da TRI e as implicações da avaliação, composta por um conjunto de dados com 94 alunos da graduação, que realizaram exame multidisciplinar, uma avaliação interna da instituição, para verificar o desempenho e conhecimento. Os estudos realizados mostraram que cada modelo proposto para aquilatar a capacidade dos alunos tem sua particularidade. Nos resultados, foram indicados três modelos escolhidos dos nove abordados pela unidimensionalidade. Os escores obtidos possibilitaram comparações pelo método da teoria clássica e pela TRI. Através dessas análises o objetivo é mostrar a importância do modelo da TRI em avaliação, capaz de levar em consideração que uma prova seja composta e consequentemente respondida por múltiplas habilidades.

Palavras-chave: Avaliação. Habilidades. Modelos da TRI.

APPLICATION OF ITEM RESPONSE THEORY IN A MULTIDISCIPLINARY EXAMINATION

Abstract: The Item Response Theory (IRT) on the educational evaluation, search to provide the teaching performance, aiming on their skills and competences, alternative with used to evaluated students, considering the development of their knowledge, on the items and not on the tests. This research provides a wide view of the IRT models and their implications on the educational evaluations, and it is composed by a set of data collected from 94 graduates, that took a multidisciplinary exam, developed by the institution, to verify the development and knowledge. The studies showed that the purposed each model is to measure the student´s capacity have their

particularity. On the evaluated results three out of nine models were chosen and analysed by the one–dimensionality. The obtained scores allowed comparison with the classic theory method and the IRT. From these investigations the aim is to evaluation for IRT model, able to take in consideration that an exam may be composed and consequently answered by many latent abilities.

Key word: Evaluation. Abilities. IRT Models

INTRODUÇÃO

As avaliações aplicadas à educação, psicologia, qualidade de vida ou à gestão da qualidade em indústrias e empresas como um todo, vêm ocupando posição de destaque. A procura por melhoria da qualidade na execução dos processos de avaliação é uma realidade.

A avaliação é um importante instrumento de medição e ensinar é buscar ao longo do período credenciar o aluno a absolver cada vez mais conhecimentos. Marcar uma opção em uma prova pode ser um ato de plena consciência do assunto em questão, ou apenas mais um item assinalado para não perder alguma chance de possível acerto.

Para tentar chegar mais próximo de uma conclusão real acerca do verdadeiro conhecimento de um indivíduo avaliado foi desenvolvida, a partir da Teoria Clássica dos Testes (TCT), a Teoria da Resposta ao Item (TRI). Definida como modelos matemáticos construídos para representar a probabilidade de um indivíduo responder corretamente um item em determinado teste.

A TRI vêm sendo utilizada nacionalmente, principalmente por órgãos de avaliação institucional como o MEC no âmbito do INEP. À medida que o contato do meio educacional com esta teoria seja antecipado ocorrerá a mais rápida disseminação de seu uso, o que pode trazer contribuição à sociedade, por tratar-se de um elemento diferencial para melhor conhecer e retratar o conhecimento.

Para avaliar o aprendizado, a aquisição de habilidades e as competências de acordo com o perfil dos estudantes, a Instituição Pontifícia Universidade Católica do Paraná (PUCPR) promove todos os anos, em seus campus, o exame multidisciplinar. A avaliação parte da ideia de que as competências profissionais são construídas ao longo do curso e 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, essa que constitui uma das mais importantes vertentes para a qualidade do ensino.

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, candidatos concorrem a vagas de emprego na área pública, entre outros. Mas um ponto crucial que se deve levar em consideração na escolha ou na aprovação de um indivíduo nestes tipos de seleções é: Será que a competência que ele realmente adquiriu mede o 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 e outras questões são de relevância para o processo de avaliação educacional, que não podem ser resumidas somente em conceitos formais e de 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 e, 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 continua, permitindo desta forma uma melhorar compreensão dos mecanismos geradores de dificuldades no aprendizado, um acompanhamento do desempenho dos cursos ao longo do tempo, contribuindo desta forma para a melhoria do ensino superior como um todo dentro da instituição.

Além disso, objetiva-se comparações entre as abordagens da TRI e TCT no sentido de verificar, a concordância ou não entre os métodos, na atribuição de notas/escores em alunos avaliados. E também destacar os pontos onde a TRI tem mais a contribuir para o processo avaliativo do que a TCT.

Espera-se que esta análise, possibilite uma reflexão e que leve os docentes de disciplinas correlatas, a ter subsídios para direcionar o esforço empreendido no processo de ensino e aprendizagem, de forma a contemplar a melhor abordagem pedagógica e o mais pertinente método didático adequado à cada disciplina.

O presente trabalho encontra-se dividido em cinco seções, esta primeira busca introduzir o problema de análise motivando o uso dos métodos da TRI. A segunda apresenta a importância da TRI no contexto institucional. A terceira aborda os modelos da TRI utilizados nesta pesquisa. A quarta apresenta os principais resultados e a quinta e última apresenta uma breve discussão e recomendações para trabalhos futuros.

CONTEXTO INSTITUCIONAL-TRI

O objetivo da PUCPR é avaliar a aquisição de habilidades e competências, mas ainda não aplica o modelo TRI no exame multidisciplinar. Com a aplicação da nova proposta, os resultados deverão oferecer subsídios contribuindo na melhoria constante da qualidade do ensino. O exame coopera com o processo de planejamento e com a auto-avaliação dos envolvidos na academia. Garantia que pode ser obtida pela TRI e assegurada em novas propostas pedagógicas, detalhadas estatisticamente, visando a qualidade do ensino.

Ao buscar fundamentos na literatura para aplicar a TRI, observa-se que é muito recente quando o enfoque é avaliação educacional. De acordo com Andrade et al. (2000) a teoria vem sendo utilizada desde 1995, ganhando popularidade devido às provas do ENEM para a seleção de candidatos em algumas universidades do país em 2010. Cito alguns autores que abordaram a TRI na área de avaliação relatando as potencialidades da teoria na validade de testes como: (NOJOSA, 2001;VENDRAMINI ET AL, 2005; ANDRIOLA, 2008 e 2009; PASQUALI, 2007, 2009 e 2011; ANDRADE ET AL, 2010; QUARESMA ET AL, 2012).

A maioria das avaliações atuais levam em considerações 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.

Os modelos da TRI unidimensionais descrevem a relação entre as respostas observadas ao item e um traço latente, usualmente simbolizado por θ, que forma a base destas respostas. Eles 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 a compreensão de como o conhecimento por disciplina esta sendo avaliado e fornecer diagnósticos, subsídios para a implementação ou manutenção das metodologias educacionais.

MODELOS DA TRI

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 onde um teste é formado por i questões e 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 envolve: (i) a habilidade latente do individuo θj, (ii) a dificuldade βi, (iii) a discriminância αi e (iv) a probabilidade de acerto casual ci, para cada um dos itens. A equação do modelo logístico de três parâmetros é:

$$P_{ij} = P(Y_{ij} = 1|\theta_j) = c_i + (1-c_i)\ \frac{1}{1 + e^{-\alpha i(\theta j - \beta i)}}. \qquad (1)$$

Pode-se identificar facilmente dois casos particulares. O primeiro quando a probabilidade de acerto casual é desprezível, ou seja, ci = 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, denote 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:

$$Y_{ij} \mid \theta_j \sim B\ (n=1, P_{ij})$$

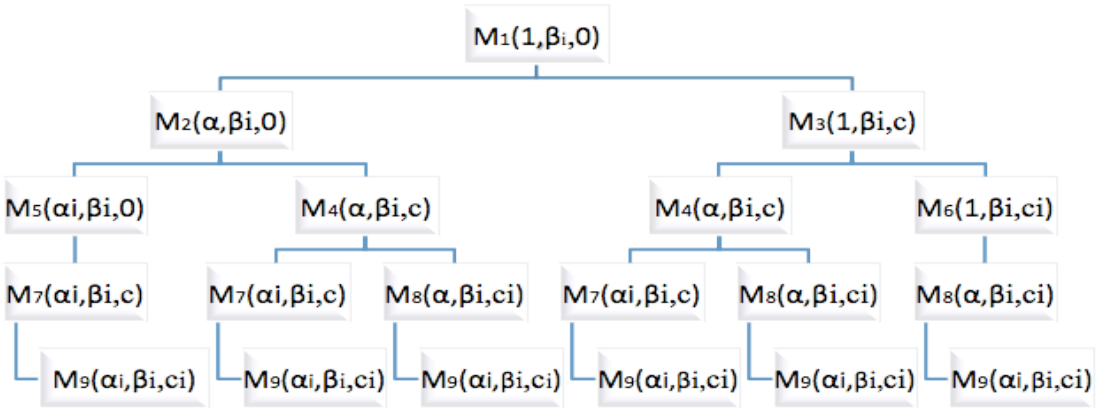
$$\theta_j \sim N\ (0, 1).$$

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 θj. O integrando desta verossimilhança marginal 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 artigo 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 TRI de Acordo com a Posição Parâmetros



FONTE: O autor (2012)

A Figura 1, mostra que partido 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. Esta comparação pode ser feita pelo teste de razão de verossimilhanças quando os modelos são encaixados, ou pelo Critério de Akaike (AIC), ou pelo Critério Baysiano (BIC) as três possibilidades foram contempladas na análise.

RESULTADOS

Foram sujeitos do presente estudo uma amostra aleatória do segundo semestre do ano de 2011 e escolhido casualmente um único curso da graduação com trezentos e quarenta e três (343) alunos que realizaram a avaliação multidisciplinar. Dentre esses alunos somente os do segundo período que realizaram a mesma prova, totalizando noventa e quatro (94) discentes da graduação.

O conjunto de dados foi processado pela instituição deixando em anonimato o nome do curso devido a ética profissional, até obter resultados considerados finais para a pesquisa. As notas foram processadas pelo método tradicional de sumarização de certo e errado pela TCT. Primeiramente analisou-se os resultados descritivamente do conjunto de dados da avaliação multidisciplinar, contemplando uma única habilidade, que é a geral e visa a formação do acadêmico.

O exame multidisciplinar é composto por seis blocos de disciplinas e sem dúvida, todas fazem parte de uma habilidade geral, que é a formação do profissional. A prova é composta

de seis habilidades, cujas as, disciplinas são : 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 á Química (IEQ). Averiguou-se os resultados para as devidas comparações observando e analisando as seis habilidades e a habilidade geral (HG). Cada disciplina abordou cinco (5) itens.

A principio os modelos TRI supõe que existe uma única habilidade latente sendo medida, porém a prova multidisciplinar é composta de seis pode-se chamar de sub habilidades que fazem parte de uma habilidade geral. Pensando nisso, a análise foi conduzida de duas formas, a primeira considerando que cada disciplina é uma habilidade latente e sua prova foi analisada separada das demais. A segunda considerou 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 mais adequada para a presente situação.

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 compõe o exame multidisciplinar. Para uma melhor compreensão numerou-se as questões dentro das habilidades de 1 a 5 e na prova como um todo de 1 a 30, essa codificação será utilizada em toda a análise.

Tabela 1 – Proporção de Acertos por Questão

HG	HQ	% acertos	HG	HG	% acerto
	QGI			CDI	
Q1	Q 1	44%	Q16	Q 1	30%
Q2	Q2	20%	Q17	Q 2	27%
Q3	Q 3	18%	Q18	Q 3	37%
Q4	Q4	17%	Q19	Q 4	40%
Q5	Q 5	30%	Q20	Q 5	24%
	IEQ			FGE	
Q6	Q 1	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 HG. Questão (Q).

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%). Essa disciplina é a que apresenta um maior distanciamento entre a proporção de acertos das questões. Nas demais disciplinas os acertos foram mais 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 especificas.

Tabela 2 – Corelação Ponto Bisserial e Alfa de Cronbach

Habilidades Questão (Q)	Ponto Bisserial		Cronbach's alpha *Todas as questões e Excluindo a questão	Habilidade Geral	Ponto Bisserial		Cronbach's alpha *Todas as questões e Excluindo a questão
	incluir	excluir			incluir	excluir	
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's 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.

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 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 a 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. 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 esta testando ou sem o item o que é de mais interesse, por isso a Tabela 2, apresenta as colunas (incluir) e (excluir).

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

Um questão interessante que surge quando aplica-se uma prova de múltipla escolha é o que aconteceria se todos os alunos estivessem respondendo ao acaso as questões. Para verificar qual seria o comportamento sob esta hipótese 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 responderam a prova esperava-se que ninguém zerasse, sob a suposição de acerto casual. Porém, observou-se 19 alunos que não tiveram nenhum acerto.

Tabela 3 – Frequências de Acertos Observado e o Número Esperado de Acertos sob a Hipótese de Respostas Aleatórias

Acertos	Observado	Esperado	Acertos	Observado	Esperado
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 (NAC) foi zero.

Esse é um resultado bastante interessante, pois ele pode indicar pelo menos duas situações que merecem atenção. A primeira é que um grupo de alunos optou por não fazer a prova e marcaram as questões que sabiam estar erradas. A segunda é que os alunos não objetivam responder a prova de forma aleatória, e sim tentam responder corretamente, porém devido a sua baixa habilidade erram. De uma forma ou de outra este é um resultado que indica que melhorias, sejam nas questões, sejam no ensino, ou mesmo na conscientização dos alunos para responderem a prova de forma responsável são necessárias.

Observa-se também que entre 0 e 9 acertos o valor observado foi sempre menor do que a esperança sob acerto casual. Observa-se que ninguém acertou da primeira a 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. Da mesma forma que a parte descritiva o ajustes dos modelos TRI será conduzido por disciplina e com a prova como um todo.

Desta forma, para cada uma das disciplinas e para a prova toda foram ajustados 9 modelos 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 conforme a suposição para os parâmetros de discriminância, dificuldade, e acerto ao acaso (α, β, c) ajustados foram: $M_1(1, \beta_i, 0)$; $M_2(\alpha, \beta_i, 0)$; $M_3(1, \beta_i, c)$; $M_4(\alpha, \beta_i, c)$; $M_5(\alpha_i, \beta_i, 0)$; $M_6(1, \beta_i, c_i)$; $M_7(\alpha_i, \beta_i, c)$; $M_8(\alpha, \beta_i, c_i)$; $M_9(\alpha_i, \beta_i, c_i)$.

Para comparação de forma geral busca-se o modelo com menor número de parâmetros (df)

por ser mais simples de interpretação e estimação. Como recomendação geral, quanto menor o AIC e BIC melhor é o ajuste. A Tabela 4, apresenta as medidas de comparação para todos os modelos ajustados para a habilidade geral e especificas.

Tabela 4 – Habilidades dos Alunos com os Diferentes Modelos da TRI

	MODELOS (α, β, c)								
	M_1 (1, βi, 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)
				GERAL					
	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					
	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	575,30	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**	507,40	506,04	507,40	511,03	514,03	511,06	514,24	520,78
BIC	**518,76**	522,66	518,76	522,66	536,47	539,47	536,50	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	**583,40**	585,21	583,40	585,34	588,64	591,07	588,64	592,64	598,77
BIC	**596,12**	600,47	596,12	600,60	614,07	616,51	614,07	620,62	636,92

FONTE: O autor (2012)

Para escolha do modelo, considere 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, isso indica que o procedimento computacional está coerente.

Observa-se na Tabela 4 que os modelos M_5 (α_i, β_i, 0); M_7(α_i, β_i, c) e 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 dos modelos possuírem o mesmo valor é porque o valor do acerto casual (c) foi estimado para ambos igual a zero.

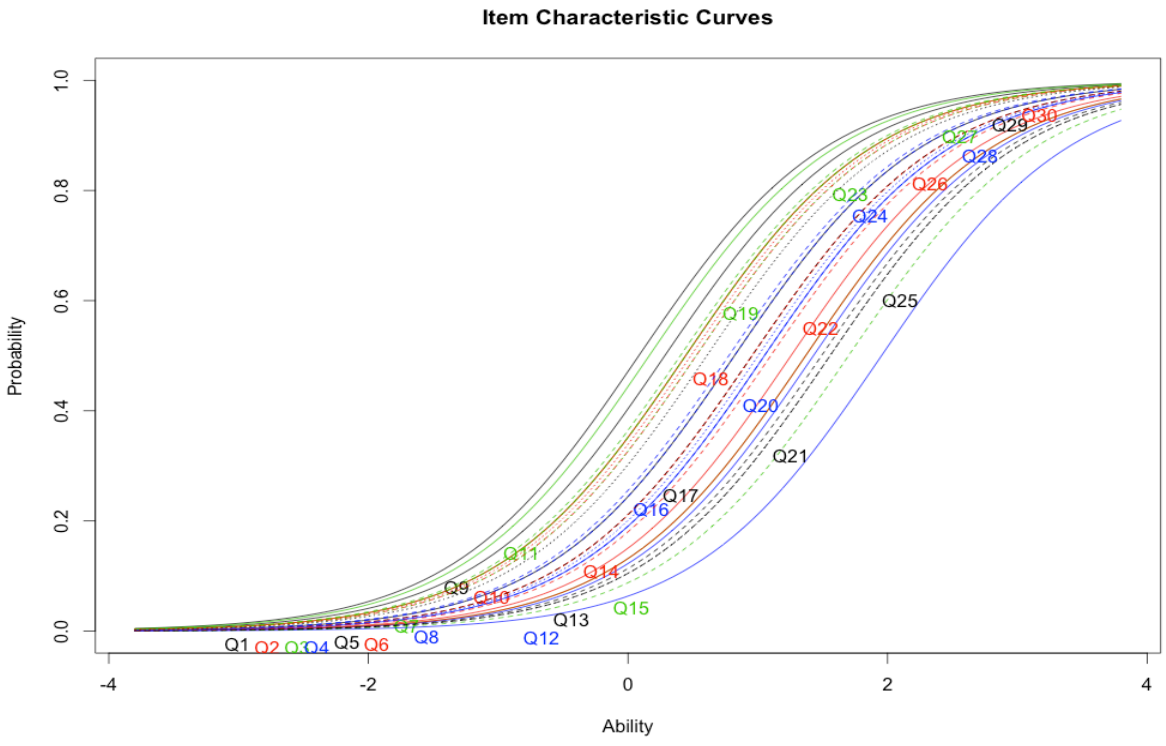
Considerando os critérios AIC e BIC, observa-se que o modelo com menor AIC e BIC é o M2 e o M4, porém o M4 apresenta a mesma logLik que o M2, pois o valor do parâmetro c foi estimado em zero, recaindo assim ao M2. Desta forma, opta-se por escolher o modelo M2, pois apresenta um ajuste não diferente do modelo mais complexo M9, porém com 59 parâmetros a mesmo para estimar.

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

O resultado do ajuste dos modelos podem ser vistos através das Curvas Características do Itens (CCI), tais curvas apresentam toda a informação relevante proveniente do modelo. A CCI relaciona através de um gráfico a habilidade dos indivíduos com a probabilidade destes 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 compõe o exame multidisciplinar. Considerando o modelo M2.

Figura 2 – Curva Característica do Item para as Trinta Questões que Compõe a Prova Multidisciplinar, Modelo M2.

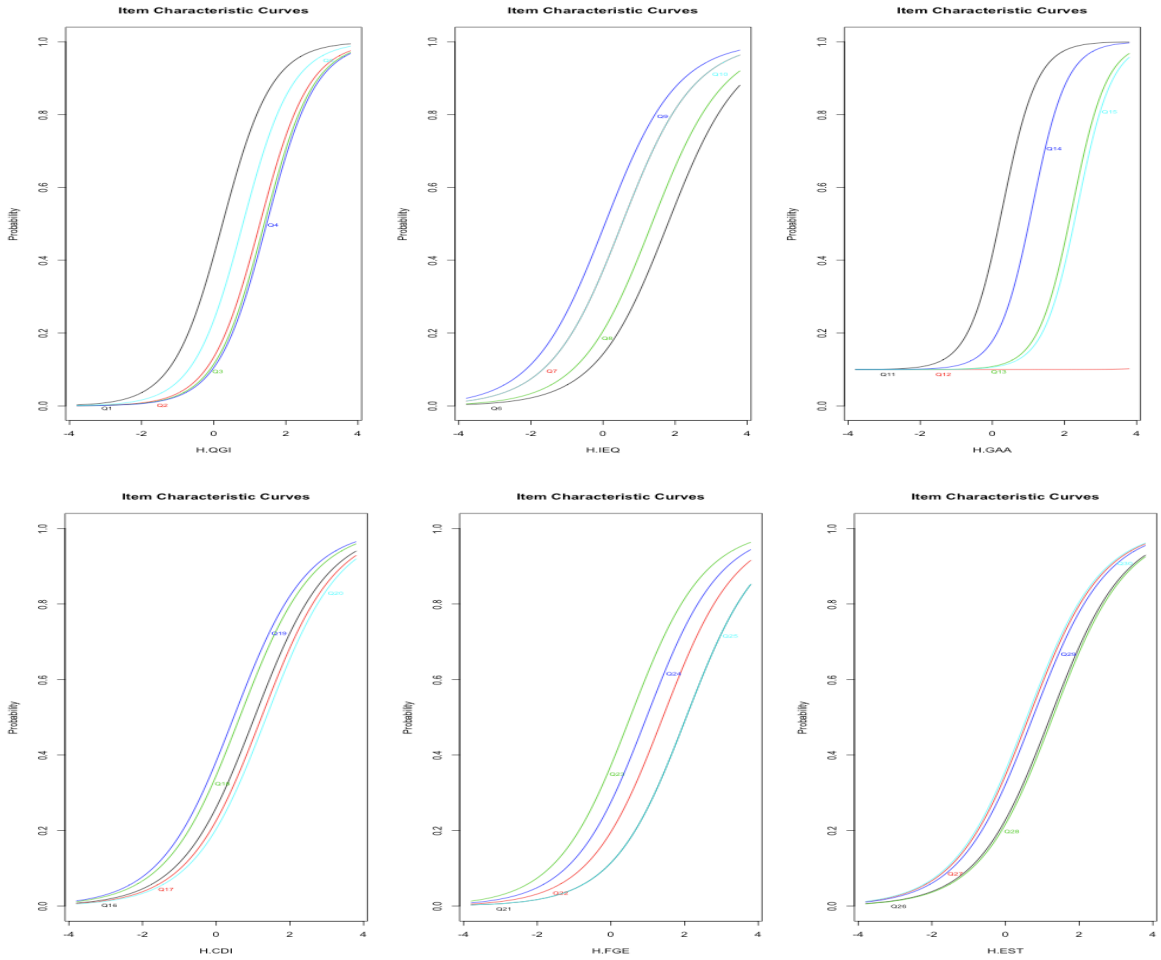


FONTE: O autor (2012)

Na Figura 2 observa-se que a calda da curva começa em zero para todas as questões conforme o $M_2(\alpha, \beta_i, 0)$. Também verifica-se 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 aos outros, é o que requer maior habilidade para ser respondido corretamente. Por sua vez a curva que esta mais distante do eixo das habilidades é a Q9, 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 as seis habilidades, ajustando o modelo conforme as disciplinas.

Figura 3 – Curva Característica do Item para QGI - $M_2(\alpha, \beta i, 0)$, IEQ - $M_1(1, \beta_i, 0)$, GAA – $M_4(\alpha, \beta_i, c)$, CDI, FGE e EST - $M_1(1, \beta_i, 0)$.



FONTE: O autor (2012)

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 uma probabilidade muito baixa de acerto e esta probabilidade é praticamente toda atribuída ao acerto casual.

Esta questão é a Q12 foi indicada como a mais difícil quando avaliou-se 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.

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 especificas. A comparação não deve ser feita por valores, mas sim pelo ranqueamento dos alunos. O objetivo é verificar se usando uma ou outra abordagem, por exemplo, para classificação dos alunos para o recebimento de uma bolsa de estudos, os resultados seriam drasticamente diferentes.

O procedimento consistiu em estimar a habilidade para todos os alunos por cada um dos modelos ajustados, com estes ranquear os alunos e comparar com o ranking obtido pela correção tradicional da prova. A Tabela 5 apresenta esta análise.

Tabela 5 – Comparação entre os Escores Obtidos pela TCT e TRI

Alunos	Ac	TCT		TRI																			
		Escore (0a10)	Po	H.G (94) M_2	Po	6 Habilidade (94)																	
						Ac	QGI M_2	Po	Ac	IEQ M_1	Po	Ac	GA A M_4	Po	Ac	CDI M_1	Po	Ac	FG E M_1	Po	Ac	EST M_1	Po
1	12	4	7	0,54	7	1	0,00	5	2	0,20	4	1	-0,40	3	3	0,74	2	3	0,97	2	2	0,23	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
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	-	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
31	10	3,3	9	0,33	9	0	1,85	0	1	-0,29	5	0	-	0	4	1,19	1	2	0,49	3	3	0,71	3
32	8	2,7	11	0,10	11	2	0,51	4	1	-0,29	5	0	-	0	3	0,74	2	1	-0,04	4	1	-	5
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	-	4	1	-0,04	4	2	0,25	4

FONTE: O autor (2012)

NOTA: Acertos (Ac) por habilidades; Posição (Po) posicionamento de ranque 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 já que acerto 40% da prova, e foi o sétimo melhor aluno. Quando avaliado pela TRI seu escore ficou em 0.54 que o levou novamente ao sétimo melhor desempenho.

Para a construção da Tabela 5, observou-se que nas habilidades estimadas pelos modelos M_1 e M_2 onde somente o parâmetro de dificuldade foi estimado, a habilidade estimada levou a um ranqueamento igual a TCT. Diferenças ocorreram somente quando as habilidades são estimadas pelo modelo M_4 que inclui o parâmetro de acerto ao acaso que é comum a todas as questões e a discriminância que foi estimado 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 ou 38 que tiveram mudanças em suas posições.

DISCUSSÃO

Este artigo apresentou a aplicação dos modelos de TRI a uma prova multidisciplinar, aplicada pela PUCPR a alunos de graduação. Os resultados mostraram que em geral a prova apresenta boa coerência interna com a maioria dos itens tendo 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 critica de sua construção e de sua capacidade em aferir o conhecimento dos alunos. Apresenta informações relevantes quanto a 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.

Em mãos destes resultados a instituição conta com mais um instrumento para a melhoria da qualidade do ensino oferecido aos seus alunos. Sendo também um instrumento que pode auxiliar os professores a identificar as áreas das disciplinas que geram mais dificuldades e que necessitam de mais atenção, para garantir a completa assimilação do conteúdo pelos alunos.

Deixa-se aqui como futuras agendas de 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 considere que várias habilidades podem influenciar a resposta de cada item.

REFERÊNCIAS

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.

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.

ANDRIOLA, W. B. **Psicometria Moderna: características e tendências**. Estudos em Avaliação Educacional, São Paulo, v. 20, n. 43, maio/ago. 2009.

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

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. Read- ing: Addison Wesley; p.17-20, 1968.

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

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

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. Centro de Formação Interdisciplinar, 2012. Disponível em: http://www.sbec.org.br/evt2012/trab16.pdf. Acessado em fevereiro 2012.

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

PASQUALI L. TRI - **Teoria de Resposta ao Item: teoria, procedimentos e aplicações.** Brasília: LabPAM/UnB; 2007.

PASQUALI L. **Psicometria.** Rev. esc. enferm. USP vol.43 no.spe São Paulo Dec. 2009.

PASQUALI, L. **Psicometria. Teoria dos testes na psicologia e na educação**. (4ª ed.). Petrópolis: Editora Vozes, 2011.

R. Development Core Team. **R: A language and environment for statistical computing. R** Foundation for Statistical Computing, Vienna, Austria. ISBN 3- 900051-07-0, URL http://www.R-project.org, 2012.

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

RECKASE, M. D. **Multidimensional Item Response Theory.** Statistical for Social and Behavioral Sciences. Springer Science Business Media: LLC, 2009.

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>

RIZOPOULOS, D. **Title Latent Trait Models under IRT**. Version 0.9-7 Date 2011-10-18.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. **Handbook of Modern Item Response Theory**. New York: Spring- Verlag, 1997.

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.