Uso da TRI para análise de um simulado

Use of TRI for analysis of a simulate

DOI:10.34117/bjdv7n1-353

Recebimento dos originais: 10/12/2020
Aceitação para publicação: 13/01/2021

Alan Kardec Messias da Silva
Mestre
Universidade do Estado de Mato Grosso - UNEMAT
Campus de Nova Xavantina - MT
Endereço: Rua Prof. Renato Figueiró Varella, Parque Municipal Mário Viana, CEP: 78690-000, Nova Xavantina – MT
E-mail: allankardec@unemat.br

Acelmo de Jesus Brito
Mestre
Universidade do Estado de Mato Grosso - UNEMAT
Campus de Barra do Bugres - MT
Endereço: Rua A, s/n, Bairro São Raimundo, CEP: 78390-000, Barra do Bugres – MT
E-mail: acelmo@unemat.br

Daniel Messias da Silva
Mestre
Instituto Federal de Mato Grosso – IFMT
Campus de Lucas do Rio Verde
Endereço: Avenida Universitária 1600 – W, Bairro: Parque das Emas, CEP: 78455-000, Lucas do Rio Verde – MT
E-mail: daniel.silva@lrv.ifmt.edu.br

Luciana Bertholdi Machado
Mestra
Universidade do Estado de Mato Grosso - UNEMAT
Campus de Barra do Bugres - MT
Endereço: Rua A, s/n, Bairro São Raimundo, CEP: 78390-000, Barra do Bugres - MT
E-mail: lucianabm@unemat.br

William Vieira Gonçalves
Doutor
Universidade do Estado de Mato Grosso - UNEMAT
Campus de Barra do Bugres – MT
Endereço: Rua A, s/n, Bairro São Raimundo, CEP: 78390-000, Barra do Bugres – MT
E-mail: williamvieira@unemat.br

RESUMO
O presente artigo exibe a análise de um simulado da Prova Brasil aplicado nas turmas de 5º ano como uma das ações do projeto Observatório da Educação com Iniciação à Ciência (OBEDUC), vinculado ao Campus da Universidade do Estado de Mato Grosso

(UNEMAT), localizado em Barra do Bugres – MT. O projeto OBEDUC desenvolvia ações nas escolas parceiras que visavam *promover o avanço na qualidade do ensino e reflexos nas melhorias da nota do IDEB dessas escolas*. Os simulados eram realizados nos 5º e 9º anos com o objetivo de *fazer um diagnóstico dos conhecimentos matemáticos desses discentes e orientar outras ações como a confecção de materiais auxiliares e demais abordagens pedagógicas*. A questão norteadora do grupo de trabalho foi *como fazer uma devolutiva para as escolas parceiras que não se pautasse apenas na estratificação de quantidades de acertos e erros ao simulado, e* desta "curiosidade" surgiu os estudos que possibilitaram fazer a transição da Teoria Clássica do Teste (TCT) para a Teoria de Resposta ao Item (TRI). Nosso objetivo foi *Escolher pela Teoria de Resposta ao Item um modelo matemático logístico que melhor se ajuste aos dados empíricos de nosso simulado* e para isso, fizemos uso do pacote estatístico *latente train model* (*ltm)* do software *R Statistc*, que aborda a metodologia psicométrica empregada na Teoria de Resposta ao Item para geração de seus modelos. Como resultado de nosso trabalho foram evidenciadas questões que contrariam o modelo logístico acumulativo pressuposto no simulado, tendo nestas questões pouca ou nenhuma consistência interna em seus padrões de respostas ao compararmos o teste como um todo, indicando haver questões que deveriam ser retiradas das análises ao passo que avançamos o processo de calibração dos dados pelos modelos de 1,2 e 3 parâmetros. As questões oriundas de melhores contribuições para a confecção de um modelo logístico mais confiável tiveram seus ajustes realizados por funções estatísticas do próprio pacote *ltm* e seus resultados indicaram o Modelo Logístico de 3 parâmetros (ML_3) sendo o mais ajustado aos dados empíricos obtidos em nosso simulado.

Palavras-chave: Teoria de Resposta ao Item, Modelos Logísticos, Avaliação em Larga Escala.

ABSTRACT

This article shows the analysis of a simulation of Prova Brasil applied to 5th grade classes as one of the actions of the Observatory of Education with Initiation to Science (OBEDUC) project, linked to the Campus of the State University of Mato Grosso (UNEMAT), located in Barra do Bugres - MT. The OBEDUC project developed actions in the partner schools that aimed to promote the advancement in the quality of teaching and reflexes in the improvement of the IDEB score of these schools. The simulations were carried out in the 5th and 9th years in order to make a diagnosis of the mathematical knowledge of these students and guide other actions such as the preparation of auxiliary materials and other pedagogical approaches. The guiding question of the working group was how to make a return to the partner schools that was not based only on stratifying the number of correct and wrong answers to the simulated, and from this "curiosity" came the studies that made it possible to make the transition from the Classic Test Theory (TCT) for Item Response Theory (TRI). Our objective was to choose, by Item Response Theory, a logistic mathematical model that best fits the empirical data of our simulation and for that, we used the latent statistical train model (ltm) package of the R Statistc software, which addresses the psychometric methodology employed in Item Response Theory to generate your models. As a result of our work, issues were found that contradict the cumulative logistic model assumed in the simulation, with these questions having little or no internal consistency in their response patterns when comparing the test as a whole, indicating that there are questions that should be removed from the analyzes step by step. that we have advanced the data calibration process using models of 1,2 and 3

parameters. The questions arising from better contributions to the making of a more reliable logistic model had their adjustments performed by statistical functions of the ltm package itself and their results indicated the 3-parameter Logistic Model (ML_3) being the most adjusted to the empirical data obtained in our simulation.

Keywords: Item Response Theory, Logistic Models, Large Scale Assessment.

1 O SIMULADO DO PROJETO OBEDUC - UNEMAT

Iniciamos este trabalho falando um pouco sobre os simulados realizados pelos professores pesquisadores em Avaliação em Larga Escala do projeto OBEDUC, vinculado a Universidade do Estado de Mato Grosso (UNEMAT) localizado no Campus de Barra do Bugres, que tinha por objetivo *fazer um diagnóstico dos conhecimentos matemáticos desses discentes e orientar outras ações como a confecção de materiais auxiliares e demais abordagens pedagógicas.*

Os simulados da Prova Brasil era uma ação desenvolvidas durante o ano letivo e aplicados aos discentes dos 5º e 9º Anos das escolas parceiras do Projeto, que atendia 6 (seis) municípios no interior do Estado de Mato Grosso, além de Barra do Bugres tínhamos também as cidades de Assari, Arenápolis, Nova Olímpia, Tangará da Serra e Nortelândia.

Essa avaliação não tinha como objetivo ranquear ou premiar os discentes ou escolas envolvidas e por isso em um primeiro momento, seus resultados estatísticos não tiveram uma abordagem metodológica mais sistemática pautada na análise de teste, eram somente verificações das quantidades de erros e acertos das questões, não levando em consideração algo muito importante, a qualidade das questões, algo que define a principal diferença entre a TRI e a TCT.

As análises consistiam em quantificar o número de acertos e erros dos respondentes e estratifica-los por turmas, por exemplo:

- Acertos até 30% das questões;
- acertos de 30% até 50% das questões;
- acertos acima de 50% das questões.

Tal movimento era para evidenciar alunos com diferentes graus de dificuldades nas mais diversas questões abordadas pelo simulado. No entanto, precisávamos melhorar nossos instrumentos de avaliação e encontrar mecanismos que pudesse direcionar uma melhor tomada de decisão, tanto na elaboração de materiais complementares, quanto na forma de

serem abordados em sala de aula, originando assim a escolha pela Teoria de Resposta ao Item.

Precisávamos agora melhor entender nosso novo instrumento de avaliação e buscar atingir nosso objetivo que é *escolher pela Teoria de Resposta ao Item um modelo matemático logístico que melhor se ajuste aos dados empíricos de nosso simulado*, e que este modelo mais favoreça a qualidade e a fidedignidade na descrição das habilidades dos alunos e qualidade dos itens.

Para tal finalidade utilizamos o *software R Statistic* e optamos pela metodologia empregada no pacote *latente train model - (ltm)* desenvolvido por Rizopoulos (2006). É importante destacar que existem hoje mais de uma dezena de pacotes no *R Statistic* capazes de auxiliarem nas análises abordadas na TRI.

Encontraremos nas seções seguintes uma descrição detalhada sobre os modelos matemáticos, gráficos e curvas que envolvem a fundamentação teórica da TRI, além de uma breve discussão sobre alguns procedimentos metodológicos e analíticos realizados pelo pacote *ltm*, onde exibimos os dados estatísticos que fazem as prévias antes de executarmos a acurácia dos modelos propostos, além de explicitarmos para qual modelo melhor se ajustará aos dados empíricos de nosso simulado.

2 SOBRE A TEORIA DE RESPOSTA AO ITEM

A TRI conhecida também como Teoria do Traço Latente teve seu esboço por Lord (1952) que viu nesta mesma década uma rápida expansão com os trabalhos de Rasch (1960), mas somente sendo finalmente formalizada anos depois com os trabalhos de Birnbaum (1968) e o mesmo Lord (1980).

A idealização desta teoria teve como sustentação problemas encontrada na TCT, um deles que podemos citar foi observado ainda na década de 30 por Thurstone (1928), que dizia.

“Um instrumento de medida, na sua função de medir, não pode ser seriamente afetado pelo objeto de medida. Na extensão em que sua função de medir for assim afetada, a validade do instrumento é prejudicada ou limitada. Se um metro mede diferentemente pelo fato de estar medindo um tapete, uma pintura ou um pedaço de papel, então nesta mesma extensão a confiança neste metro como instrumento de medida é prejudicada. Dentro dos limites de objetos para os quais o instrumento de medida foi produzido, sua função deve ser independente da medida do objeto” (THUSTONE, 1928, pg. 547).

Thustone se referia aos testes aplicados a sua época que dependiam dos itens que os componham, ou seja, a população testada teria diferentes resultados ao aplicarmos diferentes testes ou até mesmo, alterando alguns de seus itens, gerando assim desconfiança de qual ou quais instrumentos melhor realmente entregada resultados fidedignos a população testada.

Os aprimoramentos realizados na TCT durantes as décadas de 50 a 80 culminaram nos modelos matemáticos conhecidos hoje como TRI, no entanto, é importante ressaltar que segundo especialistas a TRI não veio substituir a TCT, mas sim aprimora-la, em particular, no que diz respeitos aos itens.

Apesar das técnicas de análises desenvolvidas pela TRI já estarem bem formalizadas por Rasch na década de 60, foi somente nos anos 80 que efetivamente foi possível utiliza-la de forma ampla, isto porque nela detêm algoritmos matemáticos de tal complexidade que a tecnologia da época era incapaz de resolver de maneira útil e prático, como por exemplo os estimadores de máxima verossimilhança envolvidos na calibragem dos parâmetros de dificuldades, discriminação e acerto ao acaso.

Para Andrade, Tavares e Valle (2000) a TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa reposta a um item em função dos parâmetro do item e do traço latente (habilidades) do respondente. Para concebe-la são necessários nos teste dois pressupostos existem nos itens: *a unidimensionalidade e a independência local*. Pressupostos esses que garante a homogeneidade do conjunto de itens em estar medindo um único traço latente e que para uma dada habilidade, as respostas aos diferentes itens da prova sejam independentes. No entanto, alguns estudos mostram que na verdade só existe a necessidade de uma única suposição, uma vez que a unidimensionalidade implica independência local e vice-versa. Mais detalhes vejam em Pasquali e Primi (2003).

Nossos estudos foram realizados com uma família de modelos da TRI conhecida como modelos unidimensionais dicotômicos acumulativo, isto é, modelos que analisam itens corrigido como “certo” ou “errado”, com somente um único traço latente (habilidades), onde a probabilidade de acerto aumenta proporcionalmente de forma não linear ao nível do traço latente.

Para Andrade, Tavares e Valle (2000) e Couto e Primi (2011) um teste com k itens dicotômicos e n respondentes medindo um único traço latente, pode ser estudado por um modelo logístico de 3 parâmetro (ML_3) definido por:

$$P_i(U_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i) \cdot \frac{1}{1 + e^{-D \cdot a_i \cdot (\theta_j - b_i)}}$$

com $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, n$ onde:

U_{ij} é a variável dicotômica que assume os valores 1, quando o indivíduo j responde corretamente o item i, ou 0 caso contrário;

θ_j representa o traço latente (habilidades) do $j-$ésimo indivíduo;

b_i é o parâmetro de dificuldade do item i, medido na mesma escala do traço latente θ;

a_i é o parâmetro de discriminação do item i, com valor proporcional à inclinação da curva característica do item (CCI) no ponto b_i;

c_i é o parâmetro do item que representa a probabilidade de acerto de um indivíduo, com baixa ou nenhuma habilidade responder corretamente o item i (acerto ao casual), medido na mesma escala da probabilidade;

D é fator de escala igual a 1.702 utilizado para obter uma aproximação da função ogiva normal e facilitar a compreensão aos parâmetros dos itens.

Mas além do ML_3 existem outros dois importantes modelos logísticos que podem ser aplicados em "concorrência" a este modelo. Exibimos primeiramente o modelo ML_2 que possui dois parâmetros calibrados, são eles, b_i e a_i, ou seja, neste modelo não contém o parâmetro de acerto ao acaso e é definido como,

$$P_i(U_{ij} = 1 \mid \theta_j) = \frac{1}{1 + e^{-D \cdot a_i \cdot (\theta_j - b_i)}}.$$

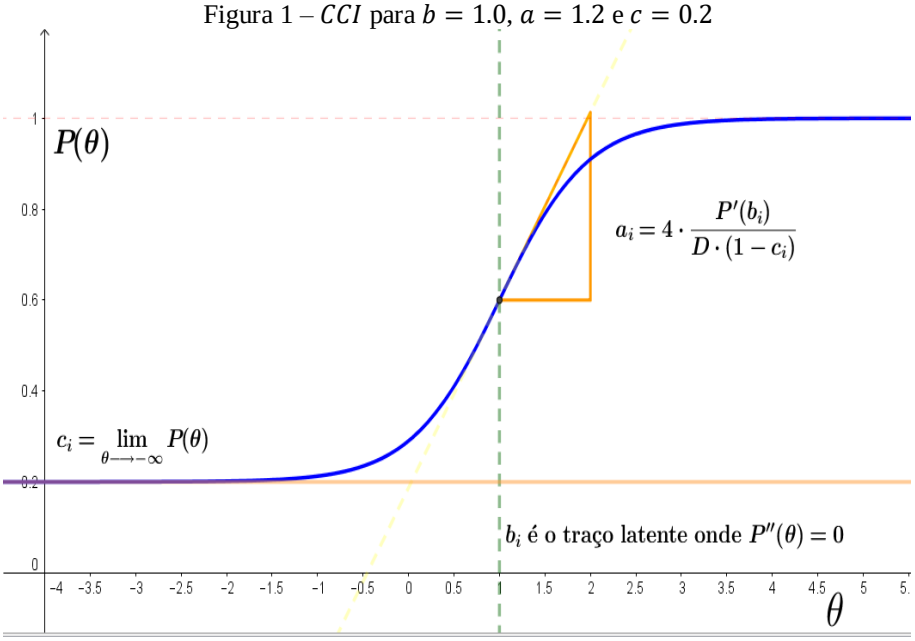
Já o outro modelo é o ML_1 que possui somente a calibração do parâmetro b_i e é descrito como,

$$P_i(U_{ij} = 1 \mid \theta_j) = \frac{1}{1 + e^{D \cdot (\theta_j - b_i)}}.$$

Ambos os modelo possuem um representação gráfica de seus itens realizada pela CCI, que é sem dúvidas a maior contribuição na inovação de análise de testes que a TRI trouxe em relação a TCT, seu objetivo é descrever graficamente a probabilidade $P_i(U_{ij} = 1 \mid \theta_j)$ de acerto de um respondente ao longo da escala do traço latente θ, além

disso, pela CCI se observa a propriedade acumulativa do item, pois indivíduos com maiores traço latente possuem maiores probabilidade de acertos, isto é, numa escala de proficiência os alunos com certos níveis de traço latente θ detêm todas as habilidades descritas para seus níveis de traço latente inferiores. No entanto, essa relação não é linear e pode ser observada no "formato de S" da CCI como exemplificada na figura abaixo.

Figura 1 – CCI para $b = 1.0$, $a = 1.2$ e $c = 0.2$



É importante ressaltar que todos os parâmetros de um item podem ser encontrados em sua CCI. Começamos pelo parâmetro de acerto ao acaso, que é definido sendo o seguinte limite $c_i = \lim_{\theta \to -\infty} P_i(U_{ij} = 1 \mid \theta_j)$ que na Figura 1 é $c_i = 0.2$, ou equivalente a dizer que a probabilidade de um respondente acertar o item com baixo ou nenhum traço latente será 20% (vinte por cento), indicando assim, um item com altas chances de acertos ao acaso (chute) no teste.

Já o parâmetro de dificuldade é encontrado sendo o ponto em θ de inflexão da função de probabilidade $P_i(U_{ij} = 1 \mid \theta_j)$, que matematicamente pode ser calculada igualando a segunda derivada a zero, ou seja, $\frac{d^2 P_i(U_{ij}=1 \mid \theta_j)}{d\theta^2} = 0$, que na Figura 1 é $b_i = 1$.

Por fim, temos o parâmetro de discriminação do item que pode ser encontrado como como a inclinação da função de probabilidade $P_i(U_{ij} = 1 \mid \theta_j)$ para quando $\theta = b_i$. Neste caso o cálculo é realizado em duas etapas. A primeira é o cálculo da derivada $\frac{dP_i(U_{ij}=1 \mid \theta_j)}{d\theta}$,

em seguida, a segunda etapa é isolar o parâmetro a_i que resultará na expressão $a_i = \frac{4 \cdot \frac{dP_i(U_{ij}=1 \mid \theta_j)}{d\theta}}{D(1-c_i)}$, este parâmetro na Figura 1 é dado por $a_i = 1.2$.

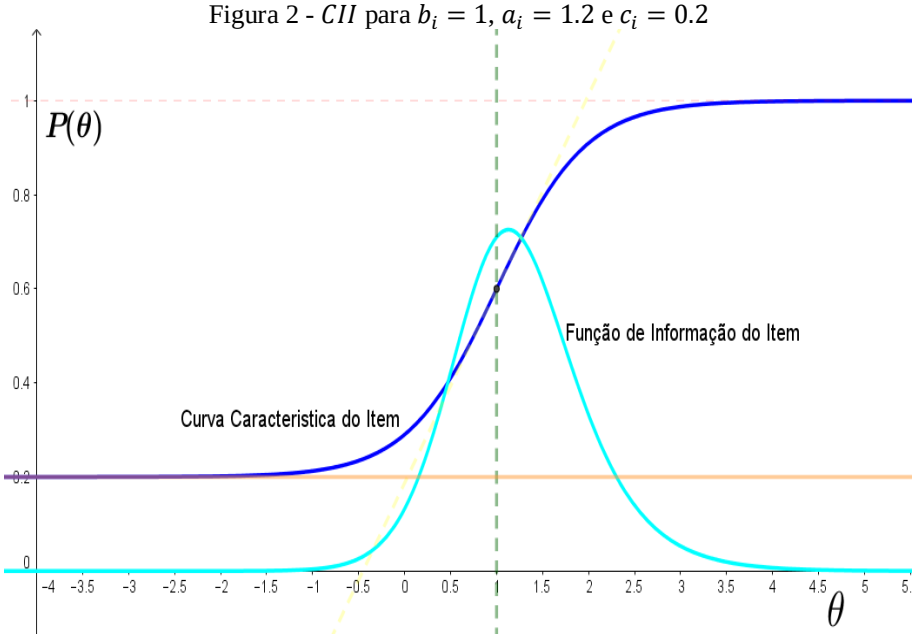
Outra novidade importante trazida pela TRI em relação a TCT é a curva que descreve a quantidade de informação ou de eficiência que um item possui sobre um determinado nível de traço latente. Ela é chamada por curva de informação do item (CII) e descreve graficamente o comportamento da função definida matematicamente abaixo,

$$FII_i(\theta) = \frac{\left[\frac{d\left(P_i(U_{ij}=1 \mid \theta_j)\right)}{d\theta}\right]^2}{P_i(U_{ij}=1 \mid \theta_j) \cdot \left(1 - P_i(U_{ij}=1 \mid \theta_j)\right)}$$

Esta equação é chamada de função de informação do item (FII) e é um escore com escala própria e obtido somente em função do traço latente, assumindo maiores valores no intervalo onde o item possui maior eficiência em descrever os dados empíricos, ou seja, próximo ao nível de dificuldade do item. Além disso, a FII também sofre impactos dos outros dois parâmetros, a discriminação e o acerto ao acaso, sendo cada vez maior em itens com alta discriminação e baixo acerto ao acaso.

Para uma melhor compreensão do comportamento da CII ao longo da escala θ, veja na Figura 2 uma sobreposição na CCI exibida na Figura 1.

Figura 2 - CII para $b_i = 1$, $a_i = 1.2$ e $c_i = 0.2$



A Figura 2 mostrar um item mais eficiente em descrever as habilidades investigadas em determinados níveis de traço latente, como por exemplo em $0.5 \leq \theta \leq 2.0$. Em contra partida o item é superestimado ao tentar descrever as habilidades de respondentes fora deste intervalo, podendo exibir informações enviesadas ou contrárias aos dados empíricos estudados.

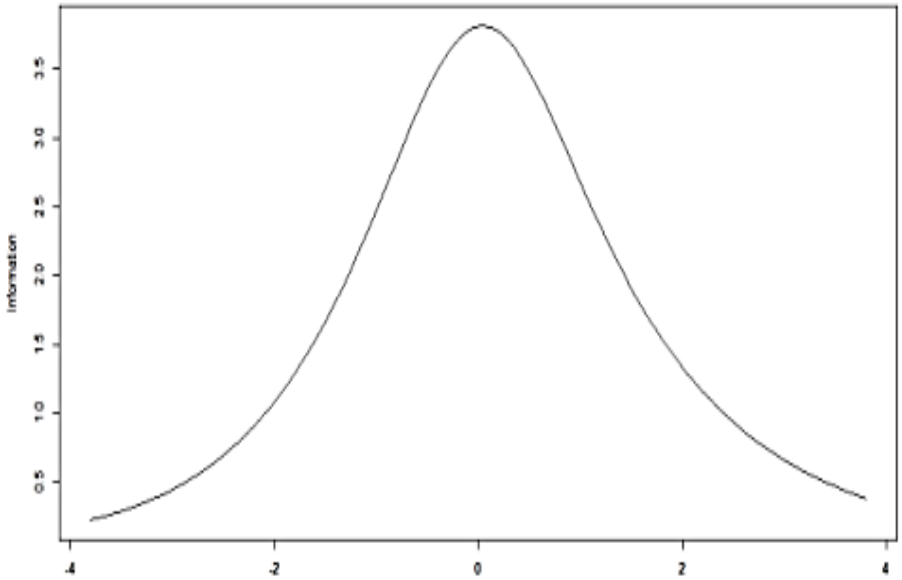
As CCI, CII e FII juntas são capazes de descrever com uma boa precisão os comportamentos e informações dos itens na composição de um teste, e servirão de apoio na escolha do melhor modelo que atenda aos dados empíricos de nosso simulado, em especial o escore dado pelo somatório das $FII's$ chamada por função de informação do teste (FIT), que para um teste com n itens sobre um escala de traço latente θ é dada pela seguinte equação,

$$FIT(\theta) = \sum_{i=1}^{n} FII_i\,(\theta)$$

O gráfico que representa a FIT é chamada por curva de informação do teste (CIT) e representa a quantidade de informação total contida no modelo logístico estudado ao longo do traço latente θ, que geralmente aumenta de acordo com a quantidade de parâmetros analisados e possui um comportamento de "sino" ao longo de θ conforme podemos ver na

Figura 3.

Figura 3 - Curva de Informação do Teste



O formato de "sino" exibido pela CIT é capaz de revelar em quais níveis de traço latente o teste calibrado mostra ter maior acuraria ao explicar as habilidades investigadas, que na Figura 3 revela ser maiores próximo ao valor de $\theta = 0$. É importante dizer que o escore FIT não deve ser comparado entre modelos de diferentes base de dados, ou seja, só podem ser comparados informações totais quando os modelos calibração ocorrem sobre os mesmos respondentes.

3 MÉTODO, RESULTADOS E DISCUSSÕES

Para este trabalho foram aplicados a 286 discentes dos 5º anos das escolas parceiras ao Projeto OBEDUC, um simulado avaliativo composto por 21 questões (itens) elaborados com base no banco de questões da Prova Brasil entre 2005 a 2013.

O simulado era composto por somente questões objetivas em matemática, dicotômicas e com quatro alternativas possíveis de respostas. As aplicações do simulado ocorreram no primeiro semestre de 2014 e de maneira assíncrona entre as escolas parceiras, isto devido ao pouco número de aplicadores disponíveis e a dificuldade em sincronizar todos os horários de nossos professores colaboradores.

A correção ocorreu via gabarito e os desempenhos dos 286 discentes no simulado foram tabulados em planilha simples do EXCEL na forma dicotômica, sendo 0 para erro e 1 para acerto. Em seguida foram realizadas algumas análises iniciais com a função $descript()$ do pacote ltm no *software R Statistic*, que exibiu alguns parâmetros estatísticos sobre a confiabilidade pelo $alpha\ de\ Cronbach\ (\alpha_c)$ e correlação existentes entre as questões pela $correlação\ ponto - bisserial\ (\rho_{pb})$.

Na tabela 1 mostramos para cada uma das questões do simulado as possíveis divergências nos padrões de respostas encontradas, indicando assim, questões que devem ser retiradas para obtenção do máximo de confiabilidades e correlação possível, antes da elaboração do modelos logísticos ao simulado.

Tabela 1 - Descrição do Simulado com 21 itens

Itens	ρ_{pb}		α_c Todos os itens 0.6139
	Incluído	Excluído	Excluído Item
Q1	0.3745	0.2386	0.5976
Q2	0.1340	0.0066	0.6249
Q3	0.3655	0.2268	0.5994
Q4	0.4566	0.3263	0.5855
Q5	0.2543	0.1128	0.6144
Q6	0.3291	0.2070	0.6022
Q7	0.1870	0.0393	0.6243
Q8	0.2394	0.1008	0.6155
Q9	0.5930	0.4827	0.5629
Q10	0.3776	0.2391	0.5977
Q11	0.4911	0.3657	0.5800
Q12	0.3586	0.2317	0.5990
Q13	0.5337	0.4139	0.5730
Q14	0.1177	0.0014	0.6238
Q15	0.4566	0.3285	0.5855
Q16	0.4210	0.2876	0.5910
Q17	0.0509	-0.0460	0.6251
Q18	0.1967	0.0534	0.6219
Q19	0.4414	0.3096	0.5879
Q20	0.1367	0.0150	0.6232
Q21	0.3906	0.2625	05949

Segundo Pasquali (2011) os valores aceitáveis para elaboração de um modelo logístico conciso para descrever os dados empíricos em um teste, necessitam mínimos de ρ_{pb} = 0.30 e α_c = 0.70. No entanto, em nosso simulado percebemos inicialmente estarmos um pouco distante desses indicadores mantendo as 21 questões em análise, mas também observamos que existem diversas questões com baixo ρ_{pb}, como por exemplo as questões

Q17 com ρ_{pb}= 0.0509 e Q2 com ρ_{pb}= 0.1340. Essas questões em particular, além de possuírem baixa correlação têm em suas exclusões um aumento de forma positiva na confiabilidade do simulado.

Seguindo essa metodologia de excluir questões com baixa correlação e contribuição negativa à confiabilidade, conseguimos filtrar 13 questões que servirão de "régua" em nossa análise e elaboração de modelos logísticos de nosso simulado, conforme podemos ver na

Tabela 2. Mais detalhes sobre o assunto pode ser visto em Klein (2013).

Tabela 2 - alpha de Cronbach

Exclusão acumulada	Alpha de Cronbach
Nenhum item	0.6139
Q17	0.6251
Q17 e Q2	0.6365
Q17, Q2 e Q14	0.6479
Q17, Q2, Q14 e Q7	0.6601
Q17, Q2, Q14, Q7 e Q20	0.6714
Q17, Q2, Q14, Q7, Q20 e Q8	0.6801
Q17, Q2, Q14, Q7, Q20, Q8, e Q18	0.6879
Q17, Q2, Q14, Q7, Q20, Q8, Q18 e Q5	0.6925

Ao eliminarmos as 8 (oito) questões da última linha na tabela 2 teremos o máximo de α_c possível em nosso simulado, dado por 0.6925, menor do que o ideal para este tipo de análise, mas próximo o suficiente para validarmos como ferramenta de diagnóstico confiável.

Além disso, todas as questões restantes possuem bons valores de ρ_{pb}, todos acima de 0.30, indicando assim uma margem considerável de correlação entre as questões e ao traço latente investigado no simulado. A Tabela 3 descreve as treze questões restantes que compuseram nossas análises, juntamente com suas proporções de acertos entre o total de respondentes ao simulado.

Tabela 3 - Descrição do Simulado com 13 itens

Itens	ρ_{pb}	Proporção de Acerto	$\alpha_c = 0.6925$
Incluindo Item			Excluindo o item
Q1	0.3915	40%	0.6861
Q3	0.3634	45%	0.6909
Q4	0.5083	49%	0.6679
Q6	0.3536	25%	0.6877
Q9	0.6308	46%	0.6461
Q10	0.3889	49%	0.6872
Q11	0.5148	55%	0.6667
Q12	0.4233	29%	0.6794
Q13	0.5897	53%	0.6536
Q15	0.4800	41%	0.6723
Q16	0.4380	44%	0.6793
Q19	0.4755	51%	0.6734
Q21	0.4181	67%	0.6808

Uma breve visão sobre os resultados iniciais nos revelam que as três questões Q6, Q12 e Q1 são as menos acertadas, indicando talvez serem mais difíceis no simulado como

um todo, enquanto os itens Q21, Q11 e Q13 foram os mais acertados entre os respondentes, No entanto, essas questões serão mais discutidas após a calibração dos parâmetros e escolha do modelo logístico da TRI que mais se ajuste aos dados empíricos de nosso simulado.

Para geração de tais modelos logísticos utilizamos as funções $rasch()$, $ltm()$ *e* $tpm()$ do pacote *ltm* do *software R Statistic*, responsáveis em calibrar os parâmetros dos modelos ML_1, ML_2 e ML_3 respectivamente. Em seguida utilizaremos a função $plot()$ para exibir todas as três CCI e CII dos respectivos modelos com seus parâmetros calibrados conforme resumimos na Tabela 4.

Tabela 4 – Modelos ML_1, ML_2, ML_3 e seus parâmetros

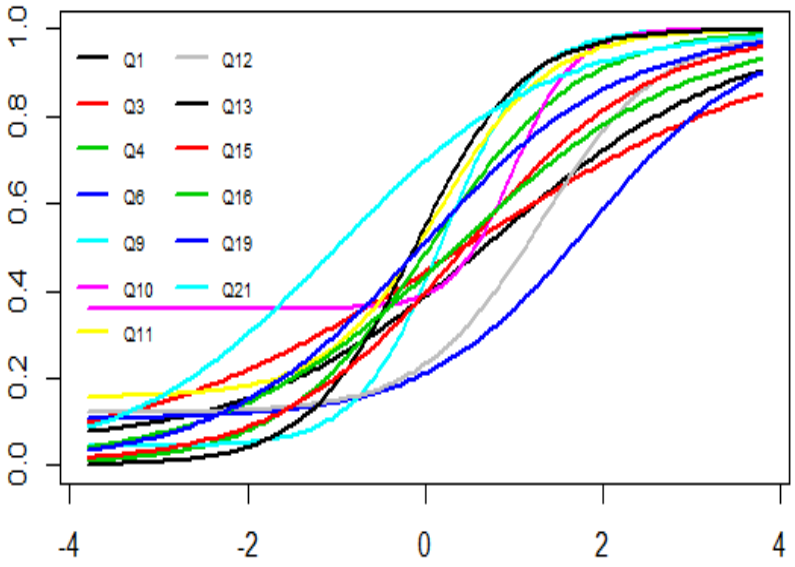
	ML_1		ML_2		ML_3		
Itens	b_i	a_i	b_i	a_i	b_i	a_i	c_i
Q1	0.514	0.914	0.662	0.659	0.791	0.736	0.046
Q3	0.274	0.914	0.456	0.488	0.431	0.522	0.000
Q4	0.056	0.914	0.046	1.187	0.052	1.187	0.000
Q6	1.418	0.914	1.934	0.620	1.861	1.081	0.106
Q9	0.183	0.914	0.121	1.844	0.205	2.079	0.043
Q10	0.074	0.914	0.100	0.609	0.997	2.911	0.359
Q11	-0.252	0.914	-0.212	1.189	0.136	1.598	0.156
Q12	1.128	0.914	1.233	0.811	1.311	1.473	0.121
Q13	-0.179	0.914	-0.125	1.673	-0.116	1.662	0.000
Q15	0.458	0.914	0.447	0.940	0.447	0.952	0.000
Q16	0.292	0.914	0.323	0.795	0.338	0.760	0.000
Q19	-0.052	0.914	-0.054	0.900	-0.051	0.889	0.000
Q21	-0.938	0.914	-0.985	0.856	-0.998	0.839	0.000

A Tabela 4 nos revela que as questões Q6, Q12 e Q1 são de fato as mais difíceis nos modelos ML_1 e ML_2, assim como previsto ao analisarmos suas porcentagens de acertos, no entanto, para o modelo ML_3 a questão Q10 tem sua dificuldade maior do que a questão Q1, isto porque ao adicionar o parâmetro de acerto ao acaso na calibração, se verificou nesta questão um alto valor de "chute", próximo a 36% (trinta e seis por cento), que fez os parâmetros de dificuldade e discriminação saltarem neste modelo. De maneira análoga temos as questões Q21, Q11 e Q13 como as questões menos difíceis nos modelo ML_1 e ML_2, mas no modelo ML_3 a questão Q11 aumenta sua dificuldade consideravelmente devido a calibração de acerto ao acaso próximo a 16% (dezesseis por cento).

Uma visão mais ampla dos resultados da Tabela 4 pode ser vista na

pelas $CCI's$ das questões calibradas de nosso simulado pelo modelo ML_3.

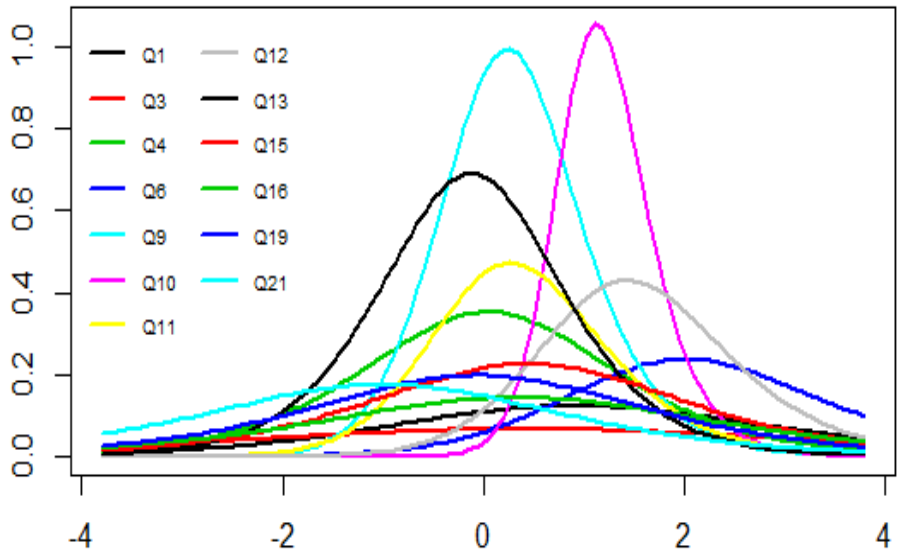
Figura 4 - $CCI's$ das 13 questões do modelo ML_3



Confirmamos pela **Erro! Fonte de referência não encontrada.** que a questão Q6 (cor azul mais a direita) é a que possui o maior parâmetro de dificuldade, enquanto a questão Q10 (cor rosa) é a que possui o maior acerto ao acaso e discriminação do modelo ML_3.

Em complementar a **Erro! Fonte de referência não encontrada.** exibimos as $CII's$ das questões calibradas do modelo ML_3 na Figura 5, onde podemos observar as questões com maiores quantidades de informações ao longo da escala θ.

Figura 5 - $CII's$ das 13 questões do modelo ML_3



A Figura 5 mostra que as questões Q10 (cor rosa) e Q9 (cor azul clara "mais alta") são as que mais possuem informações locais ao longo da escala θ, em destaque novamente a questão Q10 que possui uma grande quantidade de informação local devido ao seu alto parâmetro de discriminação, no entanto, devido ao seu também alto valor de acerto ao acaso sua curva é mais íngreme, indicando assim um poder discriminativo muito restrita na escala θ quando comparamos por exemplo, com as questões Q9 e Q13 (cor preta "mais alta").

Após a calibração dos modelos logísticos ML_1, ML_2 e ML_3 e exibição de suas $CCI's$ e $CII's$ partiremos agora para escolha do modelo logístico que mais se adequa aos dados empíricos de nosso simulado. No pacote *ltm* do *software R Statistic* existem diversas funções aplicadas de forma autônoma ou combinadas afim de auxiliar os aplicadores nesta escolha e optamos neste trabalho pela função $information()$ que é capaz de exibir o quanto de informação o modelo predito possuem sobre os dados empíricos.

Exibimos a princípio a função FIT sobre toda escala de traço latente θ, mas também o comparativo sobre o filtro no intervalo $-3 \leq \theta \leq 3$ como vista na Tabela 5.

Tabela 5 - Informação dos Modelos ML_1, ML_2 e ML_3

Modelos	FIT	
	Toda escala θ	$-3 \leq \theta \leq 3$
ML1	11.88	10.25
ML2	12.55	10.93
ML3	13.34	12.08

A Tabela 5 indica uma melhor calibração do modelo ML_3 em relação aos modelos ML_1 e ML_2, algo que já era esperado pela característica do teste e dos respondentes do simulado. É importante destacar que existem diversas outras funções no pacote *ltm* que podem complementar a escolha do modelo esperado, como por exemplo, a função $anova()$ que calcula o melhor ajuste por meio dos critérios de informações de $Akaike\ (AIC)$ e $Bayesiano\ (BIC)$, ou a função $margins()$ que calcula os resíduos do $Qui-quadrado\ (\chi^2)$ ao comparar os dados observados e estimados dos modelos. Mais detalhes sobre tais resultados podem ser vistos em (BARTHOLOMEW; STEELE; GALBRAITH; MOUSTAKI, 2002), (AKAIKE, 1974) e (SCHWARZ, 1978).

4 CONSIDERAÇÕES FINAIS

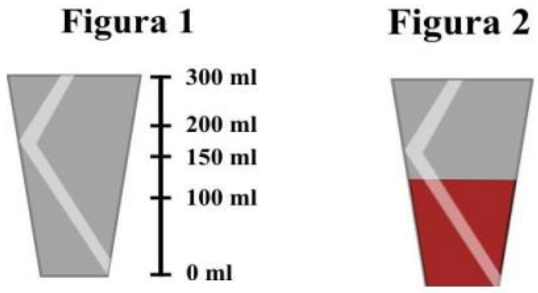
Destacarmos primeiramente a quantidade de questões que foram excluídas das análises para calibragem dos modelos, 8 (oito) questões no total, cerca de 40% (quarenta por cento) das questões do simulado, que a princípio apresentaram em seus padrões de respostas inconsistências dentro do simulado e tiveram suas exclusões necessárias para obtermos o máximo de confiabilidade e correlação aos dados observados.

Após estas exclusões foi verificado que o modelo logístico que melhor se ajustou aos dados empíricos das 13 (treze) questões resultantes foi o modelo ML_3, algo já esperado pela particularidade do tipo de avaliação e respondentes a este simulado, mostrando haver a necessidade dos parâmetros de discriminação e acerto ao acaso para uma melhor explicação dos dados empíricos produzidos pelo simulado.

A partir do modelo logístico ML_3 escolhido podemos fazer uma série de considerações e encaminhamentos sobre os resultado obtidos. Começamos pela questão Q6 que foi a questão com maior nível de dificuldade, tendo apenas 25% (vinte e cinco por cento) de acertos entre o total de respondentes ao simulado. Para melhor entender a questão segue abaixo seu enunciado.

Figura 6 - Questão Q6 do simulado da Prova Brasil - OBEDUC

06 – A Figura 1 representa um copo de 300 ml de capacidade e a Figura 2 representa este mesmo copo com suco.



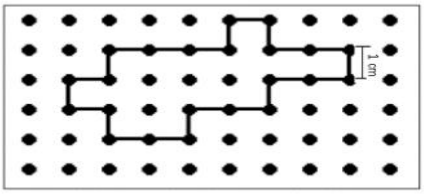
Fonte: Simulado da Prova Brasil aplicada pelo Projeto OBEDUC em 2014

A questão tinha como descritor na matriz de referência da Prova Brasil do 5º ano *Resolver problemas significativos utilizando unidades de medida padronizadas* e revelou uma fragilidade ao trabalharmos com a unidade de medida padronizada ml.

A questão com a maior discriminação no simulado foi a questão Q10, que em outras palavras consegue estratificar com mais precisão os grupos de respondentes que acertaram devido suas habilidades, daqueles que acertaram pelo "chute" (acerto ao acaso).

Figura 7 - Questão Q6 do Simulado da Prova Brasil

10 - Marina usou um elástico para representar uma figura no quadro de preguinhos que a professora levou para a sala de aula. Veja o que ela fez



Observando que a medida entre dois preguinhos é de 1 cm, qual é o perímetro da figura que Marina representou?

Fonte: Simulado da Prova Brasil aplicada pelo Projeto OBEDUC em 2014

A questão Q10 tinha como descritor *Resolver problema envolvendo o cálculo do perímetro de figuras planas, desenhadas em malha quadriculadas* e teve no simulado 49% (quarenta e nove por cento) de acerto em o total de respondentes.

Sobre acerto ao acaso foi evidenciado que as questões Q1, Q3, Q4, Q9, Q13, Q15, Q16, Q19 e Q21 zeraram ou tiveram valores relevantes em suas respostas, minimizando assim a ideia de "chute" ao simulado para esta questões. São então boas fontes de informações sobre as reais habilidades detectadas entre os diversos descritores existentes em nosso simulado.

Estes resultados provenientes das reais habilidades existentes podem auxiliar uma análise investigativa e construtiva sobre as ações ao simulado. Investigativa ao propor descobrir os principais motivos das dificuldades encontradas pelos respondes e construtiva ao propor materiais e ferramentas que possa contribuir com o ensino em sala de aula. Além disso, segue algumas ações em complementar que podem ser desenvolvidas para aqueles que pretendem replicar o método descrito neste artigo:

- Quais descritores estão associados as questões com menores e maiores índices de dificuldades, discriminação e acerto ao acaso;
- Quais grupos de respondentes possuem maiores e menores traços latentes; e

- Criação de uma escala de habilidades com base em itens ancora, originando assim um padrão mais eficiente para reconhecer as habilidades associadas a cada nível de traço latente.

Todas essas ações podem ser realizadas explorando as funções estatísticas do pacote *ltm* no *software R Statistic*, que também possui diversos outros pacotes para análise de modelos logísticos, como por exemplo o *irtoys* de Ivailo Partchev.

A escolha da TRI como metodologia de apuração do simulado está em consonância com as principais políticas públicas hoje adotadas em todo territorial nacional sobre avaliações, como a Prova Brasil, ENEM, ENADE entre outras de larga escala, e esperamos que este artigo sirva de referência para seu uso mais amplo nas diversas área da educação.

REFERÊNCIAS

AKAIKE, H. "A new look at the statistical model identification", IEEE transactions on automatic control, v. 19, n. 6, p. 716–723, 1974.

ANDRADE, D. F., TAVARES, H. R., VALLE, R. da C. Teoria da Resposta ao Item: Conceitos e Aplicações. São Paulo, ABE, 2000. Disponível em: http://egov.ufsc.br/portal/sites/default/files/livrotri.pdf. Acesso em: 6 jan. 2017.

BARTHOLOMEW D. J., STEELE F., GALBRAITH J. I., MOUSTAKI I. "The Analysis and interpretation of multivariate data for social scientists", Choice Reviews Online, v. 40, n. 01, p. 40- 0338-40–0338, 2002. DOI: 10.5860/choice.40-033.

BIRNBAUM, A. "Some latent train models and their use in inferring an examinee's ability", Statistical theories of mental test scores, p. 395–479, 1968.

COUTO, G., PRIMI, R. "Teoria de resposta ao item (TRI): Conceitos elementares dos modelos para itens dicotômicos", Boletim de Psicologia, v. 62, n. 134, p. 1–15, 2011. Disponível em: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0006-59432011000100002.

KLEIN, R. "Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências", Ensaio: Avaliação e Políticas Públicas em Educação, 2013. Disponível em: http://www.redalyc.org/articulo.oa?id=399538144003. Acesso em: 6 jan. 2017.

LORD, F. "A theory of test scores.", Psychometric Monographs, 1952.

LORD, F. M. "An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability", Psychometrika, 1953. Disponível em: http://link.springer.com/article/10.1007/BF02289028. Acesso em: 6 jan. 2017.

LORD, F. M. Applications of item response theory to practical testing problems. Routledge, Lawrence Erlbaum Associates Publishers, 1980. v. 1.

PASQUALI, L. Psicometria: teoria dos testes na psicologia e na educação. 4. ed. Petrópolis, VOZES, 2011.

PASQUALI, L., PRIMI, R. "Fundamentos da Teoria da Resposta ao Item–TRI Basic Theory of Item Response Theory–IRT", Avaliação Psicológica, 2003. Disponível em: http://hostel.ufabc.edu.br/~daniel.miranda/wp-content/uploads/v2n2a02.pdf. Acesso em: 6 jan. 2017.

RASCH, G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford, Nielsen & Lydiche, 1960. Disponível em: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1962-07791-000&site=ehost-live. Acesso em: 6 jan. 2017.

RIZOPOULOS, D. "ltm: An R package for latent variable modeling and item response theory analyses", Journal of Statistical Software, v. 17, n. 5, p. 1–25, 2006. DOI: 10.18637/jss.v017.i05. Disponível em: http://lmr.zozlak.org/SkalowanieJednoWymiarowe/Rizopoulos_2006_ltm An R Package For Latent Variable Modeling and IRT Analyses.pdf.

SCHWARZ, G. "Estimating the Dimension of a Model", The Annals of Statistics, v. 6, n. 2, p. 461–464, mar. 1978. DOI: 10.1214/aos/1176344136. Disponível em: http://projecteuclid.org/euclid.aos/1176344136. Acesso em: 16 jan. 2017.

THURSTONE, L. L. "Attitudes Can Be Measured", American Journal of Sociology, v. 33, n. 4, p. 529–554, 1928. DOI: 10.1177/014572178801400303.