Comparação de Regras de Seleção de Itens em Testes Adaptativos Computadorizados: um estudo de caso no ENEM

Victor M. G. Jatobá[1], **Karina V. Delgado**[1], **Jorge S. Farias**[2], **Valdinei Freire**[1]

[1]Escola de Artes Ciências e Humanidades – Universidade de São Paulo (USP)
Caixa Postal 03.828-000 – São Paulo – SP – Brazil

[2]Departamento de Ciências exatas e da Terra – Universidade do Estado da Bahia (UNEB)
Caixa Postal 41.150-000 – Salvador – BA – Brazil.

{victorjatoba,kvd,valdinei.freire}@usp.br, jfarias@uneb.br

Abstract. *Computerized Adaptive Testing (CAT) based on Item Response Theory allows more accurate assessments with fewer questions than the classic P&P test (Paper and Pencil). Studies showed that CAT using the Fisher Information selection rule reduced the size of the* Mathematics and its Technologies *ENEM test by 26.6% in relation to the P&P test, which has 45 items. However, the impact of the use of different item selection rules on the estimation of the examinees scores is unknown. The objective of this work is to analise this impact. The results show that the size of this test can be reduzed more with the use of other item selection rules without significant loss of accuracy in the estimation of the proficiency level.*

Resumo. *Testes Adaptativos Computadorizados (CAT), baseados na Teoria de Resposta ao Item, permitem fazer testes mais precisos com um menor número de questões que a prova clássica P&P (Paper and Pencil). Estudos mostraram que CAT, utilizando a regra de seleção de Informação de Fisher, reduziu em 26,6% o tamanho da prova de* Matemática e suas Tecnologias *do ENEM de 2012 em relação a prova P&P de 45 itens. Entretanto, não é conhecido o impacto na estimativa dos escores dos respondentes no uso de diferentes regras de seleção de itens. O objetivo deste trabalho é analisar esse impacto. Os resultados mostram que o tamanho dessa prova pode ser reduzido ainda mais com o uso de outras regras de seleção de itens, sem perda significativa da precisão na estimativa dos escores dos respondentes.*

1. Introdução

A etapa de avaliação de estudantes sempre foi muito importante no processo de aprendizagem. Na computação, os sistemas de aprendizado geradores de testes auxiliam os estudantes a identificarem se atingiram o nível adequado de conhecimento aprendido [Guzmán and Conejo 2004]. Um exemplo desse tipo de sistema é o Teste Adaptativo Computadorizado (do inglês, *Computerized Adaptive Testing* – CAT).

CATs são testes administrados por computadores que, de forma eficiente, reduzem o número de itens (questões), mantendo um melhor diagnóstico do desempenho do respondente [Kovatcheva and Nikolov 2009, Spenassato et al. 2016]. No CAT clássico, inicialmente, é selecionada uma questão e, a cada nova, é estimado o nível de proficiência do estudante. Caso o critério de parada não seja atendido, outra questão é selecionada.

DOI: 10.5753/cbie.sbie.2018.1453

Existem diversas formas de avaliar o nível de proficiência do respondente. Um modelo bastante utilizado atualmente é o da Teoria de Resposta ao Item (do inglês, *Item Response Theory* – IRT) [Lord 1980]. Esta teoria é composta por modelos matemáticos que procuram estabelecer a probabilidade de um respondente qualquer acertar uma determinada questão, dadas as características do item e as habilidades do avaliado [de Andrade et al. 2000]. Esse é o modelo adotado na prova objetiva do Exame Nacional do Ensino Médio (ENEM) para calcular o desempenho dos estudantes. Atualmente, o ENEM utiliza o modelo da IRT unidimensional de três parâmetros.

CAT, baseado na IRT, permite fazer testes mais precisos [Kovatcheva and Nikolov 2009], pois é possível identificar as áreas de carência do estudante e, assim, selecionar uma sequência de itens adaptada ao conhecimento do respondente [Chen et al. 2005]. Porém, a construção de CATs envolve alguns questionamentos-chave, como a escolha dos critérios de inicialização e finalização do teste e da regra de seleção de itens (do inglês, Item Selection Rule – ISR) [Wainer et al. 2000]. A escolha mais adequada dos questionamentos-chave, pode melhorar a precisão e a eficiência na estimativa das habilidades dos respondentes, principalmente em relação às ISRs [Chen et al. 2000, Chen and Ankenman 2004, Barrada et al. 2008, Wang et al. 2011].

A técnica de Informação de Fisher (do inglês, Fisher Information – F) é um exemplo de uma ISR. Speassato et al. (2016) mostraram que o uso dessa regra permitiu uma redução de, pelo menos, 26,6% no comprimento do teste P&P da prova de *Matemática e suas Tecnologias* do ENEM de 2012. Entretanto as técnicas de Máxima Informação ponderada a *posteriori* ($MPWI$) e a Informação de Kullback-Leibler com distribuição a *posteriori* (KLP) possuem vantagens em relação à regra F para estimar respondentes com níveis de habilidades extremos [Chen et al. 2000, Chen and Ankenman 2004]. Também há trabalhos que indicam que as regras de Informação Ponderada pela Máxima Verossimilhança ($MLWI$) e $MPWI$ apresentam um melhor desempenho geral que a regra F [van der Linden 1998, Van der Linden and Pashley 2009].

Assim o objetivo deste trabalho é analisar o impacto na estimativa dos escores dos respondentes em relação ao uso de diferentes estratégias de seleção de itens na prova *Matemática e suas Tecnologias* do ENEM de 2012.

2. Teste Adaptativo Computadorizado

Testes Adaptativos Computadorizados, também conhecidos como Sistemas Geradores de Testes, são testes administrados por computador, que, de forma eficiente, reduzem o número de itens, mantendo um melhor diagnóstico do desempenho do respondente [Kovatcheva and Nikolov 2009].

O fluxograma de um CAT clássico pode ser visto na Figura 1. Inicialmente é selecionada uma questão para o usuário, e após a resposta dele é estimado o nível de proficiência. Caso o critério de parada não seja atendido, outra questão é selecionada e o fluxo se repete. A Figura 2 apresenta um exemplo de um teste realizado por CAT para um usuário de nível de habilidade θ igual a 1,1. O termo *escore verdadeiro* também pode ser usado quando o nível de habilidade é conhecido. Nos momentos iniciais, à medida que o respondente acerta, seu *nível de habilidade estimado* $\hat{\theta}$ cresce. Mas percebe-se que logo após errar a terceira questão, a estimativa do nível de habilidade decresce. Conforme são selecionadas mais questões, a estimativa do nível de habilidade vai ficando cada vez mais

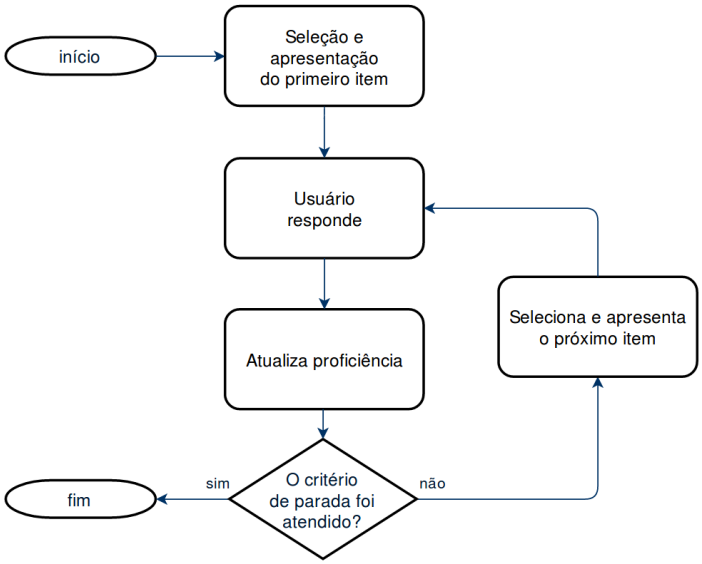

Figura 1. Fluxograma clássico de funcionamento de um CAT (Adaptado de [Olea and Ponsoda 1996])

precisa.

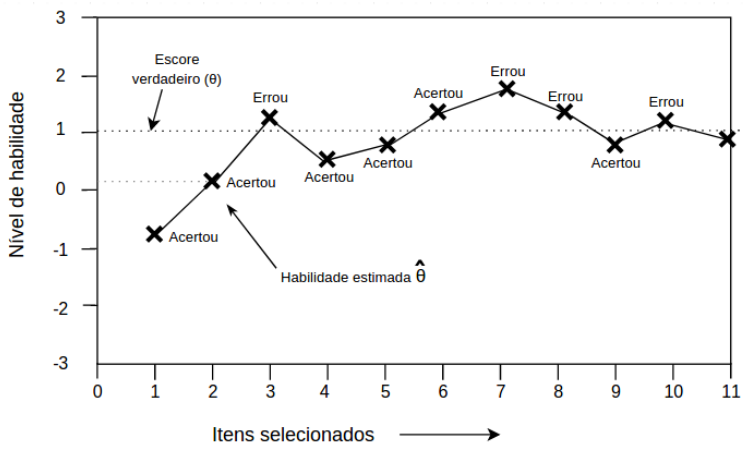

Figura 2. Exemplo do funcionamento de um teste realizado por CAT

Sistemas CAT são compostos por cinco componentes principais: (1) banco de itens; (2) critérios para iniciar o teste; (3) algoritmo ou regra para a seleção de itens; (4) método de estimação da habilidade; e (5) regra de parada do teste [Thompson and Weiss 2011]. Existem diferentes modelos que suportam a criação de CAT, entretanto a IRT é o modelo mais utilizado [López-Cuadrado et al. 2010, Guzmán and Conejo 2004, Wong et al. 2010]

3. Exame Nacional do Ensino Médio

O ENEM é uma prova de formato P&P, promovido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) [Gonçalves and Aluísio 2015]. Foi criado

em 1998, e recebeu uma reformulação a partir de 2009, após a qual passou a ser usado como meio de seleção para o ingresso nas universidades. Concomitantemente adotou a IRT para o cálculo das proficiências dos avaliados [Costa 2015].

O ENEM parte do conceito de competências, que se traduz em habilidades, conhecimentos e atitudes para resolver cada situação-problema. A prova é estruturada em quatro macroáreas: Ciências da Natureza, Ciências Humanas, Matemática e Linguagens e Códigos [Andrade 2013]. Para cada participante, são calculadas quatro proficiências, uma para cada macroárea. A prova é dicotômica, com possibilidade de múltipla escolha e o Modelo Logístico utilizado pela IRT é o de três parâmetros (ML3), com escala estabelecida em 500 e desvio-padrão 100 [Spenassato et al. 2016].

4. Regras de Seleção de Itens

A escolha da ISR influencia diretamente na maior eficiência e precisão na estimativa da proficiência dos respondentes de CAT, em comparação a testes P&P [Eggen and Straetmans 2000]. A seguir, são descritas algumas ISRs.

4.1. Informação de Fisher

A informação de Fisher (F) do item i é definida como sendo [Lord 1980]:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{1}$$

na qual $P_i(\theta)$ é a probabilidade do usuário acertar o item i dado o nível de habilidade θ. Já $P_i'(\theta)$ é a primeira derivada de $P_i(\theta)$. Considerando o Modelo Logístico com três parâmetros, a Equação 1 configura-se em:

$$I_i(\theta) = \frac{2.89(a_i)^2(1 - c_i)}{[c_i + exp[1.7a_i(\theta - b_i)]][1 + exp[-1.7a_i(\theta - b_i)]]^2}, \tag{2}$$

sendo:

- b_i é o parâmetro de dificuldade da questão;
- a_i é o poder de discriminação que cada questão possui para diferenciar os participantes que dominam dos participantes que não dominam a habilidade avaliada na questão i.
- c é o parâmetro de acerto ao acaso.

Usando a informação de Fisher, o item a ser selecionado é aquele que fornece a maior informação para um dado nível de proficiência. A regra de seleção F possui maior eficiência e precisão na estimativa dos escores, quando $\hat{\theta}$ está próximo do *escore verdadeiro* [Chen et al. 2000].

4.2. Informação de Kullback-Leibler

A Informação de Kullback-Leibler (KL) [Chang and Ying 1996] é uma medida geral para a *distância* entre duas distribuições [Van der Linden and Pashley 2009] e é definida por:

$$KL_i\left(\hat{\theta}, \theta_0\right) = P_i(\theta_0) \log\left[\frac{P_i(\theta_0)}{P_i(\hat{\theta})}\right] + [1 - P_i(\theta_0)] \log\left[\frac{1 - P_i(\theta_0)}{1 - P_i(\hat{\theta})}\right]. \tag{3}$$

na qual θ_0 representa o valor real do nível de habilidade do respondente. Quando $\hat{\theta} = \theta_0$, o valor de KL é zero. Isso significa que o item não consegue distinguir entre respondentes de mesmo nível de habilidade. Na perspectiva contrária, quando θ e θ_0 são muito diferentes, o valor de KL é muito grande. Na regra de seleção KL, é considerado um intervalo de confiança $(\hat{\theta}_l, \hat{\theta}_u)$ para a proficiência atual estimada, obtendo-se:

$$KL_i(\hat{\theta}) = \int_{\hat{\theta}_l}^{\hat{\theta}_u} KL_i(\hat{\theta}, \theta)\, d\theta. \tag{4}$$

4.3. Informação de Kullback-Leibler com Distribuição a Posteriori

A regra de Informação de Kullback-Leibler com Distribuição a Posteriori (do inglês, Kullback-Leibler Information with a Posterior Distribution – KLP) [Chang and Ying 1996] inclui uma distribuição a *posteriori* e, ao invés do intervalo de confiança finito, torna o $\theta \in (-\infty, \infty)$. Essa regra é definida por:

$$KLP_i(\hat{\theta}) = \int_{-\infty}^{+\infty} p(\theta \mid X_n)\, KL_i(\theta, \hat{\theta})\, d\theta. \tag{5}$$

Com o KLP, os altos valores de KL para níveis extremos de θ são equilibrados pela distribuição de densidade a *posteriori*, a qual, tipicamente, possui valores de probabilidade menores para níveis de θ extremos [Chen et al. 2000]. Portanto, é mais apropriado o uso de uma distribuição a *posteriori* do que um intervalo de confiança usado pela KL.

4.4. Informação ponderada pela máxima verossimilhança

A regra de informação ponderada pela máxima verossimilhança (do inglês, Maximum Likelihood Weighted Information – MLWI) é definida por [Veerkamp and Berger 1997]:

$$MLWI_i(\theta) = \int_{-\infty}^{+\infty} I_i(\theta)\, L(\theta|x_1, ..., x_k)\, d\theta, \tag{6}$$

na qual k é o numero de itens administrados, $x_1, ..., x_k$ o padrão de respostas provisórias e $L(\theta|x_1, ..., x_k)$ a função de verossimilhança avaliada em θ, dado o padrão de respostas provisórias.

4.5. A regra de Informação ponderada a *posteriori*

A regra de Máxima Informação ponderada a *posteriori* (do inglês, Maximum Posterior Weighted Information – MPWI) é definida por [van der Linden 1998]:

$$MPWI_i(\theta) = \int_{-\infty}^{+\infty} I_i(\theta)\, \pi(\theta)\, L(\theta|x_1, ..., x_k)\, d\theta, \tag{7}$$

sendo que $\pi(\theta)$ é a distribuição a *priori* do nível de habilidade.

5. Método de Pesquisa

A seguir são descritas as abordagens para a captação dos dados, a montagem do Banco de Itens, a construção do CAT e a validação do trabalho.

5.1. Captação dos Dados, Montagem do Banco de Itens e Estimação das Habilidades

Foram utilizados os dados da prova do ENEM de 2012, os quais são públicos e foram retirados do portal da transparência [INEP 2016]. O modelo de dados é composto por um conjunto de respostas dicotômicas de avaliados selecionados de forma aleatória. Cada resposta u, pode conter apenas dois possíveis valores, que são: 1 para "acertou" e 0 para "errou", isto é, $u = \{0, 1\}$, caracterizando a nossa distribuição como uma Bernoulli.

A amostra inicial foi composta por 1000.000 respondentes da prova Rosa de Matemática e suas tecnologias. Os valores dos parâmetros a, b e c dos 45 itens foram retiradas do trabalho [Spenassato et al. 2016]. Esses parâmetros serviram de base para estimar o nível de habilidade dos respondentes com o *software* ICL [Hanson 2002].

O método para a estimação foi o Esperança a Posteriori (do inglês, Expected a Posteriori – EAP), com 10 pontos de quadratura, o mesmo utilizado pelo trabalho [Spenassato et al. 2016]. Essas estimativas são aqui consideradas como os *escores verdadeiros*.

5.2. Configuração do CAT e Definição do Comprimento

Para a montagem do CAT, utilizamos o pacote catR do *software* R. Não foi preciso implementar nenhum critério de exposição de itens, pois todos os avaliados foram submetidos aos mesmos 45 itens. Também nenhuma restrição para balanceamento de conteúdo foi desenvolvida, pois não é divulgado a qual conteúdo cada questão pertence.

Após a estimação dos *escores verdadeiros*, estes foram classificados em 10 grupos entre -2 e 3.5 (ver Tabela 1). A primeira intenção foi entender o comportamento da etapa de estimação. O menor e maior θ foram, respectivamente, -1.716215 e 3.083216. Nota-se que existem poucos respondentes com θs altos (maiores que 2) e θs baixos (menores que -1.5). De cada grupo, foram retirados, de forma aleatória, 500 respondentes que totalizaram outra amostra de 5000. Isso foi importante para garantir que todos os grupos de níveis de usuários façam parte da etapa de simulação do CAT.

Tabela 1. Estatísticas descritivas relacionadas aos θs verdadeiros em cada intervalo

Intervalo dos θs verdadeiros	Tamanho da amostra	Menor θ	Maior θ	Média dos θs
[-2 ; -1.5]	12193	-1.716215	-1.500013	-1.58
] -1.5 ; -1]	121331	-1.499992	-1.000002	-1.19
] -1 ; -0.5]	211557	-0.999995	-0.500002	-0.75
] -0.5 ; 0]	193117	-0.499994	-1e-06	-0.25
] 0 ; 0.5]	163131	1.3e-05	0.499995	0.24
] 0.5 ; 1]	139804	0.500006	0.999997	0.74
] 1 ; 1.5]	95364	1.000001	1.499987	1.23
] 1.5 ; 2]	53275	1.500001	1.999989	1.71
] 2 ; 2.5]	9435	2.000034	2.499647	2.16
] 2.5 ; 3.5]	793	2.500007	3.083216	2.66

Foi adotada a estratégia de [Spenassato et al. 2016] para identificar o comprimento do CAT. Esta etapa identifica o ponto de estabilidade, ou seja, o momento da execução da prova onde a estimativa da habilidade do respondente j tende a se manter estável. O CAT então é finalizado quando a diferença entre o erro padrão do item atual ($SE_{i,j}$) e o erro padrão do item anterior ($SE_{i-1,j}$) é inferior a 1% do erro padrão do item anterior, isto é:

$$|SE_{i,j} - SE_{i-1,j}| < |0,01 \times SE_{i-1,j}|. \quad (8)$$

O cálculo foi aplicado para cada uma das regras de seleção KL, KLP, $MLWI$ e $MPWI$, com $\hat{\theta}$ inicial fixado em 0 (zero). Do grupo dos 5 mil respondentes, foram considerados apenas aqueles que responderam no mínimo a 40 itens, totalizando 4979 respondentes. O *SE* foi estimado pelo método *semTheta* disponível no pacote catR do *software* R.

5.3. Avaliação das ISRs

A Tabela 2 contém o tamanho da amostra e os resultados da identificação do ponto de estabilidade (Equação 8) para cada ISR em diferentes intervalos de θ. Os valores em negrito foram os comprimentos máximos definidos n para cada ISR. Com isso, todas as ISRs foram novamente executadas considerando a regra de parada fixada em n.

Para avaliar o desempenho das ISRs na estimação das habilidades, foram calculados o viés médio (Equação 9) e a raiz do erro quadrático médio (do inglês, Root Mean Squared Error of Estimation – RMSE. Ver Equação 10).

$$vies(n) = \frac{1}{R} \sum_{k=1}^{R} \left(\hat{\theta}_{n,k} - \theta_k \right), \quad (9)$$

$$RMSE(n) = \sqrt{\frac{1}{R} \sum_{k=1}^{R} \left(\hat{\theta}_{n,k} - \theta_k \right)^2}. \quad (10)$$

θ_k é o *escore verdadeiro* do k-ésimo respondente; R é o número total de respondentes e $\hat{\theta}_{n,k}$ é o valor estimado da habilidade do k-ésimo respondente após aplicar n itens.

6. Resultados

Em linhas gerais, as regras KLP e $MPWI$ praticamente não conseguiram estimar usuários de nível alto ($\theta \geq 2$). Com isso, colocaram todos os respondentes com θs verdadeiros maiores que 1.5 no intervalo] 1.5 ; 2]. No outro extremo, a regra $MLWI$ teve pouco êxito em estimar usuários com níveis baixos ($\theta \leq -0.5$). Para mais detalhes, consulte-se a Tabela 2.

A regra de seleção F foi a que selecionou, em média, um maior número de itens para um determinado grupo de $\hat{\theta}$ e a $MPWI$ foi a que selecionou menos, totalizando 35 e 18 itens, respectivamente. Praticamente todas as médias máximas de itens selecionados foram encontradas no intervalo de 1 a 1.5, exceto para a regra de seleção $MLWI$. O resultado obtido para a regra F é similar ao resultado em [Spenassato et al. 2016], no qual a média máxima de itens foi 33.

Tabela 2. Quantidade de respondentes (σ) e média de itens selecionados ($\bar{x}$) para cada intervalo de $\hat{\theta}$

Intervalo dos $\hat{\theta}$ estimados	F		KL		KLP		MLWI		MPWI	
	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$
] -2 ; -1.5]	502	7	48	14	301	8	0	-	0	-
] -1.5 ; -1]	439	8	548	8	552	8	3	**23**	797	8
] -1 ; -0.5]	611	10	477	8	740	9	7	17	723	8
] -0.5 ; 0]	490	14	1047	7	385	13	2544	6	517	9
] 0 ; 0.5]	666	21	629	8	506	14	771	6	585	15
] 0.5 ; 1]	281	19	369	10	708	9	377	7	395	13
] 1 ; 1.5]	665	**35**	406	**21**	242	**24**	166	10	454	**18**
] 1.5 ; 2]	152	25	554	12	1512	7	35	19	1506	9
] 2 ; 2.5]	1142	19	901	15	3	22	769	14	2	2
[2.5 ; 3]	31	27	0	-	0	-	307	9	0	-

A Figura 3 traz o desempenho de cada regra em relação ao número de questões selecionadas para cada grupo de $\hat{\theta}$ estimado. Na maioria dos casos, a regra F tem maior média de itens selecionados para $\hat{\theta}$ maiores que -0.5. Já a regra MLWI foi melhor para $\hat{\theta}$s entre -0.5 e 1.5.

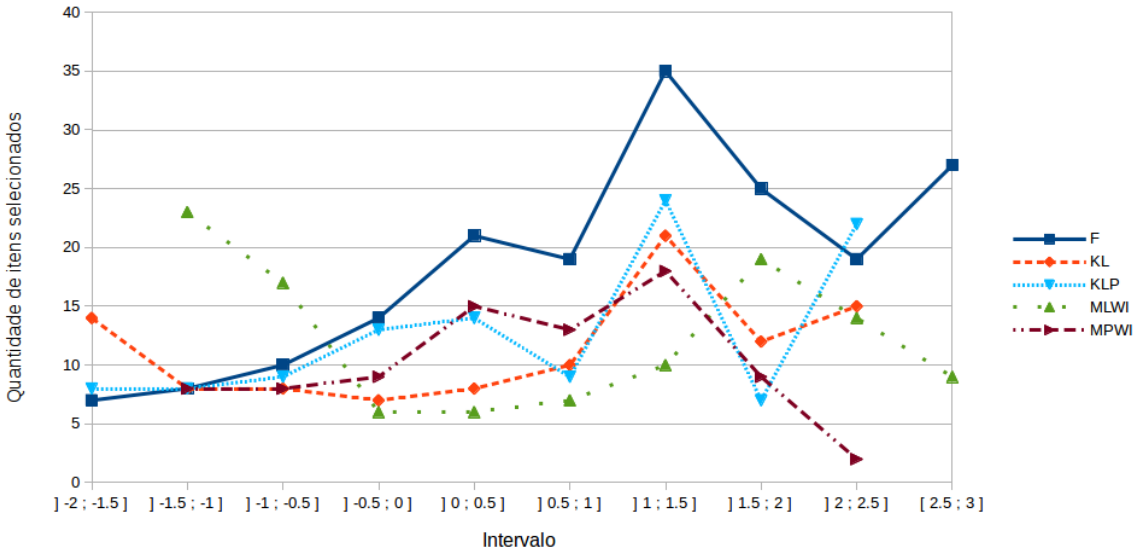

Figura 3. Média de itens selecionados pelas ISRs para cada intervalo dos escores estimados.

Os resultados do viés e do RMSE podem ser vistos na Tabela 3. Há evidências de que as regras KL e $MPWI$ estão subestimando a habilidade dos respondentes, pois estas possuem viés negativo. As regras com a menor raiz do erro quadrático médio, ou seja, as que possuem melhores estimadores são a F e a KLP.

De forma geral, a regra de seleção com maior destaque é a KLP, pois tem o

Tabela 3. Viés e RMSE para cada regra de seleção de itens

	F	KL	KLP	MLWI	MPWI
Viés	0.030	-0.002	0.001	0.067	-0.028
RMSE	0.174	0.273	0.193	0.400	0.294

menor viés, o segundo menor RMSE e permite reduzir em 46.6% o tamanho da prova, sem perda significativa da estimativa do escore do respondente, se comparado ao teste completo com 45 questões.

7. Conclusões

Testes Adaptativos Computadorizados, baseados na Teoria de Resposta ao Item, permitem fazer testes mais precisos para o diagnóstico do desempenho dos respondentes, sendo que a escolha mais adequada da regra de seleção de itens pode melhorar ainda mais a precisão e a eficiência na estimativa das habilidades dos respondentes. Neste trabalho, foram avaliadas as regras de seleção F, KL, KLP, $MPWI$ e $MLWI$ no teste de Matemática e suas Tecnologias do ENEM 2012. Os resultados mostram que, usando a regra de seleção KLP, é possível reduzir em 46.6% o tamanho da prova sem perda significativa da precisão na estimativa dos escores dos respondentes.

Referências

Andrade, G. G. (2013). A metodologia do ENEM: uma reflexão. *Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB*, (33).

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2008). Incorporating randomness in the fisher information for improving item-exposure control in cats. *British Journal of Mathematical and Statistical Psychology*, 61(2):493–513.

Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3):213–229.

Chen, C.-M., Lee, H.-M., and Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, 44(3):237–255.

Chen, S.-Y. and Ankenman, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, 41(2):149–174.

Chen, S.-Y., Ankenmann, R. D., and Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24(3):241–255.

Costa, C. E. S. (2015). *Análise da dimensionalidade e modelagem multidimensional pela TRI no ENEM (1998-2008)*. PhD thesis, Universidade Federal de Santa Catarina.

de Andrade, D. F., Tavares, H. R., and da Cunha Valle, R. (2000). Teoria da resposta ao item: conceitos e aplicações. *ABE, São Paulo*.

Eggen, T. and Straetmans, G. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, 60(5):713–734.

Gonçalves, J. P. and Aluísio, S. M. (2015). Teste adaptativo computadorizado multidimensional com propósitos educacionais: Princípios e métodos. *Revista Ensaio: Avaliação e Políticas Públicas em Educação*, 23(87):389–414.

Guzmán, E. and Conejo, R. (2004). A model for student knowledge diagnosis through adaptive testing. In *International Conference on Intelligent Tutoring Systems*, pages 12–21. Springer.

Hanson, B. A. (2002). IRT command language (ICL). Disponível em: `http://www.openirt.com/b-a-h/software/irt/icl/`. Acesso em: 21 maio 2017.

INEP (2016). Microdados do ENEM. Brasília: Inep, 2016. Disponível em: `http://portal.inep.gov.br/microdados`. Acesso em: 29 mar. 2017.

Kovatcheva, E. and Nikolov, R. (2009). An adaptive feedback approach for e-learning systems. *IEEE Technology and Engineering Education (ITEE)*, 4(1):55–57.

López-Cuadrado, J., Pérez, T. A., Vadillo, J. Á., and Gutiérrez, J. (2010). Calibration of an item bank for the assessment of basque language knowledge. *Computers & Education*, 55(3):1044–1055.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Olea, J. and Ponsoda, V. (1996). Tests adaptativos informatizados. *Psicometría*, pages 731–783.

Spenassato, D., Trierweiller, A. C., de Andrade, D. F., and Bornia, A. C. (2016). Testes adaptativos computadorizados aplicados em avaliações educacionais. *Revista Brasileira de Informática na Educação*, 24(2).

Thompson, N. A. and Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, 16(1):1–9.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2):201–216.

Van der Linden, W. J. and Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing*, pages 3–30. Springer.

Veerkamp, W. J. J. and Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22(2):203–226.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, C., Chang, H.-H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48(3):255–273.

Wong, K., Leung, K., Kwan, R., and Tsang, P. (2010). E-learning: Developing a simple web-based intelligent tutoring system using cognitive diagnostic assessment and adaptive testing technology. In *Third International Conference on Hybrid Learning (ICHL)*, pages 23–34. Springer Berlin Heidelberg.