Como os escores do ENEM são atribuídos pela TRI?

Ricardo Primi [1] & Airton A. Cicchetto [2]

[1] Universidade São Francisco, rprimi@mac.com

[2] Universidade São Francisco, airtoncicchetto@scg.com.br

Resumo

Objetivou-se avaliar como o modelo de TRI de 3 parâmetros atribui as notas aos alunos no Enem, caracterizar o efeito do parâmetro de acertos ao acaso e a relação com a unidimensionalidade da prova. Análises psicométricas numa amostra de 549.265 respondentes à prova de Matemática, indicaram que para um mesmo número de respostas corretas, há uma variabilidade superior a 3 desvios-padrão no escore padronizado estimado pela TRI-3p. Uma análise da dimensionalidade da prova indicou um fator secundário composto por itens difíceis. Há uma parcela considerável de alunos com notas acima da média nesse fator mas com baixa quantidade de acertos na prova como um todo. Em razão da lógica do modelo TRI-3p esses alunos não recebem pontuação por esses itens.

Palavras-chave: Teoria de resposta ao item, equidade, dimensionalidade de testes

1. Introdução

O ENEM é um exame de alto impacto cujo resultado tem potencial para impactar a vida estudantil e profissional de um largo contingente de pessoas. Tendo em vista sua importância julgamos importante buscar entender como se processam as pontuações das provas que constituem o exame, e, principalmente, como a lógica do modelo interpreta as respostas aos itens e processa as pontuações dos participantes.

O ENEM, utiliza a o modelo de 3 parâmetros da Teoria de Resposta ao Item (TRI-3p) o qual caracteriza cada item das provas pela sua dificuldade, discriminação, e probabilidade de acerto ao acaso (De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991 Klein, 2009). Assim, se uma prova foi mais fácil no ano anterior em relação à prova do ano corrente, a TRI utilizará as informações de dificuldade dos itens para compensar essa diferença, de modo que uma mesma quantidade de acertos nos dois anos, serão convertidos em resultados diferentes (valores menores maiores no ano posterior já que, neste, a prova foi mais difícil).

Essa lógica da TRI-3p sugere de maneira genérica que a dificuldade do item interfere na atribuição das notas. Assim supõe-se que se dois alunos A e B tiverem a mesma quantidade

de acertos em uma prova do mesmo ano, e o aluno B tiver acertado itens mais difíceis ele teria uma nota mais alta. Isto, entretanto, pode não ocorrer, pois dependendo do caso, ao aluno B poderia ser atribuída uma nota menor do que ao aluno A. Esta eventual ocorrência pode ser explicada pela lógica do modelo de TRI-3p, em que candidatos com um desempenho global baixo na prova não são pontuados por itens difíceis que tenham respondido corretamente, visto que o modelo assume que acertos em itens difíceis observados em pessoas com baixo acerto no total da prova são decorrentes do acaso (Chiu & Camilli, 2013; Klein, 2013; Lord, 1980).

Essa lógica, de não atribuir pontos a questões difíceis acertadas por estudantes com baixa pontuação na prova, se justifica quando o teste for estritamente unidimensional (Conde & Laros, 2007). Entretanto, há uma escassez de estudos indicando que provas de avaliação de extensão e profundidade de conhecimentos e competências acadêmicas são unidimensionais. Pode-se supor que, eventualmente, um determinado aluno tenha adquirido conhecimento especifico em outra dimensão diferente da dimensão geral da prova, para responder acertadamente itens caracterizados como difíceis. Se esse aluno tiver uma baixa quantidade de acertos no total da prova, os pontos nesses itens específicos não serão creditados. Assim, nesse caso hipotético, mesmo sendo justo que os pontos sejam atribuídos, pela lógica do modelo de TRI-3p, eles não o serão. Isto porque o modelo TRI-3p interpreta o teste como unidimensional, isto é, ela desconsidera a possibilidade de haver uma outra habilidade responsável para resposta a esses itens específicos. Assim assume que a única possibilidade para esse aluno seria o acerto ao acaso.

Dada a escassez de estudos sobre a unidimensionalidade dos testes educacionais, e de como esse aspecto interfere nas pontuações do ENEM o objetivo desse estudo será (a) explorar o fenômeno de pontuação diferencial para alunos com mesmo número de acertos e (b) investigar se há evidência de multidimensionalidade que justificaria acertos em itens difíceis por alunos com baixa habilidade

2. Metodologia

Participantes

Nesse estudo foram analisados dados de 549.253 pessoas que fizeram o ENEM em 2015. Desses 58.6% são mulheres, a média da idade foi M=22,1 e desvio padrão DP=7,6 (variando de 12 a 88 sendo que 70,9% tinha entre 12 e 22 anos). Em 2015 5.59 milhões de pessoas participaram do ENEM. Essa é uma amostra aleatória contendo 9,8% desse total. É uma amostra diversa com pessoas de 5029 municípios de 28 estados brasileiros. Cerca de 53.4%

já tinha concluído o ensino médio, 25.5% estava no terceiro ano e 16.7% estava cursando o primeiro ou segundo ano do ensino médio. O estudo analisou as pontuações da prova de matemática.

Análise de dados

Inicialmente se realizou uma análise descritiva da correlação entre a pontuação atribuída pelo modelo TRI-3p, que está disponível na base, com os escores totais calculados pela soma de itens corretos na prova. Em uma segunda etapa foi criado uma pequena simulação para demonstrar o esquema de ponderação dos itens. Essa simulação foi feita em linguagem R (https://www.r-project.org) e está disponível em http://www.labape.com.br/rprimi/R/3pl_ENEM.html. Na terceira e última fase foi feita uma análise explorando a questão da multidimensionalidade da prova de matemática. Foi feita uma análise de componentes principais dos resíduos obtidos a partir do ajuste do modelo de Rasch (*Rasch-residual-based Principal Components Analysis* - PCAR) implementada no programa Winsteps (Linacre, 2017). A PCAR faz a calibração do modelo de Rasch estimando-se as habilidades das pessoas e as dificuldades dos itens e calcula as respostas esperadas aos itens. A partir disso cria uma matriz de resíduos que conterá a diferença entre as respostas esperadas e aquelas, de fato, observadas. Por fim a PCAR extrai os componentes principais dessa nova matriz de resíduos e apresenta o resultado das cargas dos itens nos componentes não rotacionados, nomeados de contrastes (Linacre, 2003). Segundo Lincare essa análise tenta falsear a hipótese de que os resíduos são aleatórios, padrão esperado quando há somente uma dimensão responsável pela correlação entre os itens. Havendo componentes cuja importância é significativa pode-se suspeitar a influência de mais de uma dimensão. A escolha do modelo de Rasch foi feita para deliberadamente não modelar diferentes associações dos itens com o traço latente (índice de discriminação) e os acertos ao acaso. Assim a variância associada a esses parâmetros se refletiriam nos resíduos. Uma das intepretações das diferenças no índice de discriminação é justamente a multidimensionalidade. Além disso se o acero ao acaso for resultado de um processo aleatório não se devem encontrar correlações sistemáticas entre os resíduos.

3. Resultados e Discussão

Inicialmente serão apresentados os resultados da exploração a pontuação diferencial para alunos com mesmo número de acertos. Na Figura 1 é apresentada uma distribuição dos

alunos de acordo com seu número de acertos e respectiva pontuação no exame de matemática obtida por correção via TRI-3p.

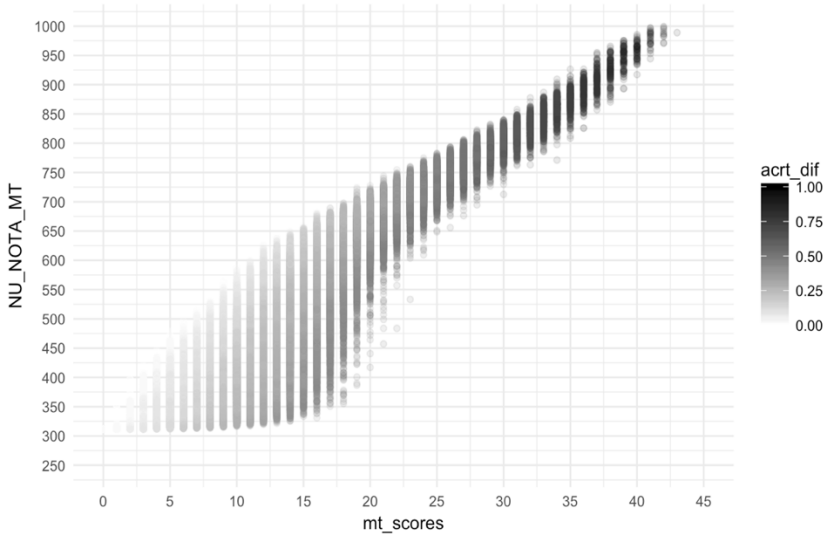

Figura 1. Diagrama de dispersão das pontuações dos sujeitos via TRI 3 parâmetros no eixo y (NU_NOTA _MT) versus número de acertos x (mt_scores) no eixo x coloridos (tons de cinza) pela proporção de acertos nos itens mais difíceis (acert_dif).

No eixo y são descritas as pontuações do exame (NU_NOTA _MT) que tem média M= 500 e desvio padrão DP= 100. No eixo x é descrito o número de acertos dos respondentes na prova de matemática (mt_escores), que pode variar entre 0 a 45 acertos. Também á apresentada uma escala de tons de cinza, acrt_dif, que varia de 0 a 1 e cujo gradiente indica a proporção de acertos nos itens difíceis do teste. Isso foi feito calculando-se a média de acertos em um subgrupo de itens mais difíceis cuja dificuldade era maior que $b > 1{,}5$.

Conforme pode ser observado pela distribuição dos candidatos na figura, para um mesmo número de acertos as notas apresentam variações. Por exemplo, se tomado no eixo X número de 15 acertos, observa-se que há no eixo Y notas que variam de um mínimo de 325 até um máximo de 675 pontos, sendo, portanto, maior que três desvios padrão. O ponto central dessa figura é que os alunos com pontuação abaixo da média, dos 15 itens que acertaram, há uma maior proporção de itens difíceis caracterizados na escala pelo tom cinza escuro. Já os alunos que tiveram pontuação acima da média acertaram uma proporção menor de itens difíceis entre os 15 itens que acertaram caracterizados polo tom cinza claro. A correlação entre as pontuações dos alunos, NU_NOTA_MT e o número de acertos, mt_scores é de $r = 0{,}88$. Nota-se que a discrepância entre as pontuações é muito maior entre os alunos com baixo desempenho.

A lógica de pontuação do modelo de três parâmetros está apoiada na idéia da unidimensionalidade e na admissão da possibilidade de acerto ao acaso por pessoas com baixa

habilidade. Assim admitindo-se a unidimentionalidade, a única explicação possível para ocorrer acertos de pessoas com baixa habilidade em itens difíceis é o acerto ao acaso. Uma explicação alternativa propõe a existência de um fator secundário ligado a subconjuntos de itens, e, especialmente subconjuntos de itens difíceis, que poderia explicar a probabilidade de acerto independentemente da primeira dimensão. Nesse caso se um aluno tivesse uma baixa habilidade no fator primário mas alta habilidade no fator secundário ele poderia acertas os itens desse fator mas ter nota geral baixa na prova. Nesse caso os acertos não seriam devidos ao acaso mas sim decorrentes de uma habilidade específica. Nessa parte do trabalho exploramos a existência de fatores específicos via análise de componentes principais dos resíduos.

Essa análise indicou que do total de variância nas respostas (51,35) 6,5% (3,22) é explicada pelo efeito principal dos parâmetros de habilidades dos sujeitos, 6,1% (3,12) pelo efeito principal dos itens somando-se 12.4% da variância explicadas pelo modelo. A análise extraiu cinco componentes da matriz de resíduos (depois de removido o fator geral dado pela dimensão Rasch) com *eigenvalues* 1,78, 1,44, 1,20, 1,20 e 1,11. Conforme indicado por Raiche (2005) valores maiores de 1,50 indicam a presença de resíduo sistemático. Assim as cargas fatoriais dos itens no primeiro componente foram inspecionadas para se verificar quais itens foram agrupados. A correlação entre cargas e dificuldade sé r = -0,81 contrastando claramente dois grupos de itens: um com itens fáceis (carga positiva no componente, por exemplo, mt43, mt6, mt9 etc.) e outro com itens difíceis (carga negativa no componente, por exemplo, mt14, mt23, mt35, etc).

Ao se examinar os itens difíceis com maior carga no componente (mt14, mt23, mt35) identificou-se tratar de itens envolvendo o cálculo de probabilidade e análise combinatória, portanto, tendo um domínio comum sugerindo a existência de um fator secundário para além do fator geral/primário na prova. A partir disso foram calculados três escores para cada sujeito a partir de três subgrupos de itens identificados na Figura 3 pelas três formas/tons de cinza. Esses três grupos consistem em: (a) c1_meas1: itens fáceis com cargas positivas no fator secundário, (b) c1_meas2: itens fáceis e médios com carga ao redor de zero no fator secundário, e (c) c1_meas3: itens médios e difíceis com cargas negativas no fator secundário. Essa cálculo é feito pelo WINSTEPS no procedimento de PCA. Utiliza-se os parâmetros do modelo do Rasch para estimação dos escores pelo método *Joint Maximum Likelihood Estimation (JMLE)*. As correlações dos escores originais da base no modelo de três parâmetros com esses três escores foram: r = 0,89 (c1_meas1), r = 0,68 (c1_meas2), e r = 0,30 (c1_meas3). Assim o terceiro fator considerado um fator secundário na prova formado pelos itens mais difíceis exibiu

uma correlação mais baixa indicando a independência do fator secundário do fator geral. Assim nota-se claramente que há uma maior discrepância entre esse fator secundário e o fator geral da prova. A média de desempenho em c1_meas3 é M= -1,23 e DP=0,71 e em NU_NOTA_MT é M=465,7 DP=107. Há um grupo considerável de alunos (183.450 pessoas, 33,4% da amostra) nessa condição com desempenho acima da média no fator secundário, mas abaixo da média na nota global do ENEM.

4. Conclusões

Esse estudo descreve a pontuação diferencial para alunos com mesmo número de acertos que ocorre no modelo TRI-3p. Notou-se que para um mesmo número de respostas corretas, há uma variabilidade superior a 3 desvios-padrão no escore padronizado estimado pela TRI-3p. Isso ocorre particularmente no grupo de sujeitos com baixa habilidade, já que, nesse modelo, pessoas com baixa capacidade não recebem pontos por itens difíceis que eventualmente acertem uma vez que o modelo assume que esses casos seriam acertos ao acaso. Essa lógica é justificada quando há uma única dimensão subjacente aos itens. Mas esse estudo demonstra a existência de um fator secundário não negligenciável indicando que há uma sistematicidade nos acertos observados por alunos com baixa habilidade em itens difíceis. Esses dados questionam a prática de não atribuir pontos a esses alunos.

Referências

CHALMERS, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29.

CHIU, T.-W., & CAMILLI, G. (2013). Comment on 3PL IRT Adjustment for Guessing. *Applied Psychological Measurement*, *37*(1), 76–86. https://doi.org/10.1177/0146621612459369

CONDE, F. N., & LAROS, J. A. (2007). Unidimensionalidade e a propriedade de invariância das estimativas da habilidade pela TRI. *Avaliação Psicológica*, *6*(2), 205–215.

DE AYALA, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications.

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, H. J. (1991). *Fundamentals of item response theory.* Thousand Oaks, CA, US: Sage Publications, Inc.

KLEIN, R. (2009). Utilização da Teoria de Resposta ao Item no Sistema Nacional de

Avaliação da Educação Básica (SAEB). *Revista Meta: Avaliação*, *1*(2), 125–140. https://doi.org/10.22347/2175-2753v1i2.38

KLEIN, R. (2013). Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências. *Ensaio: Avaliação e Políticas Públicas Em Educação*, *21*(78), 35–56. https://doi.org/10.1590/S0104-40362013005000003

LINACRE, J. M. (2003). PCA: Data Variance: Explained, Modeled and Empirical. *Rasch Measurement Transactions*, *17*(3), 942–943.

LINACRE, J. M. (2017). *Winsteps Rasch measurement computer program User's Guide.* Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/winman/

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

PRIMI, R., NAKANO, T. C., & WECHSLER, S. M. (no prelo). Using Four-Parameter Item Response Theory to model Human Figure Drawings. *Avaliação Psicológica*.

RAÎCHE, G. (2005). Critical Eigenvalue Sizes (Variances) in Standardized Residual Principal Components Analysis. *Rasch Measurement Transactions*, *19*(1), 1012.