PONTOS FRACOS NA PROVA DA FUVEST

Por Hindemburg Melão Jr.

SÍNTESE:

Neste artigo, discutiremos a efetividade e a equanimidade do tratamento estatístico dado à prova da Fuvest e outros vestibulares, bem como provas escolares, exames universitários, testes psicotécnicos, avaliações personalógicas, encaminhamentos psicológicos, inventários de interesse e todo gênero de questionário usado para os mais variados fins.

Nossas metas com esse artigo são:

- Apontar algumas deficiências nos procedimentos adotados pela Fuvest e em todos os vestibulares, concursos públicos e outros exames.
- Dar um exemplo ilustrativo de falha cometida pela Fuvest, por usar técnicas obsoletas, e comentar algumas conseqüências disso.
- Fazer uma análise comparativa das vantagens que podem ser alcançadas se forem implementadas as mudanças propostas nesse artigo.

INTRODUÇÃO:

O Vestibular da Fuvest é o mais famoso e um dos mais bem reputados do Brasil, com amplos méritos para isso, porque, em comparação a outros exames, a prova da Fuvest é uma das que apresenta questões mais bem adequadas aos objetivos propostos, tanto no conteúdo quanto na distribuição dos níveis de dificuldade, na variedade e na abrangência dos traços cognitivos e epistemológicos medidos. Não obstante, essa prova poderia e deveria ser aprimorada em alguns aspectos muito importantes, que a tornariam mais eficiente e mais justa. Mais eficiente no sentido de elevar o poder de discriminação dos itens e refinar a precisão nos escores; mais justa porque possibilitaria gerar escores individuais mais fidedignos e minimizar efeitos fortuitos de sorte/azar. Por isso, se forem aprovadas e implementadas as mudanças propostas nesse artigo, assim como nos demais artigos dessa série, teremos como resultado a maximização das chances de que as pessoas aprovadas sejam de fato as mais capacitadas, tornando a prova da Fuvest mais apropriada para selecionar os futuros profissionais que dirigirão o país.

Nossas críticas às provas da Fuvest também podem ser dirigidas a todos os demais exames nacionais, inclusive vestibulares do ITA, Unicamp, concursos públicos, exames para ingressar nas escolas de oficiais do Exército, Marinha e Aeronáutica, questionários para análise de *credit scoring* usados por grandes bancos e administradoras de cartão, além de muitos outros.

SOBRE A BIBLIOGRAFIA LUSÓGRAFA QUE VERSA SOBRE ESSE ASSUNTO:

O Prof. Dr. Luiz Pasquali é uma autoridade nacional em Psicometria, lecionou na *University of Michigan* e é professor titular na Universidade de Brasília, autor de vários livros sobre o assunto, com destaque para sua obra "Psicometria", provavelmente o melhor livro disponível em nosso idioma.

Muitas das técnicas apresentadas nesse livro são importantes e, se aplicadas corretamente, são mais eficientes do que as que estão sendo usadas na Fuvest e em praticamente todos os grandes exames educacionais brasileiros.

Na capa posterior do livro, é feita uma crítica muito acertada ao Provão, Enem e Saeb:

> *"(...) O livro vem preencher uma lacuna grave na formação daqueles que trabalham em psicologia e educação no país. Ele se torna, assim, um livro indispensável e visivelmente atual no contexto presente do Brasil, onde se vem insistindo mais e mais sistematicamente na avaliação educacional em nível nacional, através de programas do Saeb, Provão e Enem.* ***Inclusive, a obra instrumentaliza os que trabalham nestes programas no sentido de superar as críticas, muitas vezes justas, que se fazem contra a baixa qualidade científica das avaliações feitas nesses programas.*** *(...)"*

Nessas poucas palavras, Pasquali resume a situação dos exames utilizados com propósitos de triagem e recrutamento, inclusive destacando alguns dos exames mais conhecidos, nos quais seria esperado que houvesse maior preocupação com a qualidade.

Praticamente todas as críticas que Pasquali faz aos vestibulares e a outros exames são procedentes, no entanto as mudanças necessárias ainda não foram implementadas nem tampouco parece haver empenho por parte dos organizadores desses exames no sentido de aprimorar a metodologia utilizada.

VESTIBULAR DA FUVEST:

Os dados disponibilizados no site http://www.fuvest.br/ sugerem que o tratamento estatístico da prova da Fuvest foi baseado na Teoria Clássica dos Testes (TCT), não em Teoria de Resposta ao Item (TRI). Isso não constitui propriamente um erro, mas significa que estão usando métodos do século XIX e primórdios do século XX.

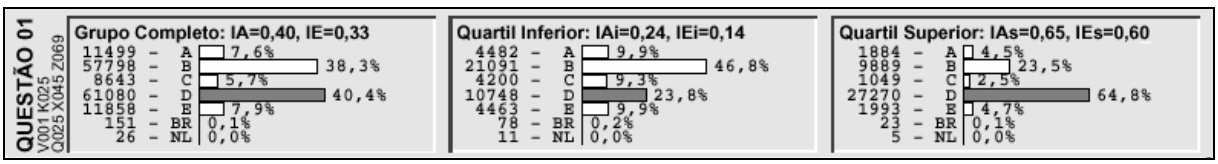
A partir de 1936, foram concebidas técnicas mais avançadas do que a TCT e que oferecem importantes vantagens no tratamento dos dados, possibilitando derivar mais informações relevantes sobre os itens individuais e sobre o teste como um todo, permitindo, entre outras coisas, fazer inferências importantes sobre as mudanças que devem ser feitas nas provas futuras, a fim de torná-las cada vez mais eficazes.

Para compreender melhor o problema, vejamos um exemplo: a questão 1 da prova de Português da Fuvest 2004, cuja resposta certa é a alternativa D, apresentou o seguinte comportamento:

Os 25% de candidatos com escores mais baixos se distribuíram assim:
9,9% dos candidatos escolheram a alternativa A.
46,8% dos candidatos escolheram a alternativa B.
9,3% dos candidatos escolheram a alternativa C.
23,8% dos candidatos escolheram a alternativa D.
9,9% dos candidatos escolheram a alternativa E.

Os 50% de candidatos com escores intermediários se distribuíram assim:
8,0% dos candidatos escolheram a alternativa A.
41,5% dos candidatos escolheram a alternativa B.
5,5% dos candidatos escolheram a alternativa C.
36,5% dos candidatos escolheram a alternativa D.
8,5% dos candidatos escolheram a alternativa E.

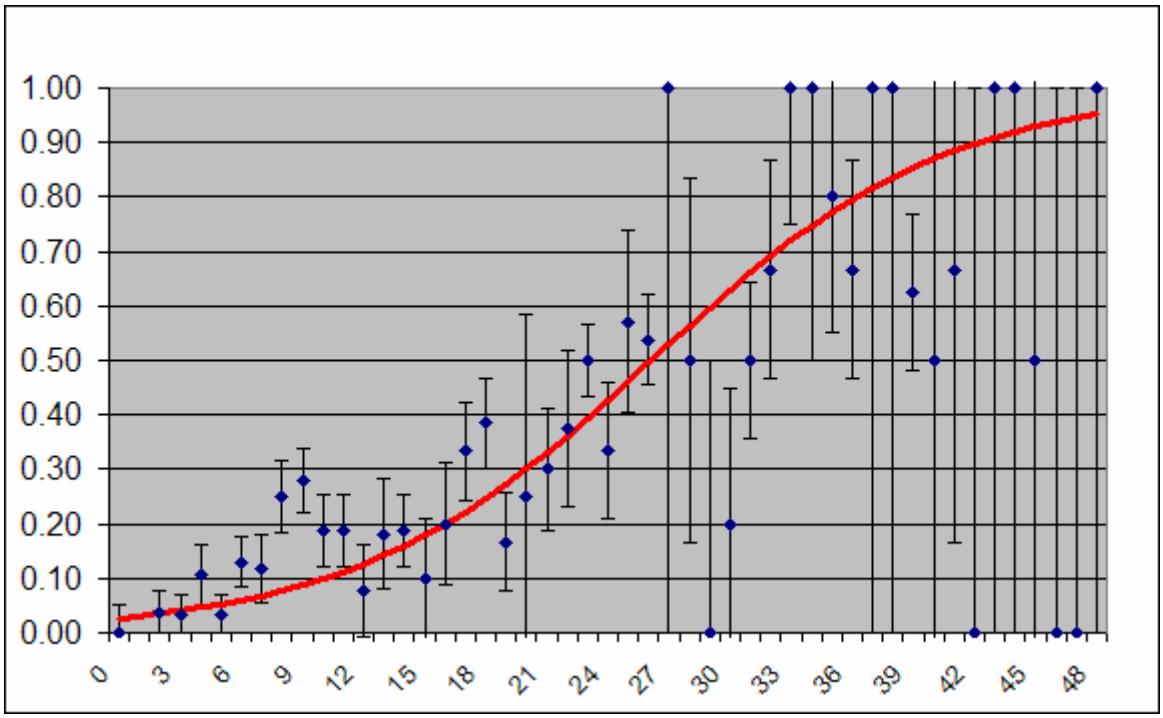
Os 25% de candidatos com escores mais altos se distribuíram assim:
4,5% dos candidatos escolheram a alternativa A.
23,5% dos candidatos escolheram a alternativa B.
2,5% dos candidatos escolheram a alternativa C.
64,8% dos candidatos escolheram a alternativa D.
4,7% dos candidatos escolheram a alternativa E.



O download do arquivo em PDF pode se feito aqui: http://www.fuvest.br/scr/id1f.asp?anofuv=2004&fase=3

O primeiro problema, mas não o mais grave, foi tetratomizar a amostra, ou seja, dividir os candidatos em quatro grupos, cada um com 25% da amostra (quartis), e comparar os desempenhos dos 25% melhores classificados aos dos 25% piores classificados e dos 50% classificados intermediários. Os dois quartis intermediários foram agrupados e por isso também poderíamos dizer que a amostra foi tricotomizada, mas esse é um detalhe semiológico de pouca importância para os nossos objetivos. O que nos interessa saber é

que um procedimento muito melhor teria sido construir uma curva de item, usando o modelo logístico de três parâmetros. Isso continuaria permitindo fazer tudo que é possível pelo método usado pela Fuvest e muito mais, como, por exemplo, verificar o comportamento do item em todos os níveis de habilidade, propiciando uma visão geral de todos os dados em conjunto, identificar mais facilmente uma grande variedade de defeitos e virtudes da questão. Infelizmente não temos como representar essa curva de item sem dispor de todos os dados brutos sobre os candidatos da Fuvest, e essa informação a Fuvest não disponibiliza no *web site*, por isso usaremos a questão 34 do Titan Test, de Ronald Hoeflin, para ilustrar visualmente as vantagens de que falamos:



A curva para a questão 34 do Titan ficaria como na figura acima. O eixo x informa os escores brutos (0 a 48) obtidos no Titan e o y informa a probabilidade de pessoas com cada escore bruto ter acertado a questão 34. Quem tiver interesse em construir as curvas de item para outras questões do Titan, pode encontrar os dados brutos nessa página: http://www.eskimo.com/~miyaguch/titandata/item_ana.html. Para conhecer o modelo logístico de 3 parâmetros, com informações e descrições ilustradas, visite: http://www.uts.psu.edu/Item_Response_Theory_frame.htm ou leia este livro: http://edresearch.org/irt/baker/final.pdf. A maneira como esse site e esse livro abordam o assunto requer algumas críticas, especialmente as afirmações que são feitas sobre os parâmetros b e c, mas isso nos desviaria de nosso tema central, por isso ficará para outra oportunidade.

No gráfico que representa a curva do item 34 do Titan, podemos visualizar claramente uma riqueza bem maior de informações do que nos histogramas de quartil usados pela Fuvest. Vemos uma função logística (curva vermelha) que representa o comportamento do item para todos os níveis de habilidade, temos os dados empíricos (pontos azuis) que indicam as quantidades de acertadores nos diferentes grupos de habilidade e suas respectivas incertezas. Muitas outras informações podem ser derivadas destas, como o nível de dificuldade, o poder de discriminação global ou local, entre outras. A representação que fizemos acima, embora seja muito superior ao histograma da Fuvest, ainda é muito simplificada; poderíamos trabalhar também com as variações dos

parâmetros c e b, e em lugar do modelo unidimensional poderíamos usar um modelo multidimensional, ou introduzir ajustes no modelo unidimensional para torná-lo mais representativo dos dados experimentais, levando em conta eventuais anomalias que afetem a forma logística ou mesmo violem a hipótese da monotonicidade não-decrescente da função. Portanto o nível de sofisticação que poderia ser alcançado é bem mais elevado e propiciaria um farto conjunto de informações que contribuiriam para eleger as questões mais adequadas ao propósito do exame.

O Titan não é um teste de múltipla escolha, por isso o parâmetro c é zero para os níveis mais baixos de habilidade, que pode ser expresso matematicamente dessa maneira: $\lim_{\theta \to -\infty} c = 0$ (c é definido em função θ), onde θ representa o nível de habilidade medido em escore padronizado z ($\mu = 0,\ \sigma = 1$). Mas na prova da Fuvest, como cada questão tem 5 alternativas, o parâmetro c é 20% para os níveis mais baixos de habilidade. Para que a prova cumprisse bem sua missão, o parâmetro c deveria crescer à medida que aumentasse o nível de habilidade, mantendo esse comportamento ao longo de todos os níveis de habilidade. Preciso justificar a afirmação que acabo de fazer, porque estou me opondo ao que diz Frank B. Baker na página 28 do livro linkado acima (*The Basics of Item Response Theory*), em que ele afirma que o parâmetro c não varia em função do nível de habilidade:

> *"The parameter c is the probability of getting the item correct by guessing alone. It is important to note that by definition, the value of c does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. The parameter c has a theoretical range of $0 \leq c \leq 1.0$, but in practice, values above .35 are not considered acceptable, hence the range (…)"*

O fato é que se uma questão tem 5 alternativas, as pessoas com nível de habilidade muito baixo "chutarão" a resposta, sem nenhum critério, portanto suas escolhas serão fortuitas e elas terão $c=0,2$, que representa 20% de chances de escolher qualquer alternativa, inclusive a certa. Mas as pessoas com nível mais alto de habilidade "chutarão" com base em critérios fundamentados em heurísticas eficientes, assim elas poderão perceber que algumas alternativas são evidentemente erradas e eliminar essas opções, aumentando suas chances de acerto, portanto aumentarão o valor do parâmetro c, podendo chegar até mesmo em 0,5 ou mais. Esse é um efeito real e desejável, porque contribui para que o item discrimine corretamente inclusive entre pessoas que não conheçam a resposta inteira. Se João não sabe nada, enquanto Maria sabe um pouco, então João chutará entre 5 opções, mas Maria conseguirá perceber que uma das alternativas não pode ser correta, logo ela chutará entre 4 opções. O efeito disso é que para grandes números de alternativas, a tendência é que Maria marque 25% de pontos a mais do que João. Mesmo que Maria não saiba algumas respostas certas, o fato de ela conhecer um pouco mais do que João possibilita que ela tenha escore "fortuito" maior, graças a essa propriedade do parâmetro c. O resultado final é que a pessoa com mais conhecimento marcará mais pontos. Se o parâmetro c não variasse com o nível de habilidade, seria impossível distinguir entre quem conhece um pouco mais de quem conhece um pouco menos, e ambos receberiam, em média, 20% dos pontos fortuitos.

Por isso afirmamos que nas boas questões o parâmetro c deveria crescer à medida que aumenta o nível de habilidade e manter esse comportamento ao longo de todos os níveis de habilidade. No entanto, algumas vezes observamos evidências de que c diminui com o aumento da habilidade, não como uma flutuação estatística, o que seria normal, mas como um fato real e em proporções grandes o bastante para prejudicar alguns candidatos e afetar a qualidade da prova. O dano será sentido principalmente ao passar de um nível de habilidade ao seguinte e o decréscimo no parâmetro c for maior do que o aumento na probabilidade de acerto não-casual, porque isso violará o comportamento monotônico não-decrescente da função, presumido pela TRI "clássica". Veremos a seguir como ocorre esse efeito, analisando a questão 1 da prova de Português da Fuvest 2004.

No histograma que vimos anteriormente (pág. 3), pudemos perceber que há fortes indícios de que não ocorre o comportamento monotônico não-decrescente, porque nos níveis baixo e intermediário de habilidade a alternativa B se mostra como um atrator de preferências anormalmente forte, que acaba sendo escolhida com maior freqüência do que a alternativa correta, e isso sugere que há um intervalo de níveis de habilidade entre θ e $\theta+x$ no qual a probabilidade de acerto pode ser menor que 20%. Podemos deduzir isso porque nos 25% menores escores observamos 23,8% de acerto, que é apenas 3,8% acima dos acertos casuais. Provavelmente, nos 10% menores escores ocorre um pico de preferência em B que atrai tantas pessoas ao ponto de deixar menos de 20% escolhendo a opção correta D. Se isso acontecer (isso pode ser conferido mediante o exame dos dados brutos), então, como também teremos um intervalo de níveis de habilidade entre $-\infty$ e θ no qual a probabilidade de acerto será 20%, podemos concluir que a função não terá monotonicidade não-decrescente e isso terá importantes implicações, porque haverá um determinado intervalo de habilidade no qual as pessoas menos capacitadas terão mais chances de acertar essa questão do que as pessoas mais capacitadas. Isso pode ser desastroso e uma questão com essas características só deveria ser mantida na prova se demonstrasse possuir outras propriedades especiais muito positivas que compensassem esse defeito. Caso contrário, a questão deveria ser reformulada, ou substituída, ou a alternativa B deveria ser modificada. Antes de decidir sobre isso, seria necessário investigar se o intervalo em que a curva decresce afeta as notas no nível de corte para algum curso. Se não afetasse, então a questão poderia ser mantida, mas pela dificuldade observada (quantidade de acertadores), temos fortes indícios de que ela afeta intervalos de habilidade que correspondem a notas de corte para algumas carreiras. Portanto temos motivos para preferir a exclusão dessa questão, ou reformulação ou substituição da alternativa B.

Alguém poderia argumentar: *mas como poderíamos fazer essa correção, se o problema só pode ser constatado após a aplicação da prova?* A resposta é simples: em primeiro lugar, é possível prognosticar o problema sem aplicar a prova; discutiremos esse tema com mais detalhes num próximo artigo. Em segundo lugar, mesmo que fosse necessário dispor de dados experimentais preliminares para conhecer as propriedades dos itens, não seria difícil conseguir esses dados, e poderiam ser obtidos sem colocar em risco a divulgação antecipada dos gabaritos. Portanto não há nenhuma justificativa aceitável para que a prova da Fuvest contenha alguns itens com propriedades psicométricas tão indesejáveis e tão nocivas como essa questão.

CONCLUSÃO:

Vimos nesse artigo que os grandes exames apresentam problemas que afetam a qualidade seletiva desses instrumentos. Vimos que existem ferramentas claramente superiores às que são usadas. Por fim, analisamos uma das falhas que poderiam ser eliminadas da prova da Fuvest, se o tratamento estatístico fosse mais apropriado e se o processo de construção da prova seguisse normas mais rigorosas.

No próximo artigo, discutiremos algumas normas que podem parametrizar vantajosamente o processo de construção das alternativas, mediante o uso de heurísticas que visam politomizar os níveis discriminados *por cada* item, contribuindo para que a prova tenha máxima eficiência ao selecionar os candidatos mais capacitados.

Sobre o autor: **Hindemburg Melão Jr.** é detentor de três recordes mundiais em atividades intelectuais, um dos quais está registrado na edição de 1998 do *Guinness Book of Records*, páginas 110-111, é membro honorário em várias associações culturais internacionais, inclusive em Pars Society, na Turquia, para pessoas com QI acima de 180 (em cada 3.500.000 de pessoas, apenas uma tem QI no nível suficiente para ser aprovada), membro honorário em ISI Society, na Inglaterra, para pessoas com QI acima de 151 (em cada 1.400 pessoas, apenas uma tem QI no nível suficiente para ser aprovada), sócio honorário em High IQ Society for Humanity, na Dinamarca, que tem como o objetivo orientar e subsidiar crianças talentosas que vivem em regiões carentes, fundador de Sigma Society, para pessoas superdotadas, que atualmente reúne mais de 200 membros provenientes de 40 países dos 5 continentes, autor do Sigma Test e do Sigma Test VI, disponíveis em 13 e 7 idiomas, respectivamente, publicados em 7 revistas internacionais especializadas em inteligência e testes de QI e em mais de 300 web sites internacionais sobre temas afins. Autor de mais de 350 trabalhos publicados em mais de 120 países, inclusive trabalhos premiados em nível TOP-10 mundial, TOP-19 mundial e TOP-26 mundial.

Mais informações sobre o autor em http://www.sigmasociety.com