Duas velhas novidades: a testagem da confiança e a análise *probit* de itens*

Nelson O' de Almeida**

1. A "testagem da confiança": afinal a informação desejada; 2. A análise *probit* de itens.

Busca-se focalizar os problemas e as conseqüências indesejáveis do "palpite" sobre o resultado de testes em geral – especificamente os educacionais – bem como as críticas mais comumente levantadas contra a clássica análise de itens.

A proposta dos Drs. E.H. Shuford e H.E. Massengill, sobre a "testagem da confiança", é apresentada como uma eficiente solução para o problema do "palpite", introduzindo-se pequena modificação na metodologia original, para torná-la ainda mais pragmática.

A solução oferecida pelo Dr. R.S. Das para a análise de itens é apresentada, também com pequenas modificações, como técnica ideal para a mais completa e informativa análise de itens.

Conclui-se citando algumas das vantagens que a adoção de tais "novidades" ofereceria.

1. A "testagem da confiança": afinal a informação desejada

A possibilidade do "palpite" tem demonstrado interferência acentuada nos testes em geral e, especificamente, naqueles ditos "objetivos", como os de múltipla escolha.

Muito, portanto, se tem escrito sobre a justificada preocupação dos administradores educacionais e professores; diversas soluções tentativas têm sido apresentadas, através dos anos, nas pesquisas dos estudiosos (Guilford, 1936; Cronbach, 1949; Gulliksen, 1950; Noll, 1957; Ebel, 1965), embora quase todas – como as utilizadas em nossa realidade educacional – estejam baseadas na penalização das respostas erradas: as célebres "correções para palpite".

Na verdade, a interferência do "palpite", da resposta aleatória tem impregnado os resultados de testes em diferentes sentidos, com as respectivas conseqüências perniciosas sobre fidedignidades, índices de dificuldade e de discriminação, validade e, pior que tudo, sobre o que podem inferir os professores e as escolas acerca do real conhecimento de seus estudantes, ao conhecer-lhes o grau obtido. As tentativas de eliminar ou controlar a influência do "palpite" – entre as mais utilizadas, as fórmulas de Cronbach e as de Gulliksen e os mui pouco conhecidos métodos de Coombs-Milholand-Womer e de Ebel – mostraram-se de eficiência limitada ou de tal modo complexas que ficaram restritas à pesquisa dos estudiosos.

Há cerca de 10 anos, entretanto, Shuford e Massengill (dois jovens doutores com que tivemos a chance de privar ligeiramente, quando cursávamos a Air University), após feliz análise do problema, propuseram uma real evolução metodológica, baseada em sólidos princípios psicológicos bastante simples e verdadeiramente informativa, que permite obter-se mensurações mais válidas, confiáveis e oferecendo, ainda, útil classificação dos estudantes em termos do grau de conhecimento que demonstraram possuir, permitindo identificar a ajuda de que possam, ainda, precisar.

* Artigo apresentado à Redação em 31.11.81.

**Endereço do autor: Rua Conde de Bonfim, 549 - 20.520 - Tijuca - Rio Comprido - RJ

A filosofia é de que sempre será melhor gratificar que penalizar; a teoria é a decisão, combinada com os clássicos da psicometria; a estratégia, a de demonstrar ao estudante que poderá maximizar seu resultado se, e somente se, responder honestamente a cada questão; a técnica é a da declaração do grau de confiança que possa ter na solução que deu ao problema, no desempenho que vem apresentar ou na alternativa escolhida.

O "palpite" é importante para o aluno na medida em que pode alterar seu resultado, em cada avaliação – mas o é ainda mais para o professor, para o sistema educacional, em termos de como podem ser (e são) interpretados os resultados obtidos pelos estudantes (já que degrada seriamente os índices, a validade, a fidedignidade e impede qualquer assistência extra necessitada).

Hoje o sistema clássico, de apontar-se uma alternativa, estimula o aluno a (e ele aprende depressa ser isso de seu suposto melhor interesse) assinalar uma das alternativas, saiba ou não qual seja a correta: o aluno que responder aleatoriamente a todos os itens de um teste – com ou sem a sagrada e inadequada "correção para o palpite" – tem chances definidas de obter resultados aceitáveis, próximos dos daqueles que respondem seus testes com absoluto conhecimento. O anedotário educacional, como o exemplo a seguir, é bastante rico na caracterização das conseqüências típicas.

Imagine-se o item seguinte, extraído de uma prova "objetiva" sobre geopolítica brasileira para estrangeiros:

A capital do Brasil é:
a) Brasília;
b) Rio de Janeiro;
c) Buenos Aires;
d) São Paulo.

Frau Berta Schatz, de Munique, assinalou o *a*; M. Jean Passepartout, de Lion, marcou *a*; Mr. John Popcorn, de Oshkosh, também escolheu o *a*.

O professor, contendo a custo sua euforia, reconhece orgulhoso o conhecimento de seus alunos; a escola, satisfeita, reconhece o valor de seus cursos e seus mestres; oficialmente os três estudantes obtêm o mesmo grau e são classificados como "conhecedores" do assunto. . . Mas verifique-se – talvez através de poderes parapsicológicos, os únicos disponíveis no método atual – o que, na realidade, se passou:

Frau Schatz, estudiosa e interessada, sabe até que a "nova" capital já ultrapassou o primeiro milhão; M. Passepartout, menos informado, lembra-se entretanto de que Buenos Aires *"c'est là bas"* e de que São Paulo é a capital industrial: entre Rio e Brasília não pode decidir-se; fecha os olhos e (. . .) *voilà*: Brasília.

Mr. Popcorn, recém-admitido ao curso, não saberia distinguir entre qualquer das quatro *(holly cow!)*: mas como era preciso (. . .) escolheu Brasília.

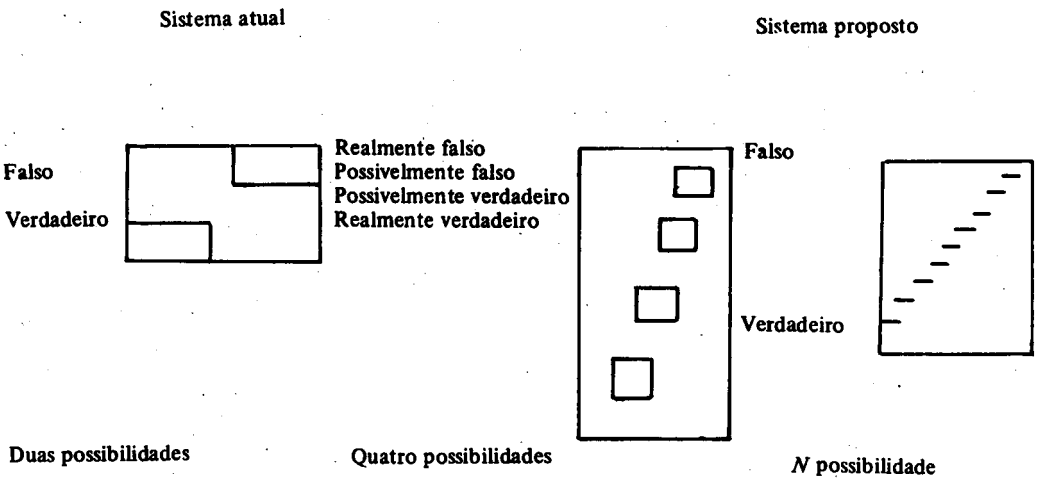
As conseqüências sobre validade e toda a sorte de controles estatísticos que devem exibir os bons testes estão bastante claras . . . e os estudantes, sem auxílio extra, entregue às respectivas sortes: professor, alunos e escola muito mal-informados e, possivelmente, felizes.

Ora: para cada item de uma prova de múltipla escolha, o estudante estará, certamente, analisando cada alternativa oferecida em termos da confiança de que se vê possuído de que seja ela a correta; se reconhece que uma delas está errada, aloca-lhe 0% de confiança e examina a seguinte; se tem dúvidas, pode alocar um certo grau desconhecido de confiança e passa à seguinte.

Nos extremos, se tem absoluta certeza de que dada alternativa é a correta, aloca-lhe 100% de confiança (e declara isso fazendo a marca prevista); se não sabe como definir-se nem consegue eliminar alguma das respostas, terá sua confiança igualmente dividida entre todas as alternativas (mas infelizmente, para todos, marcará uma delas (. . .) no "palpite"). Assim, é literalmente impossível ao professor, como ao sistema, classificar os alunos em mais de duas categorias, por item: ou "sabe tudo" ou "não sabe nada". Conhecimento, entretanto, não é "tudo-ou-nada" (. . .) e ricas informações ficam perdidas no sistema atual de resposta única.

As mais elaboradas tentativas de solução para o problema, como as apresentadas por Coombs ou Ebel, embora representando passos decisivos para o seu equacionamento e solução, não chegaram a controlar, de forma prática, exeqüível, o fenômeno. Shuford e Massengill imaginaram, então, a resposta como uma superfície contínua de confiança e observaram, inteligentemente, que o sistema de resposta em uso poderia ser o responsável pelo estímulo ao "palpite". Para um teste falso-verdadeiro, por exemplo, ter-se-iam as possibilidades apresentadas da figura 1.

Figura 1



A escala bipolar leva ao "tudo-ou-nada"; a de quatro categorias oferece discriminação bem melhor. A declaração de confiança na correção da resposta, entretanto, permite maior discriminação possível, indicando, de fato, o grau de conhecimento (e do socorro necessário ao aluno), já que os palpites estarão eliminados, pois foram transformados em diferentes graus de confiança.

Como principal estímulo para que os estudantes viessem a aderir ao "teste de confiança", ofereceu-se uma transformação logarítmica das percentagens de confiança em notas, que garantem aos alunos "honestos" um ganho de pontos sempre que chegue a alocar alguma confiança na resposta correta.

Ao mensurar-se uma *performance* qualquer, devem os concorrentes declinar seu grau de confiança em que obterão determinado resultado, e passar à ação. Um tal processo permite, ainda, que se obtenha informações valiosas sobre o grau de realismo com que cada concorrente se auto-avalia – e foi possível demonstrar que, após pequena experiência com o método, pode-se prescindir aleatoriamente da execução da *performance* – que está altamente relacionada com a simples declaração do grau de confiança.

Em uma nova resposta livre – como uma dissertação, em 30 linhas, sobre o efeito da inflação sobre as pequenas economias ou os fatos capitais da proposta da gestalterapia – deve o aluno redigir seu ensaio e, *ex post facto*, declarar o grau de confiança que tem de ter coberto devidamente o assunto, ter sido claro, etc.

O processo oferece, ainda, duas outras informações padronizadas, de extremo interesse educacional. A primeira trata de classificação do grau de conhecimento do estudante – o que permite socorrer os necessitados – em seis classes, conforme regras bastante explícitas; conforme o grau de confiança alocada à resposta certa, dir-se-á serem, os estudantes, "bem informados", "desinformados", etc. até a classe dos "muito mal-informados". A segunda propõe regras também simples que permitem mensurar-se o grau de realismo com que os alunos alocam a sua confiança – o que é excepcional valor para testes de *performance* – tanto nas respostas corretas como nas outras alternativas.

Quando se atenta para o fato de que a qualificação do conhecimento, buscada por testes, provas e desempenhos, deveria ter por principal objetivo o "aprendizado" e seu aperfeiçoamento, é lógico compreender-se as diversas vantagens de um método que evita palpites, oferece tantas informações a mais e permite o cálculo de índices e coeficientes bem mais precisos.

O sistema proposto pelos Drs. Shuford e Massengill, afinal, exige que cada estudante receba e manipule uma espécie de calculador de mão, o *scorule*, o que lhes permitirá transformar seu grau de confiança na nota logarítmica prevista. A experiência aqui vivida, entretanto, desaconselha tal processo, que pode ser tensional para muitos estudantes – já que representa uma operação a mais para responder a cada item – e, realmente, aumenta o tempo necessário para calcular e registrar a resposta. Mas viu-se ser desnecessário, senão indesejável, que o estudante tenha de preocupar-se com o cálculo de sua nota: mais eficiente é pedir-lhe, apenas, que registre a percentagem de confiança, ficando todas as outras operações para serem realizadas *a posteriori*, pelo professor ou pelo órgão de avaliação da escola – sem nenhum prejuízo para o método e com todas as vantagens. Um órgão de avaliação do aprendizado, ou o professor, fará isso mais fácil e propriamente, inclusive podendo também computar o resultado clássico, para fins de comparação; além disso, a classificação dos estudantes em termos de conhecimento e de realismo, embora muito simples e exigindo, apenas, certas comparações, seria mais propriamente realizada pelo professor ou agente de avaliação, usualmente mais treinados para tanto.

Foram também criadas técnicas especiais, e muito ricas, para a análise de itens de uma prova *confidence*, oferecendo os clássicos índices de dificuldades e de discriminação, respectivas variâncas, fornecendo a informação estatística típica – com as restrições clássicas, hoje, solucionadas pela análise *probit* de itens, conforme proposta por Das.

Os resultados obtidos em continuada experimentação, como apresentados pelos autores, demonstram uma redução da variança, em prognoses, da ordem de 35% – o que é bastante interessante e bem próximo do que se obteve aqui.

Por tudo isso pode-se pensar em conseguir, agora, a informação há tanto procurada e a qualidade pretendida.

2. A análise probit de itens

Guilford (1936) apresentava 19 métodos diferentes para a análise de itens, embora recomendasse apenas dois, muito pouco conhecidos entre nós; Anstey, sistematizando um bom e geral estudo, reconheceu seis abordagens típicas, oferecendo um total de 27 índices diferentes. Perma-

nece, entretanto, o fato de que qualquer método de análise de itens terá, por principal objetivo, responder às cinco perguntas básicas seguintes:

1. quão difícil é o item (índice de dificuldade);
2. quão bem discrimina (índice de discriminação);
3. como se relaciona a um critério (r_{it} ou r_{ic});
4. quais os melhores itens;
5. quão "paralelos" são os itens.

Os métodos clássicos têm respondido até a terceira pergunta, mas, ao fazê-lo, utilizam artifícios sujeitos a críticas. Os índices de discriminação, por exemplo, são normalmente obtidos com base nas propostas de Kelley e de Flanagan, comparando-se o número de acertos no "grupo superior" (usualmente 27% de *N*) e no "grupo inferior" da amostra, abandonando, assim, as informações contidas na parte central da distribuição em estudo; os índices gerais *(ID* e *PD)* e as correlações item-resultado total ou item-critério são obtidos em operações independentes, quando normalmente vai-se tentar integrá-los na análise lógica final. Além do mais, nenhum índice específico é oferecido para a seleção ou "paralelice" dos itens, ficando tais decisões a cargo da habilidade ou experiência do técnico – nem sempre com resultados defensáveis. E apenas uma das abordagens reconhecidas avalia os resultados em termos de diferentes níveis de habilidade (e, assim mesmo, parcialmente): o célebre método de Burt, considerado, entretanto, complexo e, infelizmente, mui pouco conhecido por nós.

Resumindo: nenhuma das seis abordagens codificadas por Anstey (ou dos 19 métodos citados por Guilford) responde a todas as cinco perguntas básicas, como nenhuma delas obtém resultados de forma integrada; nenhuma, ainda, oferece a possibilidade de poder-se calcular os erros de estimativa incorridos, todas têm os resultados radicalmente relacionados com a amostra estudada e apenas uma delas examina, com restrições, os níveis de habilidade.

A proposta de Das transporta para a atualidade um método baseado em processos psicofísicos, de precisão matemática, condensando antigas tentativas (Ferguson, 1942; Lawley, 1943, Finney 1944; Lord, 1952) em técnicas simples e diretas. Responde às cinco perguntas básicas e obtém os índices de forma integrada, numa só operação, justificando e facilitando sua utilização na análise final; além de mais ricas informações, oferece meios de calcular-se os erros de estimativa em que se incorrerá, examina toda a distribuição de resultados, diversos níveis de habilidade e, assim, é mais sadia em termos de psicometria clássica, além de não depender radicalmente da amostra, apresentando índices mais estáveis em validações cruzadas. O princípio fundamental é o de transformar-se uma distribuição acumulada das percentagens de acertos por classe *(sigmóide)* em uma reta, através da integral normal, obtendo uma reta de regressão *probit*.

A operação inicial compreende a distribuição dos resultados de 7 a 10 classes (nossa experiência mostrou que, para turmas de 35/40 alunos, 7 classes oferecem o melhor resultado), a determinação de acertos percentuais por classe e a correspondente transformação em resultados *probits*.[1] Busca-se, agora, determinar a regressão dos níveis de habilidade (*Pm* das classes) sobre os *probits*, determinar o resultado equivalente ao *probit* 5 (média observada, ou índice de dificuldade) e o desvio-padrão observado (índice de discriminação). A equação de regressão *probit* representa, *au grand complet*, a relação existente entre as classes (níveis de habilidade) e os re-

1 A distribuição *probit* é uma distribuição normal, limitada a $-4\,\sigma$ e $+4\,\sigma$, $\overline{X} = 5$ e $DP = 1$; para obter-se o valor *probit* correspondente a uma dada proporção de acertos basta somar-se 5 ao valor do z encontrado numa tabela de curva normal, correspondente ao valor de p. Na prática só se utilizam os *probits* com valor mínimo de 1,91 e máximo de 8,09 (respectivamente 0,001001 e 0,998999).

sultados no item, o que é bem mais informativo que apenas um coeficiente (ó r_{it} ou r_{ic}). Através, ainda, da equação de regressão, calcula-se o *probit* teórico, correspondente a cada ponto médio; segundo Finney pode-se então obter um teste de χ^2 para o item (quão bom ele é), observando-se, mesmo, qual o nível de habilidade que possa ter apresentado resultados anormais. Quando parece necessário calcular-se o erro das estimativas, incorrido quando do uso da equação de regressão *probit*, utiliza-se a análise fracional, gráfica, como proposta por Mahalanobis (1960). Finalizando a análise, a plotagem dos índices de dificuldade contra os de discriminação (que significam, então, "grau de habilidade"), proporciona padrões para a escolha de itens paralelos.

Dois importantes pontos tem-se apresentado como de mais difícil compreensão e aceitação pelo neófito no método de análise *probit*: o primeiro refere-se à possibilidade de obter-se, por meios gráficos (a olho), a equação de regressão *probit* (relação item-habilidade) e, também, a média observada (índice de dificuldade) e o desvio-padrão (índice de discriminação); de fato, o processo gráfico pode apresentar algumas diferenças, mas sempre de características monotônicas e sem interferência notável na análise final. A experiência tem demonstrado que, após duas ou três plotagens, as equações de regressão obtidas diferenciam-se apenas na segunda ou terceira casas decimais, para o operador mediano. Além do mais o método não é particularmente sensível ao número de classes ou à freqüência em cada classe[2].

O segundo ponto, mais incômodo, trata do valor numérico dos índices de dificuldade e de discriminação que são, no método, apresentados em termos das unidades utilizadas na correção das provas, tornando-se, assim, não-comparáveis com outros índices *probits*, calculados em outras provas. Resolveu-se, pois, calcular tais índices de modo ligeiramente diferente, reduzindo-se a unidades abstratas que permitissem a comparação, com valores variando de 0 a 1 (- 1 a + 1 para o índice de discriminação) e limitados ao dobro do valor básico observado (média ou desvio-padrão). Passou-se, pois, a calcular o índice de dificuldade como $\frac{1 - |\bar{X}_o|}{2\bar{X}_1}$, adotando, como limite, o valor 1, e o índice de discriminação com $\frac{1 - |S_o|}{2S_1}$ voltando-se após a operação e a limitação em 1, a apor-se o sinal inicial, para indicar as discriminações inversas. Tais resultados assim modificados são mais compreensíveis em nossa cultura educacional.

Como extensão deve-se apontar que, apesar de só apresentar vantagens sobre os métodos clássicos, a análise *probit* de itens oferece índices que podem chocar o analista superficial, de vez que não concordam sempre com os índices clássicos usualmente calculados. Há que notar, entretanto, que os índices *probits* são baseados em *toda* a distribuição – o que, por si só, já apresentaria diferenças notáveis. Uma razoável experiência, com cerca de um ano de cálculo sistemático desses índices, veio a demonstrar a superioridade do método, embora realmente tenha provocado algumas "alergias".

Summary

This paper recalls the problems and consequence of guessing at tests in general – specially at achievement ones – as well as the main critical points put foward against the classical methods of item analysis.

[2] Segundo Finney, deve-se traçar a linha de regressão procurando minimizar as distâncias verticais entre ela e os diversos *probits* já plotados. A existência, hoje, de calculadoras de mão com capacidade para calcular diretamente a equação de regressão terá condenado o processo gráfico à situação de notícia histórica.

Drs. E. H. Shuford & H. E. Masengill's proposition for the "confidence testing" is presented – with some minor adaptations for the sake of simplicity – as a very efficient solution to the guessing problem.

Dr. R.S. Das "probit analysis" is also offered – again with some small modifications – as an ideal method for a more complete item analysis.

As a conclusion some advantages of both methods are listed and a its use recommended.

Referências bibliográficas

Anstey, E. *Psychological tests.* Trad. de seis dedos, Marova, 1976.
Burt, C. Item analysis in educational tests. inédito.
Cronbach, L.J. *Essentials of psychological testing.* Harper, 1949.
Das, R.S. Item analysis by probit methods. *Br. J. Psy., 17,* 1964.
Ebel. R. L. *Measuring educational achievement.* Prentice Hall, 1965.
Ferguson, G.A. Item selection by the constant process; *Psk.*, 1942.
Finney, D.J. *Application of probit analysis to mental tests;* Psk., 9, 1944.
Guillford, J.P. *Psychometric methods*; Mc Graw Hill, 1936.
Gardner, Jr. W.C. Confidence testing. In: *Usaf instructors J.*, 1970.
Gulliksen, H. *Theory of mental tests.* Wiley, 1950.
Lawley, D.N. On problems with item selection. *Proc. Roy Soc.*, 1943.
Lord, F.M. Theory of test scores; *Psk. Monogr.*, n. 7, 1952.
Noll, V.H. *Introduction to educational measurement.* Mifflin, 1957.
Mahalanobis. P.C. Method of frectile graphical analysis. *Econ., 6,* 1960.
Shuford, E.H. & Massengill, H.E. *Technical Reports,* Usaf, 1965/68.