TESTE INTERNACIONAL DE HABILIDADES DESENVOLVIDAS (TIHD): UMA PROPOSTA DA ASSOCIAÇÃO INTERNACIONAL DE AVALIAÇÃO EDUCACIONAL *

Protase E. Woodford**

INTRODUÇÃO(1)

A Associação Internacional de Avaliação Educacional (IAEA), formada por 44 entidades dedicadas à avaliação, em 28 países, por ocasião de seu encontro anual, realizado em Nairobi, em maio de 1977, considerou que o conceito de um teste internacional de aptidão para ingresso em universidades constituía um mecanismo promissor no sentido de facilitar o intercâmbio internacional de estudantes.

MOTIVOS E FINALIDADES DE UM TESTE INTERNACIONAL DE APTIDÃO

O movimento internacional de estudantes do ensino superior é um fato. Mais de 250.000 estrangeiros estão atualmente estudando nos EUA. Anualmente, a França acolhe cerca de 120.000, o Canadá 100.000, e o Reino Unido, a República Federal Alemã e a URSS cerca de 50.000 cada. À medida que um número crescente de indivíduos deixa seu país de origem para estudar em outro, os mecanismos que visam a assegurar este intercâmbio vêem-se comprometidos. Os procedimentos de admissão de um país estrangeiro freqüentemente não se coadunam com o sistema educacional e a informação disponível sobre o país de origem. Muitas vezes, aplicam-se testes ou exames a fim de se superar esta dificuldade, mas, nos casos em que as línguas

* Tradução de Francis Henrik Aubert, da Universidade de São Paulo.

** Do *Educational Testing Service* (ETS), Princeton, Nova Jersei.

(1) O presente artigo é uma versão condensada do relatório apresentado ao Dr. William W. Turnbull, Presidente do *Educational Testing Service*, para discussão durante a reunião anual da *International Association for Assessment in Education*, realizada em Manilha (Filipinas), no período de 25 a 29 de maio de 1981. *(Nota da Redação)*

dos dois países são diferentes, as informações fornecidas pelos testes podem facilmente ser mal interpretadas.

Um elemento crítico na tomada de decisões sobre admissão é a estimativa da capacidade do candidato de realizar o tipo e o nível de trabalho exigido no país e na instituição hospedeira. Testes de aptidão ou de habilidades desenvolvidas comprovaram sua validade para fornecer indícios úteis sobre tal capacidade. Contudo, dependem de um domínio da língua. Quando a língua do teste e a do candidato não coincidem, é difícil saber se um fraco desempenho resulta de um déficit lingüístico, superável através de treino suplementar, ou de uma falta mais fundamental de habilidades desenvolvidas para realizar o trabalho acadêmico, independentemente da língua de ensino.

A questão de como lidar com tal situação constituiu um dos tópicos da pauta de discussões na reunião da Associação Internacional de Avaliação Educacional, realizada em 1977. Os participantes concordaram em que uma solução parcial para este problema seria desenvolver um teste internacional de aptidão que seria aplicado no país de origem do estudante. Seus resultados seriam transmitidos aos responsáveis pelas decisões sobre admissão antes de o candidato deixar seu país. Formas equivalentes do teste seriam desenvolvidas em tantas línguas quantas fossem necessárias para que a aptidão dos candidatos pudesse ser testada na língua em que fossem mais versados, fornecendo desta forma uma medida mais "pura" do seu desempenho em um contexto acadêmico.

Dado que o trabalho acadêmico no país estrangeiro seria realizado em língua não-materna do estudante, também seria necessária alguma medida de suas habilidades nesta língua. O procedimento proposto possibilitaria estimativas separadas destas duas características, evitando confundir fluência em uma segunda língua com a habilidade mais fundamental para o trabalho acadêmico. Os estudantes com elevado desempenho em habilidades acadêmicas, mas com deficiências em fluência lingüística, seriam encorajados a desenvolverem esta fluência, enquanto que os estudantes com fraco desempenho em habilidades desenvolvidas e, portanto, com poucas chances de serem bem sucedidos, qualquer que fosse a língua, poderiam evitar um investimento inútil em tempo e esforço.

A idéia proposta é a de que o protótipo do Teste Internacional de Aptidão seja projetado para medir, basicamente, apenas duas áreas gerais de aptidão: – verbal e quantitativa. Outro aspecto considerado refere-se ao desenvolvimento de medidas do "raciocínio" ou do "pensamento analítico". O desenvolvimento de itens que fossem ao mesmo tempo diretos, em termos de forma, e "universais", em termos dos processos exigidos para sua solução (de modo que a tarefa a ser exigida de todos os candidatos fosse essencialmente a mesma, independentemente de sua formação anterior), pareceu extravasar o escopo do projeto proposto, embora deva ser retomado em uma etapa posterior.

Os testes-piloto resultarão em pelo menos 4 TIHD equivalentes em Árabe, Chinês, Inglês e Português (um para cada língua). Se as análises de dados indicarem um número suficiente de questões com o mesmo comportamento estatístico desejado, talvez seja possível elaborar mais de um conjunto final de "questões comuns". As questões não utilizadas para os conjuntos de "questões comuns" e seus parâmetros de resposta, no processo de equivalência, permitirão a cada uma das organizações colaboradoras* escolherem questões do conjunto geral para construir outros conjuntos de testes equiparáveis para seu próprio usó no futuro.

QUATRO FORMAS DO TESTE EM CADA LÍNGUA

Assim que cada conjunto de itens for recebido pelo ETS, será enviado a todas as organizações colaboradoras para tradução e adaptação à sua própria língua. O processo de tradução resultará em um total de 2.560 (4x640) questões. As questões verbais e quantitativas serão ordena-

* As organizações colaboradoras, no momento, são: *American University of Cairo, Educational Testing Service,* Fundação Carlos Chagas e a *Hong Kong Examinations Authority. (Nota da Redação)*

das aleatoriamente, de modo a reunir todas as questões em 4 formas do teste-piloto. Cada forma conterá, aproximadamente, o mesmo número de questões de cada uma das quatro organizações colaboradoras.

O conteúdo das formas em Inglês, I-1, I-2, I-3 e I-4, relacionado por número predeterminado, será enviado às organizações colaboradoras. Por sua vez, cada uma preparará quatro matrizes para impressão de todas as formas de testes, na mesma ordem, a serem impressas localmente ou pelo ETS. Os testes serão empacotados alternadamente e dispostos sistematicamente, de tal forma que cada teste seja distribuído aleatoriamente entre os estudantes e que duas pessoas vizinhas dificilmente tenham o mesmo exemplar. As instruções de aplicação do teste serão as mesmas para os quatro conjuntos, de modo a possibilitar a aplicação dos quatro testes em um mesmo período.

AMOSTRAS DO TESTE-PILOTO

Diversos princípios orientaram as recomendações para a seleção de amostras dos testes-piloto.

1. O TIHD tenciona facilitar o deslocamento internacional de estudantes. As amostras devem, portanto, ser representativas da população universitária nos quatro grupos lingüísticos em questão.
2. O TIHD tenciona promover a ampliação das oportunidades educacionais, de forma que seria adequado basear-se numa definição não-restritiva da população alvo.
3. Um estudo de validade concorrente está incluído no projeto, assim uma vasta gama de habilidades desenvolvidas e desempenhos deve estar representada para evitar a atenuação dos coeficientes de validade, em decorrência da diminuição da amplitude das variáveis dependentes ou independentes.
4. As pesquisas têm revelado que os primeiros 5% da população estudantil, em muitos países, são bastante semelhantes em termos de aproveitamento (Tyler, 1981). Os dados dessas populações poderiam, pois, ser especialmente confiáveis para realizar os procedimentos de equivalência dos testes.

Recomenda-se um número de 800 sujeitos para compor a amostra do teste-piloto em cada língua, 400 representando estudantes do último ano do 2º grau, que tencionam ingressar em instituições de ensino superior, e 400 representando estudantes do último ano de graduação, com qualificação suficiente para ingressarem em cursos de pós-graduação. 100 sujeitos do 2º grau e 100 sujeitos da amostra universitária seriam submetidos a cada uma das formas do teste-piloto. Recomenda-se uma sobreamostragem de estudantes de alto grau de aproveitamento entre as populações alvo: – 20 a 25% da amostra total deveriam ser constituídos por este grupo. A sobreamostragem no ponto mais alto da escala poderá reduzir a necessidade de interpolações nos intervalos mais elevados de desempenho, caso se encontrem diferenças no desempenho das amostras (Angoff e Modu, 1973).

ANÁLISE DOS DADOS DO TESTE-PILOTO E MONTAGEM DOS CONJUNTOS DEFINITIVOS

Se os recursos disponíveis assim permitirem, dois métodos de equivalência das escalas de escores serão usados na análise dos dados recolhidos e para preparar as formas de "questões quase comuns". Planeja-se desenvolver as seguintes etapas:

1. Far-se-ão pequenos ajustes para equiparar os índices de dificuldade em cada grupo lingüístico a partir de diferentes aplicações, de modo a tratar todas as questões verbais (V) de cada língua como um único conjunto e todas as questões quantitativas (Q) de cada língua também como um único conjunto. Disto resultarão 8 conjuntos, 4 V e 4 Q.
2. Empregar-se-á a Fase I do método de equivalência de Angoff e Modu a fim de comparar o desempenho dos 4 grupos lingüísticos em todas as combinações binárias possíveis, a saber:

Av : Cv	(árabe, verbal: chinês, verbal)	Aq : Cq
Av : Iv		Aq : Iq
Av : Pv		Aq : Pq
Cv : Iv		Cq : Iq
Cv : Pv		Cq : Pq
Iv : Pv		Iq : Pq

Obter-se-ão dados relativos à dificuldade, poder discriminativo e adequação de cada questão a cada um dos grupos lingüísticos comparados.

3. Selecionar, com base em todos os dados disponíveis, dois conjuntos de questões, um V e um Q, como as melhores "questões quase comuns" a serem utilizadas em cada uma das quatro línguas como formas equivalentes definitivas. As melhores "questões quase comuns" serão aquelas mais próximas do eixo principal dos pontos Δ (delta) para o número máximo de conjuntos de pares de questões que também se situem entre limites especiais de dificuldade e de discriminação.

4. Retraduzir cada questão para a língua de origem de forma independente, por tradutores diferentes, e comparar o resultado com as versões originais para assegurar a adequação da tradução, sem perda significativa de sentido.

5. Reanalisar os dados originais, usando apenas "questões quase comuns", para ajustes em função de diferenças no desempenho das amostras, e permitir a equivalência das versões em quatro línguas pelos métodos linear, equipercentil e da teoria da resposta de item. Produzir tabelas de conversão de escores. Apresentar-se-ão coeficientes de validade concorrente.

6. No caso em que um número inesperadamente grande de questões se revelarem aceitáveis como "questões quase comuns", será possível elaborar duas versões definitivas dos testes – com algumas questões comuns –, mas sendo um teste mais difícil que o outro.

RELATÓRIO DOS RESULTADOS

Todos os resultados brutos, sua análise, bem como a equivalência dos resultados da forma final serão compartilhados entre as organizações colaboradoras. Cada qual receberá um conjunto da forma final do teste usada para fins de equivalência.

Sugerimos duas formas de expressar os escores. A primeira, para fins de interpretação transnacional de escores (como, por exemplo, para informar o escore equivalente obtido por um indivíduo, em uma língua, a uma instituição que opera com outra língua), deveria ser um escore padrão, de um dígito, tal como o estanino. Uma escala deste tipo desestimularia a sobreinterpretação e o excesso de dependência em relação ao escore como única fonte de informação para a tomada de decisão. Dado que os escores brutos serão disponíveis, não há nada que impeça que cada organização colaboradora desenvolva uma ou mais escalas locais de avaliação, que poderão ser expressas no tipo de unidade mais facilmente compreensível pelos utilizadores em potencial.

ETAPA OPCIONAL DE VALIDAÇÃO

Se as organizações colaboradoras assim o desejarem, estudos de validade preditiva das formas finais do teste poderão ser iniciados, aplicando-se o teste a grupos ingressantes no primeiro ano de graduação e de pós-graduação, e obtendo-se, longitudinalmente, informações sobre os critérios, por intermédio do exame de históricos escolares ou de outras fontes relacionadas com o desempenho acadêmico.

DESENVOLVIMENTO DE FUTUROS TESTES

Deverá ser constituída uma comissão para rever e revisar, sempre que necessário, as especificações do teste, e conceder uma aprovação geral às novas formas de testes à medida que fo-

rem sendo preparadas. Assim sendo, a comissão poderia ser composta de representantes das organizações colaboradoras envolvidas no desenvolvimento dos testes, possivelmente complementada por indivíduos vinculados a outras organizações associadas.

A organização local, com a responsabilidade por uma determinada língua, produziria, inicialmente, as questões naquela língua. Se possível, encarregar-se-ia, também, da tradução para as demais línguas; caso contrário, a coordenadoria central providenciaria as traduções necessárias. Via de regra, haveria uma retradução independente para a língua original, como primeira medida para determinar-se a adequação da questão. O objetivo seria produzir cada questão em todas as línguas em que os testes estão sendo desenvolvidos, e a comissão de testes seria responsável pela revisão e pela decisão final sobre a satisfatoriedade de uma determinada questão.

É particularmente desejável que todas as questões sejam pretestadas em todas as línguas, mas, em alguns casos, o juízo de especialistas pode ser suficiente. Desenvolver-se-á um procedimento que permita que as questões sejam pretestadas nas formas operacionais do teste sem que afetem o escore.

FORMAS EQUIVALENTES E ESCALAS

A fim de atender a seu propósito principal, o programa deverá apresentar os escores em uma escala comum, independentemente da língua em que o teste foi elaborado. Um dado escore deverá representar a mesma capacitação para o trabalho acadêmico, dada uma fluência adequada na língua de aprendizagem. Para atingir tal objetivo será necessária a equivalência dos testes em cada uma das línguas. Embora nenhum método possa produzir resultados precisos, existem várias aproximações que devem proporcionar informações suficientemente úteis.

Um dos objetivos do estudo de viabilidade é o de explorar algumas possibilidades a fim de formular alguns procedimentos desejáveis. Uma das principais atribuições das organizações será, portanto, a de rever as alternativas e os problemas levantados, e a de decidir sobre a melhor forma de tornar equivalentes os diferentes testes, nas diversas línguas. Haverá também a necessidade de equiparar os testes subseqüentes aos primeiros em cada língua, o que, no entanto, constituirá uma tarefa mais simples. Será necessária, contudo, uma verificação constante para que tanto a escala para uma dada língua quanto a escala interlingüística permaneçam sempre constantes. Um dos problemas refere-se à determinação da medida em que será possível pretestar amostras adequadas nas várias regiões; um outro refere-se ao fato de saber se questões comuns poderão ser utilizadas em testes sucessivos e ainda manter sua segurança.

A escala de estaninos apresenta uma série de vantagens para sua utilização neste programa. Variando apenas de um a nove, evita a impressão de extrema precisão, como seria o caso de uma escala de dois ou três dígitos. Seria, ainda, relativamente fácil de interpretar, e haveria pouca probabilidade de ser confundida com uma porcentagem de acertos, o que poderia levar a uma interpretação inadequada.

Além dos próprios escores escalares, várias normas deverão ser fornecidas. Uma vez definidos adequadamente os grupos de referência, o percentual pode tornar-se bastante significativo, e mais, ajudar a superar eventuais falhas na equivalência das formas. A auto-seleção dos estudantes para a realização do teste deverá variar substancialmente de região para região e poderá provocar problemas na interpretação dos resultados. Isto pode acarretar dificuldades, não apenas para as instituições que utilizam o teste para seleção translingüística e transregional, mas também em outros níveis. Um grupo poderá parecer inferior a outro, quando, na realidade, a diferença não resulta de diferenças regionais em habilidades, mas apenas no critério de auto-seleção dos candidatos. Deve-se cuidar para que o material interpretativo explique inteiramente a natureza da informação, a fim de evitar ou minimizar quaisquer possíveis comparações ofensivas.

Será provavelmente desejável ter normas combinadas para todos os grupos e em todas as línguas, mas resta examinar a sua possível utilidade.

VALIDADE

Será necessário investigar a validade do teste para vários grupos e em diferentes instituições. Apesar de ser sempre verdadeiro que a validade refere-se à utilização do teste e que ela não é uma propriedade intrínseca do teste, tal afirmação torna-se particularmente verdadeira no que diz respeito ao Teste Internacional de Habilidades Desenvolvidas. Diferentes versões lingüísticas podem produzir efeitos diferentes. Grupos com diferenças a nível de formação e cultura, no interior de um mesmo grupo lingüístico, podem ter desempenhos diferentes. E quaisquer previsões para instituições diferentes, com programas educacionais e culturais variáveis, podem apresentar diferenças. Por outro lado, é lícito esperar que haja certas generalizações de uma situação para outra e que evidências não completas poderiam ser suficientes para encorajar uma utilização mais ampla.

É possível que se venha a obter, a partir do estudo de viabilidade, evidências suficientes para iniciar as atividades operacionais, sem aguardar a realização de estudos adicionais de validação. Caso isto não ocorra, será necessário providenciar experimentos em estudantes ingressantes ou prestes a ingressar em cursos de graduação ou de pós-graduação. Estes experimentos podem ser realizados em várias regiões, cada um numa única versão lingüística, como podem também ser realizados onde se matriculem estudantes provenientes de diversas regiões lingüísticas. Há várias instituições nos EUA em que esta segunda possibilidade ocorre efetivamente. Contudo, evidências de instituições norte-americanas não seriam suficientes. Seria altamente desejável obter evidências da validade preditiva antes de se montar um programa operacional, com uma indicação da relação entre desempenho no teste e uma medida do êxito acadêmico pelo menos um ano mais tarde. Por outro lado, evidências adequadas de validade concorrente (relação entre escores no teste e em uma medida simultânea do êxito acadêmico) provavelmente bastariam para justificar o prosseguimento. A menos que haja alguma evidência de diferenças óbvias entre as medidas por critérios contemporâneos e por critérios a prazo mais longo, seria de espantar que se encontrassem grandes discrepâncias entre a validade concorrente e a preditiva.

Uma característica essencial do programa em andamento deveria ser alguma providência para a realização de estudos rotineiros de validade, mesmo que haja evidências suficientes de validade para dar início à operação do programa. Haverá numerosas questões sem resposta e, com o decorrer do tempo, será necessária a reconfirmação da validade. É bem possível que haja uma faixa de validades, em função das circunstâncias. É pouco provável que as diferentes versões lingüísticas tenham a mesma validade, mesmo em função de um critério comum. A natureza da língua, da cultura, dos programas educacionais e a auto-seleção do grupo podem provocar diferenças que devem constituir assuntos a serem investigados. O mesmo teste pode apresentar validades bastante diferentes para instituições diferentes em diversas partes do mundo. Novamente, a cultura, a língua e o sistema educacional podem originar tais diferenças. Em suma, será necessário um controle ininterrupto a fim de se saber onde o teste funciona e onde não funciona. A organização coordenadora local deverá assumir uma grande responsabilidade nestas investigações e trabalhar em conjunto com as instituições utilizadoras do teste em sua região, a fim de obter os dados necessários e realizar as análises devidas.

Deverá também haver análises contínuas dos testes e de cada questão. Estas deverão ser feitas por língua e por região, para determinar se há ou não vieses que estejam afetando um ou outro grupo e se há mudanças que possam corrigir tais vieses. Tudo isso, bem como a evidência proveniente dos estudos de validade, poderia realimentar o processo de desenvolvimento de futuros testes.

TESTE NA LÍNGUA DE APRENDIZAGEM

A habilidade básica, medida na língua nativa do estudante, pode muito bem constituir uma evidência insuficiente caso ele seja incapaz de lidar com a língua de aprendizagem. Assim, o

Teste de Habilidades Desenvolvidas precisa ser complementado por evidências referentes à competência na língua de ensino. Em muitos casos, tal evidência pode ser facilmente obtida de fontes já existentes (por exemplo, o TOEFL nos EUA). O programa, através de suas organizações colaboradoras locais, deverá assumir a responsabilidade de verificar *o que* pode ser obtido e certificar-se de que sejam tomadas medidas adequadas para sua utilização em conjunto com o TIHD.