FGV PROJETOS
ARTIGOS

CADERNOS FGV PROJETOS Nº 25:

CONCURSOS PÚBLICOS, AVALIAÇÕES E CERTIFICAÇÕES

ANÁLISE "TEORIA DE RESPOSTA AO ITEM" E "DISTRACTOR ANALYSIS"

Keith A. Rowland, diretor de sistemas da Prepona S.A.
Timothy J. Perry, diretor técnico da Prepona S.A.

A IMPORTÂNCIA DE ASSEGURAR A QUALIDADE DE AVALIAÇÃO EM PROCESSOS DE CERTIFICAÇÃO DE PESSOAS

Em processos de certificação de pessoas, que envolvem a avaliação de conhecimentos em grande escala e têm longa duração, o organismo responsável deve garantir não somente a integridade dos exames (especialmente contra tentativas de defraudá-los), mas também um campo nivelado (no sentido de "fair play"), a todos os candidatos, ao longo dos anos. Neste sentido, uma prova de múltipla escolha tradicional apresenta dificuldades, pois como é possível garantir que seja igual, ou suficientemente similar, à prova aplicada em outros anos? Ou seja, será que a "régua" usada é a mesma?

O fato é que há vários bons argumentos para o uso de um exame de múltipla escolha. Aprendemos na Prepona sobre a necessidade de se manter a mente aberta, a fim de escolher a melhor modalidade de avaliação de acordo com os objetivos dos organismos de certificação. Usar uma prova de múltipla escolha em uma das suas diversas variedades – inclusive *Computer Adaptive Testing* (CAT) –, uma prova discursiva ou até uma avaliação de execução de tarefas vai depender do volume de candidatos, dos recursos disponíveis (se existem examinadores suficientes, por exemplo) e dos objetivos do processo de certificação.

Voltando para o uso de provas de múltipla escolha, é preciso considerar, então, não necessariamente o valor intrínseco desse tipo de avaliação, mas – partindo do pressuposto de que a múltipla escolha é, sim, apropriada para os desideratos de um organismo de certificação – avaliar o que pode ser feito para assegurar a qualidade da "régua" a ser utilizada. Uma das maneiras para atingir esse objetivo é submeter a um processo de "validação" todos os itens que se queira incluir no banco de itens de uma prova. Esse processo envolve uma análise estatística que, no ramo de testagem, se chama calibração. E o método mais eficaz se baseia na Teoria de Resposta ao Item (TRI).

TEORIA DE RESPOSTA AO ITEM

A TRI representa matematicamente a interface entre um candidato e o item. Tem suas raízes nas ideias de Loevinger,[1] quando afirma que todos os itens numa prova deverão medir a mesma coisa ou o mesmo traço latente. A TRI formaliza isso de uma maneira explícita, assumindo uma única dimensão de conhecimento ou habilidade de que todos os itens da prova dependem para ser respondidos corretamente. Exemplos dessas características incluem:

- Competência linguística;
- Habilidade matemática; e
- Raciocínio lógico.

A posição que cada item ocupa nessa dimensão é chamada de **dificuldade** do item e é denominada pelo parâmetro **b**.

A posição de cada candidato nesta dimensão, denominada como sua **competência** ou **habilidade**, costuma ser posicionada na escala denominada θ.

O modelo TRI dá a probabilidade de um candidato de nível de competência θ responder corretamente a um item de dificuldade **b**. Na sua forma mais simples, a TRI combina somente essas duas variáveis, e, já que caractacteriza o item com um só parâmetro (a dificuldade **b**), leva o nome de Modelo Logístico Unidimensional de 1 parâmetro (ML1). Este modelo foi desenvolvido em 1960[2] por Georg Rasch, e, por isso, leva o seu nome.

O ML1 é representado:

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

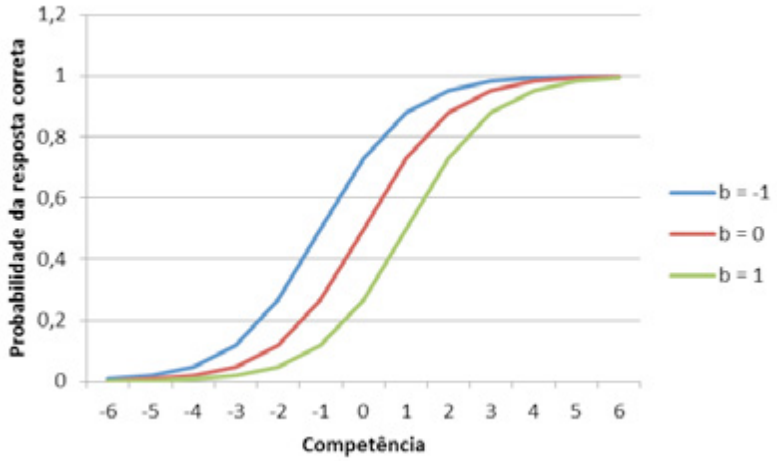
em que $P(\theta)$ é a probabilidade de um candidato com nível de competência θ responder corretamente a um item de dificuldade **b**.

A figura a seguir mostra graficamente a estrutura deste modelo para três itens de dificuldades diferentes.

[1] LOEVINGER, J. A systematic approach to the construction of and evaluation of tests of ability, Psychological Monographs, 61, 4 – Uma demonstração de que todos os itens num teste deveriam medir o mesmo característico – ou seja, o teste deverá ser homogêneo.

[2] RASCH, GEORG. Probabilistic models for some intelligence and attainment tests. Copenhagen: Denmark Paedagogiske Institut.

Gráfico 1



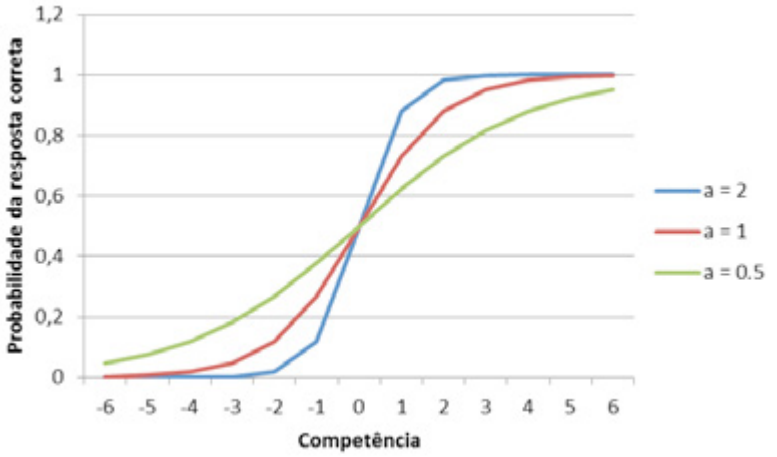
Esses gráficos chamam-se **Curvas Características do Item (CCI).**

Percebe-se que, durante a maior parte da curva, as CCI dos itens ficam mais ou menos paralelas. Infelizmente, essa aproximação faz com que o modelo falhe em muitos casos, pois o comportamento das CCI não está bem descrito. Nesses casos, temos duas alternativas: podemos retirar itens com comportamento divergente do banco de itens ou podemos generalizar o modelo para acomodar inclinações diferentes. Isso se faz através da inclusão de um segundo parâmetro para cada item. Este parâmetro, chamado **a**, caracteriza a inclinação da CCI, e mede a **discriminação** do item. O modelo matemático resultante, chamado ML2, é agora representado:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

Outra vez, uma representação gráfica ajuda a esclarecer. O gráfico a seguir mostra as CCI para três itens com o mesmo valor de **b** – um item discrimina muito bem (a = 2), outro está mais para a média (a = 1), e o último tem discriminação fraca (a = 0,5):

Gráfico 2



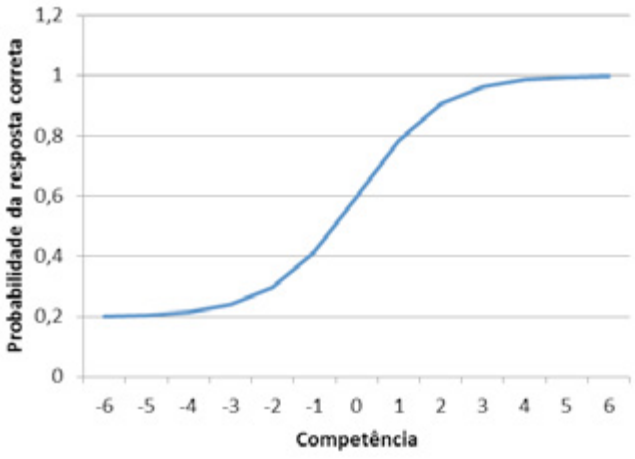
O acréscimo desse parâmetro aumentou muito a aplicabilidade da TRI. Porém, ainda há um fator a considerar: para testes do tipo múltipla escolha sempre existe a possibilidade significativa de um candidato acertar um item "no chute". Nenhum dos modelos apresentados acima considera esta possibilidade: há dois motivos para que um candidato acerte um item cuja dificuldade é muito além da sua competência – o teste não é unidimensional, e o item foi respondido utilizando

conhecimentos de outra base de conhecimento que não seja aquela que estamos testando ou o candidato tenha "chutado". Neste caso, podemos eliminar estes itens, mas é pouco provável que este tipo de acerto funcione, porque candidatos diferentes acertariam itens diferentes e acabaríamos retirando todos os itens difíceis da prova! Uma segunda solução é generalizar o modelo para acomodar "chutes". O modelo resultante, descrito no texto de Lord e Novick,[3] é chamado Modelo Logístico Unidimensional de 3 parâmetros (ML3), e é descrito pela equação a seguir:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}}$$

Mais uma vez, a estrutura do ML3 pode ser mais bem compreendida graficamente:

Gráfico 3



O ML3 é o modelo TRI mais usado em testagens de grande escala. Embora o parâmetro de chute (**c**) seja raramente necessário no contexto de um teste CAT - porque, se o teste estiver funcionando corretamente, candidatos raramente encontrarão itens que são muito difíceis para eles -, ele é necessário durante o processo de calibração e na fase de testagem inicial.

Enfim, se podemos validar os itens de uma prova usando TRI, teremos várias vantagens, não somente para a construção e aplicação de um teste CAT, mas, também, para o uso de uma prova de múltipla escolha de proporção correta. Sabendo os valores dos itens, teremos uma régua confiável para medir a competência ou habilidades dos candidatos.

ESTIMATIVA DE COMPETÊNCIA OU HABILIDADE

De posse de um banco de itens com os parâmetros a, b e c devidamente calculados, aplicamos o mesmo aos nossos candidatos, e calculamos o nível de competência (θ) utilizando o método de máxima verossimilhança.

Consideremos:

x_i a ser o vetor de respostas do candidato i, onde o valor é 1 se for correto e 0 se for errado. Os elementos deste vetor são dados por $\{x_{ij}\}$, onde os itens são indexados por j

β_j é o vetor dos parâmetros (a_j b_j c_j) do item j

$Q(\theta) = 1 - P(\theta)$

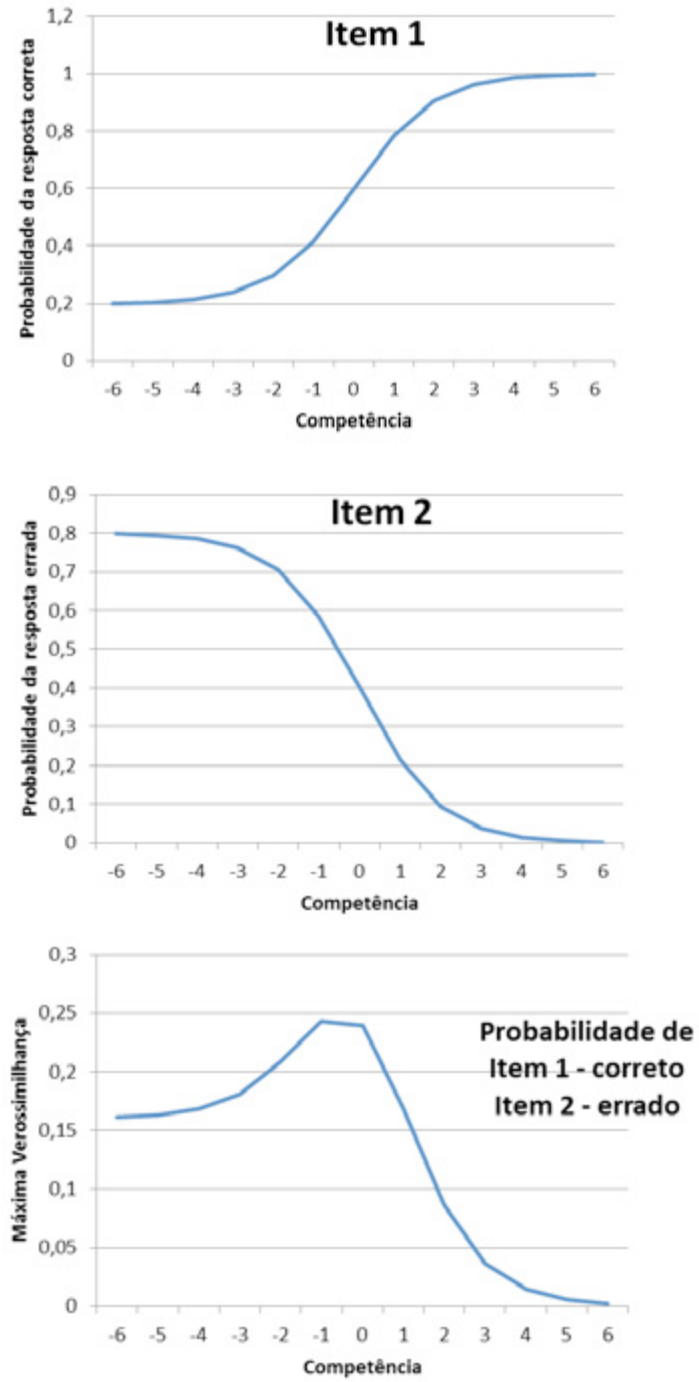
A probabilidade condicional de x_i dado θ e β é:

$$P(x_i \mid \theta_i, \beta) = \Pi\, P_j(\theta_i)^{x_{ij}}\, Q_j(\theta_i)^{1 - x_{ij}}$$

[3] *LORD, F. & NOVICK, M. Statistical Theories of Mental Test Scores. Reading. MA Addison-Wesley (1968).*

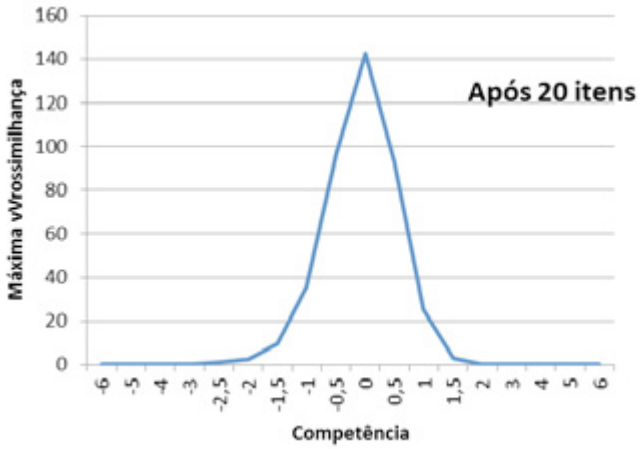
Esta equação representa o produto do candidato com nível de competência acertar os itens que acertou, e errar os itens que errou para todos os itens. Outra vez, uma representação gráfica ajuda a esclarecer. Consideremos um teste de dois itens, em que o candidato acerta o primeiro item e erra o segundo:

Gráfico 4



Para um candidato, portanto, o nível de competência ou máxima verossimilhança é representado pelo máximo no gráfico de probabilidade. É claro que para um teste de dois itens, a distribução do gráfico é muito larga. Porém, se aplicarmos mais itens, perceberemos que a estimativa começa a afunilar. No exemplo a seguir, o candidato acertou 12 e errou 8 itens num teste de 20 itens. Esse afunilamento permite o uso da tecnologia CAT para a aplicação de provas mais “inteligentes”.

Gráfico 5



TRI E TESTES CAT

Um teste CAT, por exemplo, usa um algoritmo que sempre procura aplicar os itens que fornecem mais informação a respeito do candidato. Se, por exemplo, aplicarmos uma série de itens muito fáceis, o candidato acertará todos, e não saberemos nada sobre ele. O mesmo ocorre se aplicarmos uma série de itens difíceis demais, pois o único recurso do candidato será "chutar". O CAT escolhe os itens com a meta de afunilar o gráfico ao máximo, e, sendo assim, permite uma avaliação mais exata com menos itens.

Esse fator permite concentrar a aplicação, de maneira gradativa, de itens cada vez mais aproximados ao verdadeiro nível de competência do candidato, eliminando, progressivamente, a necessidade de aplicar itens fáceis ou difíceis demais. A grande vantagem disso para o candidato é:

- Ter a percepção de que a prova foi "customizada" para ele, o que provoca a diminuição de estresse (causado ao enfrentar itens difíceis demais) e/ou tédio (causado ao enfrentar itens fáceis demais); e
- Expor-se a um total menor de itens, que expressam o seu verdadeiro nível de competência, assegurando assim um resultado mais preciso, e até mesmo reduzindo o tempo total do processo, o que também reduz o estresse.

Antes da adoção do CAT, a Prepona aplicava provas de múltipla escolha – proporção correta – nas suas avaliações de inglês. Considerando que este tipo de prova deve tentar refletir a realidade do ensino, é preciso avaliar em qual nível uma pessoa está escolhendo entre 10 possíveis níveis. Para isso, são aplicadas provas de múltipla escolha, proporção correta, contendo 120 itens. Isso permitiu a aplicação máxima de somente **12** itens por nível de competência.

Com a introdução do CAT, foi possível "acertar" o nível de um candidato com a aplicação de muito menos itens, mas, ao mesmo tempo, com a aplicação de mais itens ao redor do nível de competência do candidato, gerando um resultado mais preciso. Em resumo, a maior precisão do resultado final, com o nivelamento de competência dos candidatos, e a redução do estresse são dois dos fatores principais na adoção da metodologia CAT em provas de certificação de pessoas.

O fato de ter um banco de itens calibrados, usando TRI e a aplicação através do CAT, ajuda o organismo de certificação a manter a integridade do conteúdo das provas. Se a escolha do próximo item na prova depender da escolha da resposta do(s) item(ns) anterior(es), um candidato "fraco", por exemplo, não vai ter os mesmos itens aplicados em um candidato mais "forte", e vice-versa. Além disso, gera a sensação entre os candidatos de terem feito uma prova "customizada" para cada um deles.

O outro grande benefício, ou vantagem, é que, com itens devidamente calibrados, usando TRI, o organismo de certificação pode manter a consistência (a qualidade) da sua "régua". Se, por

exemplo, um item sofre de "over exposure" (ou seja, se for escolhido repetidas vezes pelo algoritmo do CAT), pode ser substituído por outro item com os valores psicométricos (a, b e c) mais próximos. Ou seja, em vez de substituir um item escolhendo outro qualquer, podemos escolher um item que tem grau de dificuldade similar, com o mesmo poder de discriminação e com a mesma robustez contra "chutes". Assim, o ato de substituição de itens não deverá afetar a consistência da "régua".

Essas vantagens, oriundas de um processo de análise TRI, não se aplicam somente às provas CAT, mas também trazem benefícios significativos ao desejo de assegurar, ou melhorar, a qualidade de provas de múltipla escolha de proporção correta.

Na sua aplicação, usando itens calibrados e um sistema computadorizado para a escolha aleatória de itens através de parâmetros preestabelecidos, pode-se aumentar muito as combinações possíveis, sem detrimento da qualidade da prova como um todo e com o benefício de dificultar eventuais tentativas de defraudar o processo de certificação. Somente não podemos diminuir o estresse do candidato, pois esse tipo de prova exige a aplicação de um número fixo de itens em um tempo determinado e com a abrangência de todos os níveis.

Mas, se o objetivo do organismo de certificação for avaliar áreas diferentes na mesma prova, por exemplo, ou mostrar por meio do resultado em que área um candidato é forte ou fraco – o que é de grande utilidade para orientação de seus estudos futuros –, o uso de itens calibrados com TRI se mostra excelente para assegurar a qualidade do processo. Também será possível retirar ou substituir itens com a mesma facilidade e confiança que é possível num teste CAT.

Há ainda mais uma vantagem: combinar a análise TRI com a metodologia CAT permite aproveitar as provas para calibrar itens novos a serem incluídos no banco de itens para uso futuro. Alguns deles são "itens sementes"; não influenciam o processo de avaliação e constam para fins de calibração. Ou seja, podemos usar os candidatos verdadeiros para calibrar itens novos.

CALIBRAÇÃO DOS ITENS

Como sabemos, cada item é definido por três parâmetros:

- A discriminação (**a**) – mede o poder que o item tem para diferenciar os candidatos que sabem mais daqueles que sabem menos;
- O grau de dificuldade (**b**); e
- O fator de "chute" (**c**) – leva em conta que um candidato fraco pode acertar um item difícil no "chute."

No início de um processo de testagem, não podemos nem estimar esses valores; então é necessário coletar dados de duas formas:

1. Aplicando provas simuladas em candidatos futuros; e/ou
2. Executando a análise TRI em provas antigas.

PRÉ-CALIBRAÇÃO EM PROVAS SIMULADAS

De posse do banco inicial de itens, formam-se diversas provas em que as versões diferentes têm alguns itens em comum. Consideremos, como um exemplo simples, um banco com 250 itens: poderíamos dividi-lo em dez pacotes com 25 itens e, em seguida, construir dez provas de 50 itens cada a seguir:

Prova 1 = Pacote 1 + Pacote 2
Prova 2 = Pacote 2 + Pacote 3

(...)

Prova 9 = Pacote 9 + Pacote 10
Prova 10 = Pacote 10 + Pacote 1

Sistemas mais complexos podem ser utilizados, se necessário, mas é importante que sejam aplicados a candidatos do mesmo nível daqueles que farão o teste de verdade.

CALIBRAÇÃO

O processo de estimativa dos parâmetros é computacionalmente intensivo. Utilizamos um programa, desenvolvido exclusivamente para esta finalidade, que usa uma abordagem algorítmica:

Fase 1 : Estimativas iniciais são calculadas baseadas em transformações de estatísticas clássicas.
Fase 2 : Estas estimativas são sintonizadas utilizando o algoritmo de expectativa-maximização **(EM).**[4] O ciclo EM se repete até os parâmetros permanecerem constantes. Caso algum item não convirja, o sistema alerta e o item deverá ser avaliado.

Obs.: se um organismo de certificação tiver guardado exames anteriores e as respostas dos candidatos, podemos executar um processo de calibração, acelerando a geração de itens chave calibrados e prontos para uso.

A CONSTRUÇÃO DE UMA PROVA DE MÚLTIPLA ESCOLHA DE PROPORÇÃO CORRETA

Com os valores a, b e c calculados, sabemos quais itens são os melhores para serem incluídos na prova final. A prova costuma ter um perfil que poderá ser da seguinte forma:

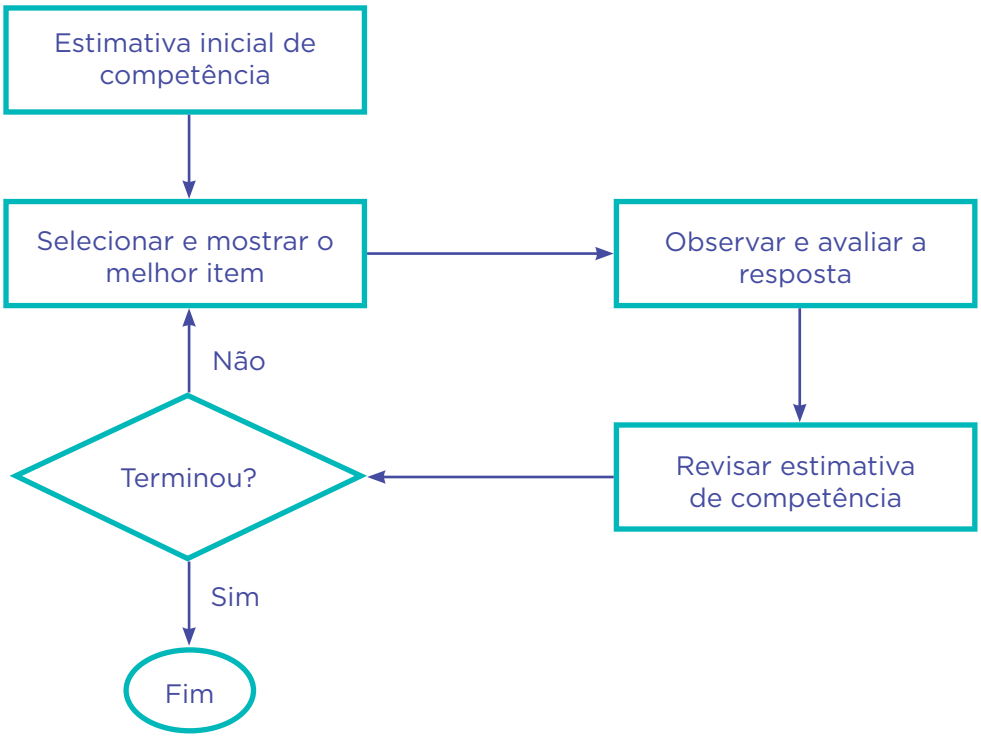
Tema 1 – Fáceis - x_1 itens, Médias – y_1 itens, Difíceis – z_1 itens
Tema 2 – Fáceis - x_2 itens, Médias – y_2 itens, Difíceis – z_2 itens
Tema 3 – Fáceis - x_3 itens, Médias – y_3 itens, Difíceis – z_3 itens
etc.

Obs. 1: O sistema poderá escolher itens aleatoriamente dentre aqueles classificados em cada faixa de dificuldade/tema.

Obs. 2: Poderá haver mais de três faixas de dificuldade.

A CONSTRUÇÃO DE UMA PROVA CAT

Nesse caso, todos os itens "aprovados" são colocados no banco. O primeiro item é escolhido de acordo com um critério preestabelecido, e os itens seguintes de acordo com todas as respostas do candidato aos itens anteriores. A lógica do teste pode ser vista no fluxograma a seguir:



[4] *DEMPSTER, A., LAIRD, N. & RUBIN, D. «Maximum Likelihood from Incomplete Data via the EM Algorithm». Journal of the Royal Statistical Society,Series B (Methodological) 39 (1): 1–38 (1977).*

Enfim, a análise TRI é uma função vital ao processo de certificação de pessoas, pois permite assegurar ao mercado a qualidade exigida, não importando se o organismo opta por aplicar exames usando a metodologia CAT ou provas computadorizadas de múltipla escolha de proporção correta.

Vimos que o processo de calibração deve ser executado antes do lançamento da prova, mas uma vez que o banco de itens estiver calibrado e a prova sendo aplicada, ainda existe a possibilidade de calibrar itens novos a serem incluídos no banco para uso futuro.

A vantagem disso se reflete principalmente na diminuição da produção de conteúdo novo, pois podemos utilizar ao máximo o banco existente com o mínimo de manutenção e o máximo de segurança.

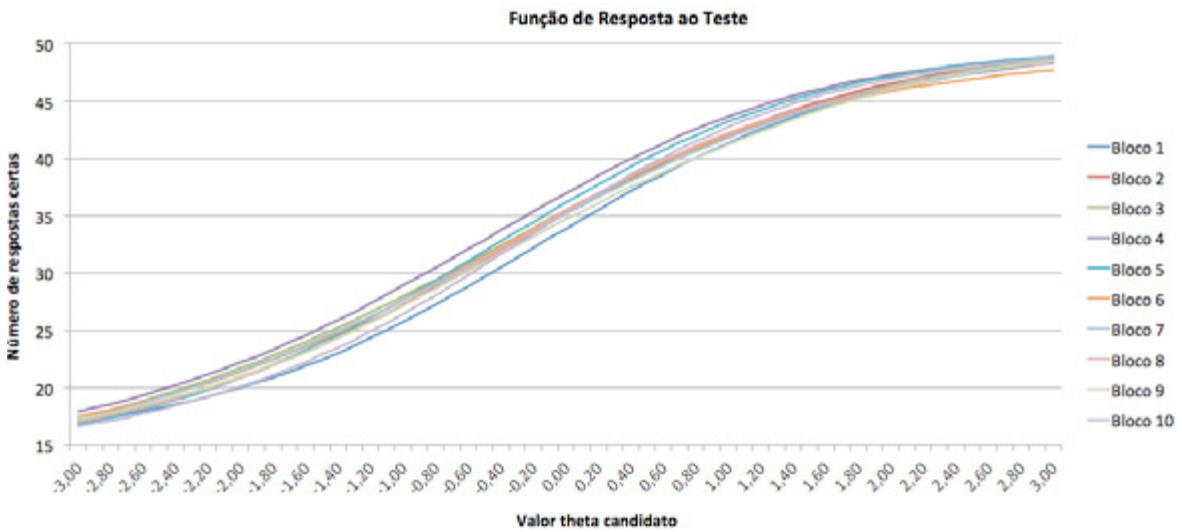
A análise TRI também oferece a oportunidade de avaliar diferentes bancos de itens do mesmo exame ou comparar um exame feito por um cliente *versus* os exames feitos por outros candidatos no mesmo processo. Isso se chama Função de Resposta ao Teste (*Test Response Function*), e a Prepona pode usá-la para ajudar seus clientes a verificar se tem fundamento um recurso com reclamação sobre o nível de dificuldade maior na prova de um candidato.

FUNÇÃO DE RESPOSTA AO TESTE PARA EQUIVALÊNCIA DE BLOCOS

Alguns clientes preferem o uso de vários blocos de itens na aplicação dos seus exames de certificação, e escolhem essa opção para manter maior controle sobre o que está sendo aplicado. A TRI é importante, na primeira instância, para avaliar se existe uma distribuição homogênea de itens de graus de dificuldade similares em todos os blocos. Além de usar a TRI, então, para fazer uma distribuição justa entre os blocos e assegurar a qualidade e consistência da "régua", podemos avaliar como se comparam ao receber de volta as respostas dadas pelos candidatos. Com esses dados, é possível demonstrar, em forma de gráfico, o agrupamento dos blocos e analisar se algum fica fora do esperado.

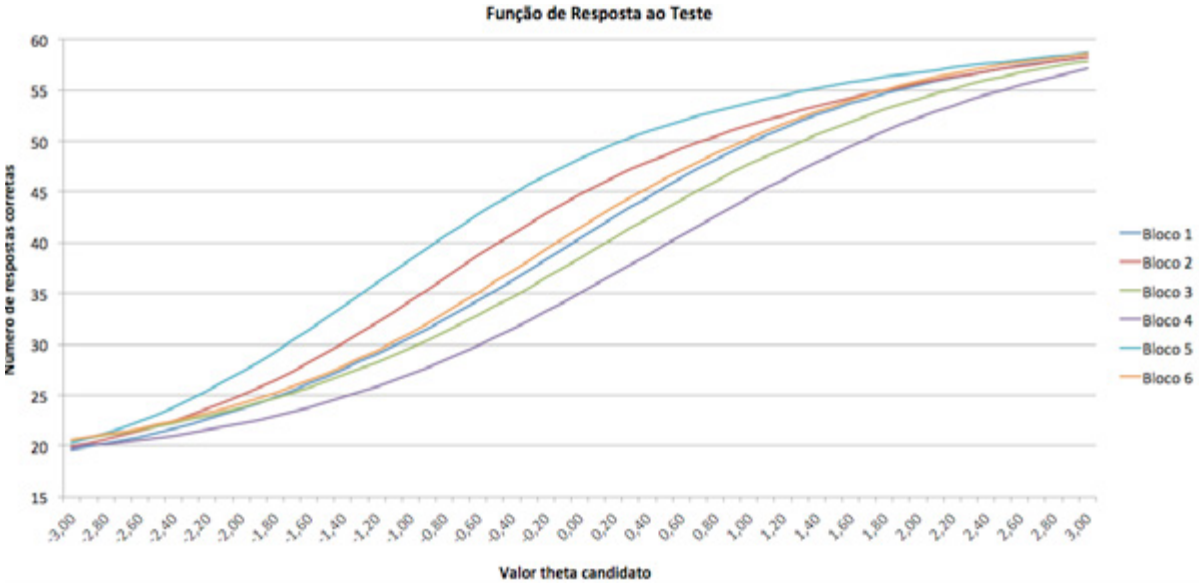
O gráfico abaixo mostra como são bem agrupados os dez blocos usados por um cliente no mesmo exame. Nesse exame, é necessário, para ser aprovado, responder corretamente a 35 dos 50 itens (ou 70%).

Gráfico 6



Como se pode ver, as curvas que representam os blocos são bem próximas umas das outras, especialmente no patamar de 35 respostas certas. É possível concluir que, em termos práticos, nenhum candidato está sendo prejudicado por ter recebido um bloco mais difícil do que os outros. Se passarmos a avaliar os blocos de itens em outro exame, podemos ver que talvez exista uma base para reclamação, e, consequentemente, uma necessidade de tomar providências corretivas.

Gráfico 7



Além de perceber que, como um todo, o agrupamento é um pouco mais disperso, existe um bloco que está bem fora da curva. Nesse caso, recomendamos que esse bloco receba uma análise mais cuidadosa dos itens contidos nele.

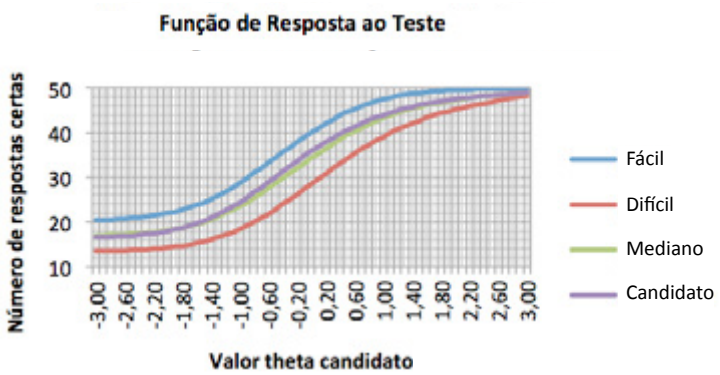
Uma variação dessa análise pode ser usada em outro tipo de aplicação de exames de múltipla escolha – o tipo de teste em que o cliente não quer utilizar blocos, mas selecionar itens randomicamente de um banco principal.

FUNÇÃO DE RESPOSTA AO TESTE PARA EQUIVALÊNCIA DE EXAMES

Outros clientes preferem que o sistema use os parâmetros estabelecidos para gerar um número quase ilimitado de testes do mesmo banco de itens. A TRI é muito importante nesse caso, especialmente caso algum candidato entre com um recurso alegando que seu exame foi mais difícil do que o dos outros candidatos. Usando o sistema, podemos gerar a combinação de itens para produzir o exame mais fácil daquele banco de itens. E fazer o mesmo processo para gerar o exame mais difícil possível. Assim, estabelecemos as duas extremidades, além de podermos estabelecer o mediano. Isto feito, podemos processar os itens respondidos pelo candidato para, então, comparar o "grau de dificuldade" do exame daquele candidato com o mediano e as duas extremidades.

No caso a seguir, o gráfico mostra que, pela maior parte do exame, o candidato ficou no lado mais fácil (mas bem próximo) do mediano, então não há fundamento para reclamar.

Gráfico 8



Esse tipo de análise torna-se parte do arsenal de dados e ferramentas de análise com que os clientes, como organismos de certificação, podem contar para assegurar a qualidade dos seus exames, tendo como base a análise da TRI.

Mas, além da metodologia TRI, a Prepona executa em tempo real outra análise que visa analisar o desempenho das respostas incorretas em cada item – ou seja, uma análise no nível micro ou dentro de cada item.

Sabemos que a parte mais difícil na redação de um item é a criação dessas respostas. Chamamos essas respostas de "distraidores", pois devem atuar para atrair os candidatos que não têm o grau de competência necessário para escolher a resposta certa.

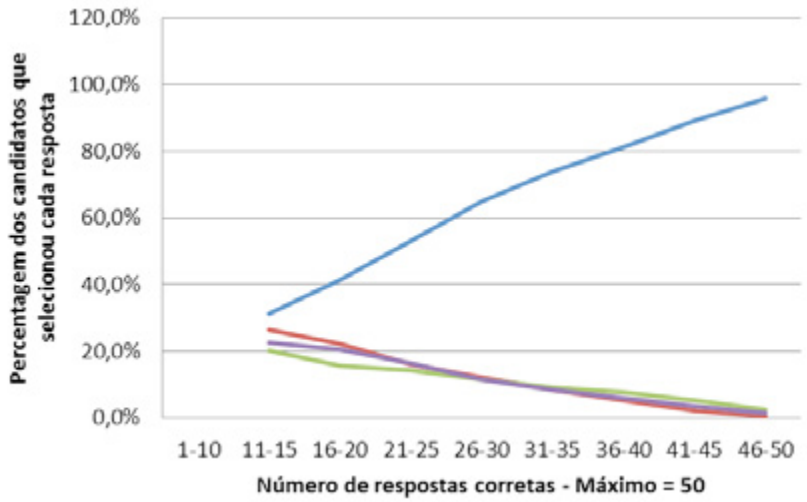
ANÁLISE DE "DISTRAIDORES" (*DISTRACTOR ANALYSIS*)

Embora os itens selecionados para compor o teste já tenham sido submetidos a uma avaliação estatística para determinar os que melhor avaliam os candidatos, é interessante avaliá-los em maior detalhe para ver como podem ser melhorados.

Como já foi explicado, o comportamento ideal de um item é representado pelo Modelo Logístico Unidimensional de 3 parâmetros. Porém, é importante também considerar as alternativas erradas e o seu comportamento. O ideal é que todas as respostas erradas atraiam um número significativo de candidatos fracos e números menores, conforme a competência dos candidatos sobe.

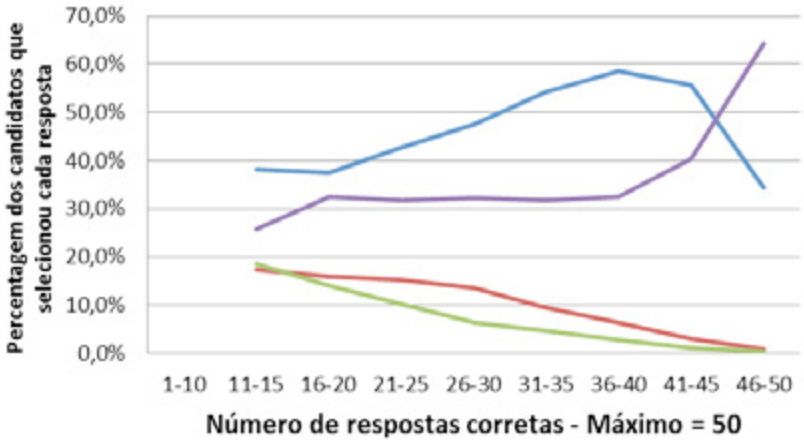
A Prepona analisa os itens respondidos pelos candidatos para verificar o seu comportamento, mostrando possíveis falhas nos itens que podem ser melhorados. O gráfico a seguir é um exemplo de um item quase ideal. A linha correspondente à resposta correta (linha azul) sobe bruscamente enquanto todas as respostas erradas caem regularmente, chegando a quase zero. Os candidatos mais fracos pareciam estar totalmente perdidos, pois cada resposta errada foi escolhida por pelo menos 20% deles.

Gráfico 9



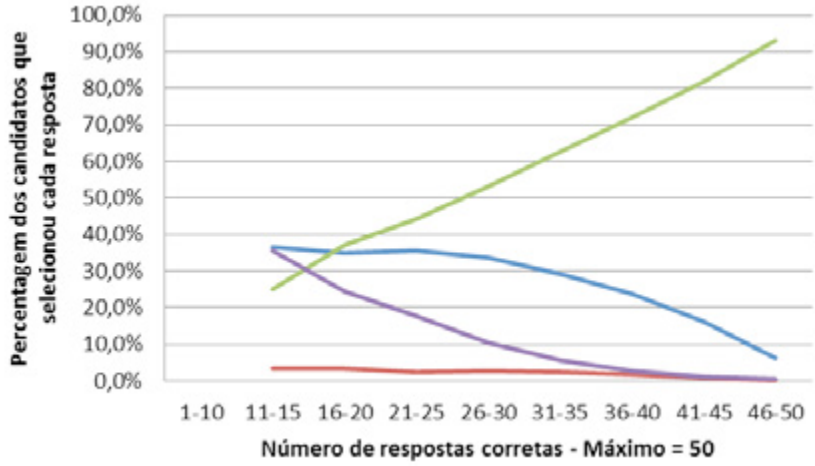
O item a seguir, porém, deveria ser examinado e modificado, pois apresenta uma resposta errada (representada pela linha azul) com sérios problemas. Ela está atraindo cada vez mais candidatos de competência mediana, conforme o nível de competência sobe. A resposta certa só começa a subir entre os candidatos muito bons. Em resumo, essa resposta errada está "confundindo" demais.

Gráfico 10



O item a seguir mostra outro comportamento:

Gráfico 11



A resposta certa (linha verde) se comporta bem, duas das respostas erradas funcionam bem (linha roxa) e satisfatoriamente (linha azul). A resposta errada, representada pela linha vermelha, porém, não atrai quase ninguém. Uma resposta obviamente errada até para os candidatos mais fracos não contribui em nada e deverá ser removida ou substituída.

Em suma, a análise de "distraidores" permite à Prepona ajudar seus parceiros a aprimorar, mais ainda, a qualidade do conteúdo das provas de certificação de pessoas.

Todas essas técnicas visam assegurar a qualidade da "régua" a ser utilizada, para tornar o processo de avaliação justo, eficiente e, acima de tudo, preciso.