

TEORIA DA RESPOSTA AO ITEM: VANTAGENS E OPORTUNIDADES PARA A ENGENHARIA DE PRODUÇÃO

Rafael Tezza (UFSC)

rafaeltezza@yahoo.com.br

Antonio Cezar Bornia (UFSC)

cezar@inf.ufsc.br



Neste artigo, são apresentadas as definições básicas e fundamentais da Teoria da Resposta ao Item (TRI), que é uma abordagem utilizada para o desenvolvimento, avaliação e administração de medidas padronizadas, muito utilizada nas áreas de psicologia e educação. O objetivo geral do artigo é apresentar as vantagens e oportunidades da TRI no contexto de engenharia de produção, levantando um paralelo com o método tradicional da Teoria Clássica da Medida (TCM). O foco na precisão das medidas em engenharia de produção se dá pelo fato de que neste mercado globalizado, há necessidade de tomar decisões com informações cada vez mais acuradas. Para exemplificar a viabilidade e as vantagens da TRI no contexto empresarial, é apresentada uma aplicação da TRI no mercado de e-business, criando uma escala padronizada para medir o grau de usabilidade de sites de e-commerce. Baseado na fundamentação teórica e no estudo de caso conclui-se que a TRI pode ser uma valiosa ferramenta de gestão, capaz de gerar informações quantitativas, precisas e robustas.

Palavras-chaves: engenharia de produção, TRI, TCM, ergonomia de interface

1. Introdução

A identificação de problemas, a construção de entendimento acerca destes e a verificação de oportunidades de melhoria, na maioria das vezes, depende de pesquisas aprofundadas que requerem um bom levantamento de dados. Sendo assim, todo este esforço, bem como os resultados da tomada de decisão, dependem do método de análise escolhido, sua adequação ao contexto e à confiabilidade das informações geradas por este. Segundo Drucker (1995), as informações geradas neste processo tornam possíveis diagnósticos, estratégias e decisões. Por isso, deve-se atentar para a relevância destas informações, bem como sua confiabilidade.

Escalas de medida são instrumentos de análise bastante utilizados na engenharia de produção para gerar entendimento e identificar oportunidades de melhoria. Segundo Gil (2002), escalas são instrumentos que objetivam medir traços latentes, intensidade de opiniões ou atitudes da maneira mais objetiva possível, utilizando séries graduadas e conjuntos de itens.

Praticamente em todas as áreas da engenharia de produção, são utilizadas escalas de medida, seja para controlar indicadores de desempenho, medir clima organizacional, identificar fatores críticos na gestão de intangíveis, em questões ambientais, da qualidade, maturidade organizacional, aceitação da inovação, satisfação de clientes ou parceiros etc. Estas escalas permitem, quando bem elaboradas, constatar eventuais discrepâncias no processo, identificar oportunidades de melhoria e poder de comparabilidade (benchmarking), entre outros.

Todavia, a forma tradicional de construir e analisar estes instrumentos (Teoria Clássica da Medida) restringe sua validade ao contexto em que este foi elaborado, ou seja, os resultados dependem da amostra e as conclusões são baseadas no escore total do teste (PASQUALI, PRIMI, 2003; EMBRESTON, REISE, 2000, SINGH, 2004).

Dentro deste cenário, surge uma teoria paralela para construção de escalas padronizadas que independe da amostra e do escore total – a Teoria da Resposta ao Item. Esta abordagem tem sido recentemente explorada nas mais diversas áreas do conhecimento tais como educação, psicologia.

Este artigo tem como objetivo apresentar os fundamentos básicos da Teoria da Resposta ao Item assim como suas vantagens e oportunidades dentro da engenharia de produção. A pesquisa classifica-se como aplicada exploratória e quantitativa. É utilizada a pesquisa bibliográfica como técnica de pesquisa e o estudo de caso como procedimento técnico. O artigo está estruturado em cinco tópicos, no primeiro é apresentada uma introdução do assunto, no segundo a definição de Teoria Clássica da Medida, no terceiro as definições da Teoria da Resposta ao Item, no tópico 4 é apresentado um estudo de caso realizado em empresas de e-commerce. Por fim, no tópico 5 são apresentadas as conclusões seguidas as referências.

2. Teoria Clássica de Medida

O ato ou ação de gerenciar um processo ou uma organização requer um conhecimento sobre os elementos envolvidos neste. Uma abordagem bastante utilizada para se conhecer, controlar e melhorar processo ou organizações é a criação de escalas padronizadas de medidas, pois, segundo Kaplan e Norton (1997), o que não se pode medir não se pode gerenciar.

No âmbito industrial ou comercial, grande parte das medidas é indireta. Dentro deste contexto, Churchill (1979) afirma que todo processo de mensuração envolve regras para atribuir números aos fenômenos, de forma a representar as quantidades de seus atributos.

Entretanto, para Nunnally (1978), o termo atributo indica que não se medem objetos diretamente e sim suas características. Portanto, a medição requer um processo de abstração. Este processo de abstração pode ser realizado a partir da elaboração de um conjunto de itens correlacionados com a característica que se objetiva medir. Sendo assim, o que se deseja medir é um traço latente, ou seja, uma variável que não pode ser medida diretamente, como por exemplo, maturidade organizacional, resistência a mudanças, satisfação do consumidor, valor dos intangíveis de uma empresa, grau de usabilidade de um produto, entre outros.

A Teoria Clássica da Medida (TCM) é o modelo mais usual para medição de traços latentes em ciências sociais, psicologia, educação e áreas afins (EMBRETSON, REISE, 2000, PASQUALI, 2003; REISE, ET AL, 2005). Segundo Hayes (1992), a equação básica da TCM, conhecida como modelo clássico de medida, descreve a relação entre os escores observados, os escores verdadeiros e o erro:

$$X = T + E \quad (\text{equação 1})$$

Onde: X é o escore observado; T, o escore verdadeiro; e E, o erro de medida.

Dentro da TCM, Weiss (1983) afirma que o escore verdadeiro (T) representa o traço latente. O erro (E), segundo Hambleton et al. (1991), pode estar associado aos seguintes fatores:

- Impossibilidade de incluir um número infinito de itens no conjunto de itens;
- Impossibilidade de aplicar o conjunto de itens um número infinito de vezes.
- Impossibilidade de submeter o conjunto de itens a um número infinito de organizações.

Na TCM, informações do item e suas correlações, bem como a confiabilidade dos resultados (consistência interna da escala), validade e erros-padrões, são dependentes das características particulares da população. Sendo assim, o escore observado (X) é definido por um conjunto particular de itens em uma simples medida (REISE ET AL., 2005).

Esta dependência direta dos resultados do teste com a amostra e o instrumento configura uma preocupação do ponto de vista da validade do modelo gerado. Thurstone (1928, p.547) já atentava para esta limitação: “um instrumento de medida, na sua função de medir, não pode ser seriamente afetado pelo objeto de medida. Na extensão em que sua função de medir for assim afetada, a validade do instrumento é prejudicada ou limitada.” Pasquali e Primi (2003) afirmam que, na elaboração de um instrumento de medida utilizando-se a TCM, os parâmetros dos itens dependem da amostra em que eles foram calculados. Assim, um item qualquer do conjunto se torna mais difícil ou mais fácil dependendo da amostra e, portanto, o parâmetro de dificuldade do item vai variar de pesquisa para pesquisa em função da amostra. Um exemplo desta limitação pode ser verificado na avaliação de intangíveis de uma empresa (VARGAS ET AL., 2008), se o instrumento não for independente da amostra, um conjunto qualquer de itens que avalie, por exemplo, o grau de desenvolvimento do capital humano será considerado muito fácil se a empresa possuir uma estrutura de treinamento e muito difícil se a empresa não possuir. Todavia, este problema configura mais um problema de amostragem do que de análise. Segundo Pasquali (1996), um problema mais grave ocorre com o cálculo do parâmetro de discriminação do item. Na TCM, a estimação deste parâmetro é feita baseada no escore total de um teste, seja utilizando grupos-critérios ou coeficientes de correlação. Sendo assim, ao se utilizar o escore total, supõe-se que todos os itens sejam adequados, o que nem sempre é verdade. Hambleton e Swaminathan (1985) levantam outra limitação da TCM, que é

a suposição de que a variação dos erros de medida (E) é a mesma para todos os respondentes, o que parece ser uma suposição de difícil sustentação. Embretson e Reise (2000) exemplificam que um conjunto com itens medianamente fáceis poderá diferenciar mais a prática organizacional ou o sujeito com habilidade média, mas não irá diferenciar da mesma maneira os que tiverem habilidade superior, que provavelmente obterão escores perto dos mais altos. Conseqüentemente, o erro de medida (E) neste segundo grupo será maior que no primeiro.

Tendo em vista estas limitações, Pasquali e Primi (2003) argumentam que, dentro dos moldes da TCM, a tarefa de comparar os sujeitos ou organizações de forma vertical ou longitudinal é difícil, uma vez que obter um escore total em um conjunto de itens fáceis não é a mesma coisa do que obter o mesmo escore em um conjunto de itens difíceis. Para equacionar estes e outros problemas da TCM, tem-se trabalhado, mas mais diversas áreas do conhecimento com a Teoria da Resposta ao Item, a apresenta várias vantagens em relação à TCM.

3. Teoria da Resposta ao Item

A Teoria da Resposta ao Item (TRI) é um conjunto de modelos matemáticos e estatísticos que são utilizados para (a) análise de itens e escalas, (b) criar e administrar medidas, e (c) medir indivíduos ou organizações em um construto (traço latente) de interesse. A TRI é sustentada basicamente por três fundamentos: (a) função resposta de um item, (b) função de informação e (c) invariância (REISE ET AL., 2005).

Apesar de ter tido suas raízes elaboradas no início do século passado com os trabalhos de Thurstone (1927), Richardson (1936), Mosier (1942), Ferguson (1942), Lawley (1943), Tucker (1946) e Lazarsfeld (1950) entre outros, a TRI, nos moldes que se conhece hoje, foi desenvolvida por Lord (1952) o qual elaborou o modelo teórico e o método para estimar os parâmetros dos itens utilizando o modelo da ogiva normal. Mais tarde, Birnbaum (1968), tornou o tratamento matemático mais fácil ao substituir o modelo de ogiva normal pelo modelo logístico (baseado em logaritmo). Entretanto, um ponto crítico na utilização da TRI é a estimação dos parâmetros, que necessita de um grande esforço computacional (ANDRADE ET AL., 2000), o que dificultou a popularização da TRI naquela época, sendo retomada recentemente em paralelo aos avanços na área computacional. No Brasil a TRI foi aplicada pela primeira vez em 1995 na área educacional na análise dos resultados do Sistema Nacional de Ensino Básico (SAEB) e no Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP) (ANDRADE ET AL., 2000).

No Brasil, além da área educacional, a TRI vem sendo aplicada em outras áreas, tais como: serviços (COSTA, 2001), gestão pela qualidade total (ALEXANDRE ET AL., 2001), qualidade de vida (GUEWEHR, 2007), avaliação de intangíveis nas organizações (VARGAS ET AL, 2008), usabilidade em sites de e-commerce (TEZZA, 2009).

Segundo Singh (2004), um modelo matemático, que especifica a relação entre variáveis observáveis e variáveis latentes, é especificado por três características básicas. A primeira considera a forma do modelo. Neste caso, observa-se que a TRI é mais geral que a TCM, uma vez que postula um modelo não-linear, enquanto a TCM utiliza um modelo linear.

A segunda característica é a descrição do modelo através de uma equação específica. Na TCM, tal representação é feita pela equação 1. Na TRI, existem vários modelos não-lineares. Entretanto, é discutido no presente trabalho o modelo logístico de 2 parâmetros (ML2P) que é comumente utilizado para modelar construtos cumulativos com variáveis dicotômicas. O ML2P (equação 2) pressupõe a relação monótona entre o valor da variável latente (θ) e a sua

probabilidade de escolha entre duas alternativas segundo uma função de distribuição logística parametrizada por coeficientes que representam as características do item.

$$P_i(\theta_j) = P(U_{ij} = 1/\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad (\text{equação 2})$$

Onde, b_i representa a dificuldade do item i em uma determinada escala e representa o valor da variável latente θ , para o qual há 0,5 de probabilidade do indivíduo j escolher a resposta representada por $U = 1$. O coeficiente a_i é o parâmetro de discriminação do item i , proporcional à inclinação da Curva Característica do Item – CCI no ponto b_i .

A terceira característica enfatiza os parâmetros do modelo. Na TCM, um único parâmetro, o escore verdadeiro (X), associa a variável observada com a latente. Em contrapartida, na TRI, vários parâmetros distintos fazem esta associação. Por exemplo, o modelo logístico de dois parâmetros considera o parâmetros de discriminação do item (a_i), que define qual o poder que um item tem de diferenciar indivíduos ou organizações com habilidades semelhantes, e o parâmetro de dificuldade do item (b_i), que define o ponto na escala gerada onde a probabilidade de resposta positiva (ou acerto) é de 50%.

Durante a criação de um instrumento de medida, podem ocorrer alguns equívocos, tais como (a) dificuldade dos itens inadequada (ou muito difícil ou muito fácil), (b) a área de incerteza entre acertar ou errar um item é muito grande e (c) o item pode indicar que um atributo está presente enquanto ele realmente não está. Estes equívocos podem ser identificados na TRI através de seus parâmetros, de forma que pode-se selecionar itens que estejam mais alinhados com o contexto em que se deseja medir (DEVELLIS, 2003).

A aplicação da TRI requer algumas considerações iniciais. Um deles é a determinação do modelo da TRI que melhor se encaixa ao conjunto de dados a serem analisados e aos objetivos da investigação. Outra consideração importante, do qual também depende a determinação do modelo, é a determinação da dimensionalidade do traço latente, ou seja, a quantidade de traços latentes que serão medidos. Segundo Singh (2004), a dimensionalidade refere-se ao número de fatores necessários para explicar a variabilidade dos dados e constituir uma hipótese a ser verificada. A dimensionalidade pode assumir um caráter unidimensional, quando existir apenas um fator em análise, por exemplo, a usabilidade de um site de e-commerce ou multidimensional, quando existirem mais de um fator. De um modo geral, segundo Schmitt (1996), quanto mais estritamente unidimensional for o construto, menos ambíguas as interpretações dos resultados e mais legítimas são as correlações. Entretanto, embora a suposição de unidimensionalidade destina-se a medir apenas um traço latente, geralmente a natureza dos testes estatísticos introduz, muitas vezes, elementos multidimensionais. Quando existir dúvida quanto à unidimensionalidade de um construto, segundo Edelen e Reeve (2007), é importante verificar os resultados do modelo da TRI adotado, a fim de identificar eventuais anomalias que possam surgir devido à violação deste pressuposto, como, por exemplo, um ou mais itens com parâmetro de discriminação com baixíssima declividade pode indicar problema de dimensionalidade.

A TRI cria uma escala, relacionada ao traço latente, com média zero e desvio padrão um, onde são estimados os parâmetros dos itens. Durante estas estimações, são geradas as curvas características dos itens (CCI), nas quais pode-se visualizar a probabilidade de acerto do item ($P_i(\theta_j)$) em função da habilidade do respondente (θ). A figura 1 exemplifica CCI de dois itens hipotéticos.

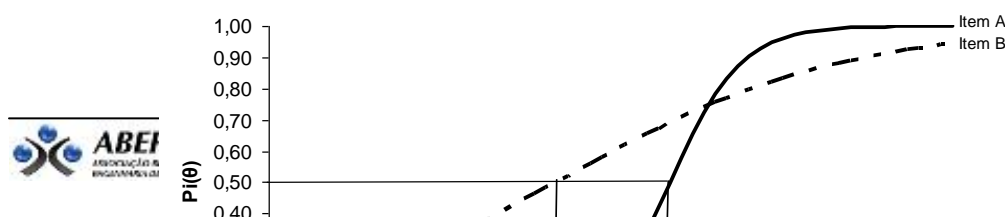


Figura 1 : Curva Característica do Item (CCI)

Observa-se que o eixo y representa os valores da função de probabilidade $P_i(\theta)$ que variam de zero a um. No eixo x, tem-se a escala de habilidade (θ) gerada pela TRI. Nesta escala, são posicionados os itens e os respondentes. Assim, é possível comparar os desempenhos dos respondentes e a qualidade dos itens verticalmente e longitudinalmente. Pode-se visualizar, na figura 1, a representação gráfica dos parâmetros. Nota-se que o parâmetro de dificuldade (b) de um item representa a posição na escala onde a probabilidade de acerto é de 0,5. Sendo assim, o item A tem um $b = 0,5$ e o item B tem um $b = -0,5$ na escala gerada. Portanto, o item A possui um grau de exigência maior que o item B, ou seja, para um indivíduo com habilidade (θ) igual a $-0,5$ tem 50% de probabilidade de acertar o item B e 4% de probabilidade de acertar o item A. O parâmetro de discriminação (a) de um item, segundo Andrade et al. (2000), é proporcional à derivada da tangente da curva no ponto de inflexão, ou seja no ponto b . Verifica-se que o item A em um inclinação maior que o item B, portanto um maior poder de discriminação.

Outro gráfico que a TRI fornece é a curva de informação de um item (CII), a qual representa quanto um item possui de informação em toda a escala e está, segundo Baker (2001), associada à precisão do instrumento. A CII é representada pela equação 3.

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad \text{equação 3}$$

Onde, $I_i(\theta)$ é a informação fornecida pelo item i ao nível da aptidão θ ; $P_i(\theta)$ é o modelo de probabilidade (equação 2); $P_i'(\theta)$ é a derivada de $P_i(\theta)$ em relação a θ e $Q_i(\theta) = 1 - P_i(\theta)$.

4. Aplicação

Com o objetivo de exemplificar as definições da TRI discutidas na seção anterior, esta seção apresenta e discute brevemente uma aplicação da TRI na área de ergonomia de interface, voltada para o mercado de *e-business*.

Na área de ergonomia de interface, um ponto chave é a usabilidade de uma interface. Segundo Nielsen e Loranger (2006), a usabilidade é um atributo de qualidade relacionado à facilidade de uso deste. Dentro do mercado de *e-business*, esta característica assume um importante papel. Winckler (2001) argumenta que problemas de usabilidade em site de *e-commerce* podem significar redução nas vendas ou mesmo venda nenhuma. Atualmente, as medidas de

usabilidade em *sites* de e-commerce vêm sendo realizadas de diversas maneiras, utilizando metodologias ou ferramentas que envolvem ou não o usuário (NILSEN; MACK, 1994; CYBIS, 2007). Estas medidas envolvem características objetivas e subjetivas, baseadas em critérios recomendados por especialistas ou em opiniões de usuários, gerando falta de sistematização e de precisão nos resultados (CYBIS, 2007). Sendo assim, a TRI pode significar uma oportunidade para criação de um instrumento de medida que possua uma escala padronizada e permita sistematização e precisão dos resultados e, conseqüentemente, comparabilidade dos mesmos.

4.1. Método

Para elaboração de um instrumento utilizando a Teoria da Resposta ao Item, é necessário primeiramente definir qual traço latente de interesse, ou seja, é necessário ter um conhecimento prévio do que se deseja medir. A TRI faz a análise dos itens e, a partir daí, constrói a escala padronizada. A elaboração dos itens deve contar com um conhecimento a priori do assunto e seguir técnicas de elaboração de itens (PASQUALI, 2003; WILSON, 2005). Geralmente a elaboração dos itens é realizada por especialistas ou extraída de literatura específica. No caso da presente aplicação, a elaboração dos itens esteve embasada em referências específicas de usabilidade. Foram elaborados inicialmente 73 itens, todos com resposta dicotômica, do tipo “possui” ou “não possui” a característica em questão. A aplicação do conjunto de itens aos sites foi realizada pelos autores. Para verificar a operacionalidade do conjunto bem como seu alinhamento com os pré requisitos da TRI (unidimensionalidade, interdependência, clareza e objetividade) foi realizado um pré-teste, com cerca de 20 *sites*, quando foram eliminados 29 itens. Após este refinamento, os 44 itens restantes (tabela 1) foram aplicados em 361 sites de *e-commerce* de origem brasileira. O tratamento dos dados foi feito no *software* BILOG-MG[®] produzido pela *Scientific Software, Inc*, Mislevy, R. J. e Bock, R. D, (1990).

4.2. Resultados e Análises

Os resultados gerados pela Teoria da Resposta ao Item com o auxílio computacional do *software* BILOG-MG[®] é sistematizada em três fases. Na primeira fase, é feito o reconhecimento dos dados e identificação do que é correto e do que não é correto, gerando com isso algumas estatísticas descritivas tais como: número de *sites* submetidos a cada item, o número de acertos em cada item e algumas correlações de interesse, como correlação de Pearson e correlação bisserial, bastante usadas na Teoria Clássica dos Testes. Segundo Andrade et al. (2000), a importância desta fase está no fato de ela verificar se a leitura dos dados está sendo feita de maneira correta e estas estatísticas geradas são utilizadas posteriormente como valores iniciais para os processos de estimação realizados nas fases seguintes.

Na segunda fase, são estimados os parâmetros de discriminação “a” e de dificuldade “b” de cada item e seus respectivos erros padrão, que podem ser visualizados na tabela 1.

Item	Descrição	Parâmetro		Item	Descrição	Parâmetro	
		a	b			a	b
1	Ao abrir a homepage não há abertura de janelas pop-up?	0,597	-3,61	23	A maioria dos produtos possui informações sobre eles?	1,022	-2,11
2	Homepage deixa claro o que o site faz (demonstra os principais produtos e/ou uma breve descrição de seu objetivo e/ou benefícios que oferece), sem precisar usar a rolagem?	0,756	-1,55	24	O tamanho da fonte dos textos é de dez pontos ou mais?	0,402	1,99

3	A homepage possui menos de 2,3 paginas de rolagem?	0,246	2,42	25	É possível ampliar as fotos dos produtos para visualizar detalhes?	0,931	-0,98
4	As palavras clicáveis (coloridas ou sublinhadas) apresentam uma forma diferenciada quando são selecionadas?	0,979	-1,35	26	Em produtos em que existam mais de uma perspectiva, é possível visualizar todas as perspectivas?	0,972	1,81
5	O layout da página é adaptável ao tamanho da janela, pelo menos horizontalmente (ou seja não há rolagem horizontal)?	0,505	0,10	27	Os grupos de botões de comando estão dispostos em coluna e à direita, ou em linha e abaixo dos objetos aos quais estão associados?	0,909	-2,90
6	As opções principais do site estão visíveis? Ou seja não há necessidade de passar o mouse sobre uma área gráfica para ver opções ocultas.	0,812	-2,97	28	Quando há rolagem, não existe elementos de design (na tela inicial) que pareçam com marcadores de final de página?	0,938	-3,24
7	O site possui a opção "ajuda" ?	0,648	1,80	29	Todos os campos e mostradores de dados possuem rótulos identificativos?	0,906	-2,40
8	Na área principal do site existem textos com no máximo quatro cores diferentes?	0,497	-0,56	30	O botão de finalização de compra está no final da lista?	0,752	-4,49
9	Na área principal do site existem textos com no máximo três fontes diferentes?	0,527	-0,72	31	É possível saber os custos totais antes de fazer cadastro?	0,792	-2,21
10	A disposição dos objetos de interação de uma caixa de diálogo segue uma ordem lógica?	0,854	-0,85	32	Os campos destinados a formulários de preenchimento pelo usuário estão agrupados linearmente, evitando espaços desnecessários?	0,490	-3,36
11	Os rótulos de campos começam com uma letra maiúscula, e as letras restantes são minúsculas?	0,869	-1,46	33	No preenchimento de um formulário, as opções que não são válidas ou não estão disponíveis estão visivelmente desativadas, prevenindo assim erros?	0,595	0,82
12	O site possui opção de acesso com outras línguas?	0,79	4,65	34	No preenchimento de um formulário, é informado a forma de preenchimento?	1,043	-0,55
13	É possível encontrar informações sobre a empresa (endereço físico, telefone, etc)?	0,549	0,35	35	Os dados obrigatórios são diferenciados dos dados opcionais de forma visualmente clara?	0,860	-1,04
14	Os títulos estão alinhados a esquerda?	0,972	-2,42	36	É possível efetivar uma compra sem realizar um cadastro (que inclua login e senha)?	0,349	5,95
15	Os parágrafos de texto são separados?	0,891	-3,03	37	O sistema emite sinais sonoros quando ocorrem problemas na entrada de dados?	0,382	-3,22
16	As palavras aparentemente clicáveis são de fato clicáveis ?	1,075	0,16	38	As mensagens de erro estão isentas de abreviaturas e/ ou códigos gerados pelo sistema operacional?	1,073	-2,88
17	Os títulos de telas, janelas e caixas de diálogo estão no alto, centrados ou justificados à esquerda?	1,376	-2,55	39	Todas as páginas possuem os mesmos layouts e exibem ao usuário as mesmas características?	1,337	-1,58
18	Todas as páginas possuem um campo de busca?	0,755	-2,70	40	O logotipo da empresa está no canto superior esquerdo em todas as páginas do site?	1,53	-1,74
19	Os resultados de busca permitem classificação por outros critérios além de custo?	1,204	0,46	41	Existe um link de um único clique que conduz a homepage?	1,178	-2,19
20	Listas longas apresentam indicadores de continuação, de quantidade de itens e de páginas?	0,791	-0,28	42	Qualquer ação do usuário pode ser revertida através da opção DESFAZER?	1,350	-1,24
21	O preço de um produto consta ao lado da imagem ou do link do produto?	1,120	-2,63	43	O site permite navegação em suas páginas em apenas uma janela, ou seja, não há abertura de novas janelas em meio a navegação?	1,056	-1,74
22	Existe uma orientação ao usuário quanto ao restante do site?	1,000	0,17	44	Os links já visitados mudam de cor?	0,716	4,29

Tabela 1 - Estimação dos parâmetros dos itens.

Quanto ao grau de dificuldade 'b', os três itens com maior dificuldade foram, em ordem crescente, os itens 44, 12 e 36. Já os itens com menor grau de dificuldade, também em ordem crescente, foram 30, 01 e 32. É importante ressaltar que os valores referentes ao grau de dificuldade são referenciados a uma escala baseada na distribuição normal de métrica (0,1), ou seja, média 0 (zero) e desvio-padrão 1 (um). Sendo assim, os valores, não possuem um

significado prático em termos de usabilidade. Entretanto, possuem a capacidade de comparabilidade entre os itens e em relação a escala de usabilidade.

O parâmetro grau de dificuldade não avalia a “qualidade” de um item, ele apenas indica qual a região, na escala proposta, em que o item possui maior informação. Já o grau de discriminação de um item, determina a “qualidade” do item, ou seja, quanto maior o valor de “a” maior o grau de discriminação na região de maior informação (“b”), conseqüentemente melhor o item. Itens com grau de discriminação inferiores a 0,700 foram considerados pouco informativos, e desta forma eliminados do construto, foram eles: 01, 03, 05, 07, 08, 09, 13, 24, 32, 33, 36 e 37. Os itens com maior grau de discriminação foram, em ordem crescente, 42, 17 e 40.

A figura 2 demonstra a curva característica e a curva de informação do item 40. Este item possui a seguinte descrição: O logotipo da empresa está no canto superior esquerdo em todas as páginas do *site*?. Uma vez que o parâmetro de dificuldade deste item, na escala (0, 1), é igual a -1,742, sites com grau de usabilidade igual -1,742 tem 50% de chance de possuir esta característica. Observa-se pela CCI que sites com grau de usabilidade maior que 1,000 tem praticamente 100% de probabilidade de acertar este item. Sendo assim, como pode ser visualizado na CII, o item 40 é um item relativamente fácil, pois concentra toda sua informação na região compreendida entre -4 e 1.

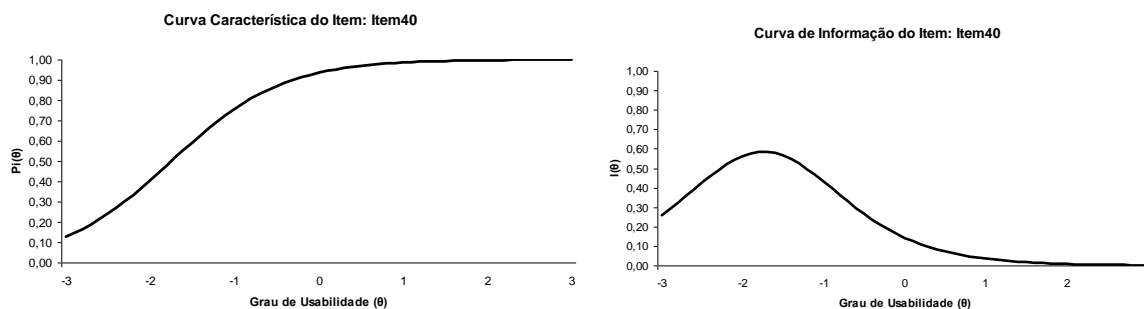


Figura 2 : Curva Característica e curva de informação do item 40

Este tipo de análise é realizada com todos os itens e por fim é gerado um gráfico total de informação (figura 3) que é a soma da informação contida em cada item e o erro associado em cada região da escala. Esta curva informa em qual região da escala existe maior precisão do instrumento. Pode-se concluir, pela análise da figura 3, que o instrumento gerado com os 32 itens é um bom instrumento para avaliar sites com grau de usabilidade entre -4 e 1, na escala gerada, ou seja, sites com grau de usabilidade baixo. Na região acima de 1 pode-se verificar que a curva do erro padrão (linha pontilhada) é maior que a curva de informação (linha contínua). Desta forma, pode-se verificar que na TRI o erro padrão varia de item para item e em cada região da escala. O que não é verificado na TCM.

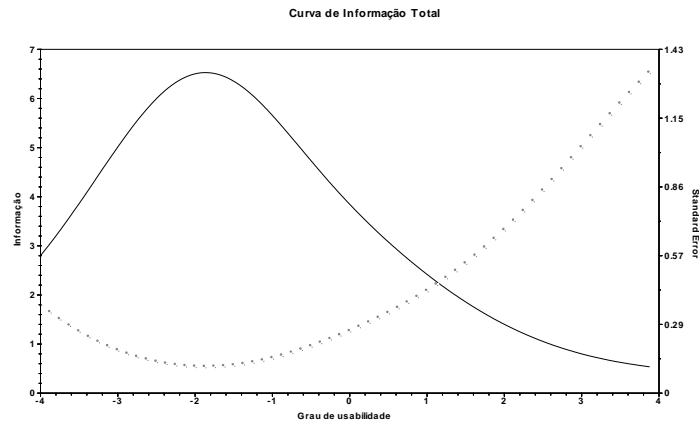


Figura 3: Curva de Informação Total dos 32 itens.

A figura 4 demonstra a distribuição de frequência dos sites analisados, dentro da escala gerada. Nota-se que, como já era de se esperar, numa escala (0, 1) a média ficou muito próxima de zero e o desvio padrão próximo de um.

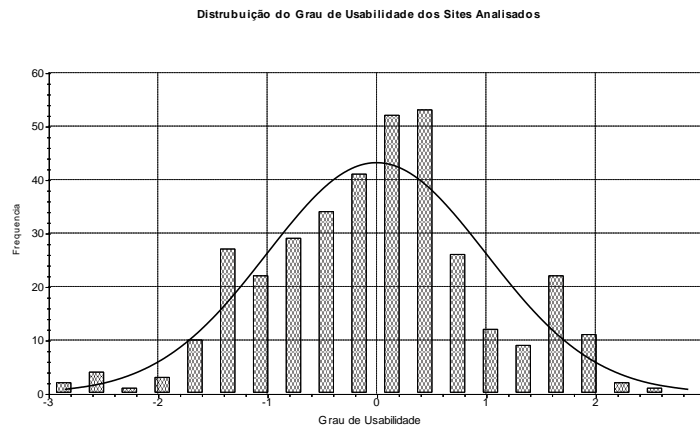


Figura 4: Distribuição do grau de usabilidade dos sites analisados

A figura 3 e a figura 4 apresentam uma importante vantagem da TRI, que é a independência da amostra com os resultados finais, pois embora a maioria dos sites utilizados para estimação dos parâmetros e geração da escala estejam em torno de zero, a maioria dos itens apresentou maior poder de discriminação na região negativa da escala. Ou seja, na TRI, os parâmetros de dificuldade do item é independente de quem está respondendo o item.

Com dos resultados obtidos na estimação dos parâmetros dos itens é possível fazer uma análise aprofundada do ponto de vista estatístico e prático de cada um dos itens e verificar se existe convergência teórico-prática destes resultados. E a partir daí gerar a escala padronizada onde são identificados os aspectos mais relevantes sobre o assunto, no caso a usabilidade em sites de e-commerce, ordenados na escala.

A análise realizada de cada item mostrou que os conceitos de navegabilidade e busca assumem uma importância grande na usabilidade de sites de e-commerce, isto foi verificado pelo fato de itens tratando destes aspectos (itens 16, 22, 39, 40, 41, 42, 43) estarem mais concentrados na região a direita da escala gerada (dentro da curva de precisão). Os aspectos relacionados com a parte puramente visual, ou de design dos sites (itens 06, 15, 17, 27, 28)

mostraram-se uma característica básica e foi caracterizada por itens de baixo grau de dificuldade.

5. Conclusão

A Teoria da Resposta ao Item representa um avanço analítico para todas as áreas do conhecimento, incluindo a engenharia de produção, onde algumas destas vantagens são fundamentais para facilitar o controle e a gestão de recursos e processos. Entre estas vantagens, destacam-se: (a) poder de posicionar indivíduos ou processos de diferentes grupos em uma escala comum, mesmo que estes tenham respondido a itens diferentes, permitindo a identificação de oportunidades de melhoria ou até mesmo benchmarking, (b) permitir uma mais precisa avaliação das propriedades dos itens e seus resultados e, conseqüentemente, permitir maior precisão na aplicação de técnicas estatísticas, (c) compreender adequadamente as propriedades psicométricas dos instrumentos, (d) possibilidade de desenvolvimento de indicadores mais eficientes para avaliar diferenças individuais de processo, práticas, sistemas ou indivíduos, (e) maior robustez dos resultados.

No estudo de caso apresentado, criou-se uma escala padronizada para medir o grau de usabilidade em sites de e-commerce permitindo identificar o quanto aspectos de navegabilidade e ferramentas de busca tem impacto na usabilidade este mercado. Sendo assim, a escala gerada permitiu (a) quantificar a relevância de cada aspecto de usabilidade avaliado, (b) identificar quais empresas possuem maior grau de usabilidade em relação às concorrentes, (c) identificar quais aspectos a empresa necessita desenvolver para melhorar sua usabilidade (d) identificar claramente as limitações do instrumento, principalmente no que diz respeito a sua área de precisão.

Talvez as maiores desvantagens da TRI sejam sua complexidade e a necessidade de amostras grandes. Entretanto, estas desvantagens, quando confrontadas com suas vantagens, tornam-se perfeitamente contornáveis, e isto vem ocorrendo de forma acelerada em algumas áreas, como psicologia e educação. Em paralelo a isso, este artigo reforça a grande potencialidade de TRI para a engenharia de produção, não só na parte de ergonomia de interface mas também no âmbito gerencial, de qualidade, tomada de decisão, mensuração de intangíveis entre outros.

De uma forma geral, a TRI é uma alternativa muito mais elaborada que a TCM para construção de instrumentos, principalmente quando se necessitam medidas consistentes em toda a escala, e em contextos em que se decisões de grande impacto são tomadas com base na pontuação de um indivíduo ou sistema

Referências

ALEXANDRE, J. W. C.; ANDRADE, D.F.; VASCONCELOS, AP.; ARAUJO, A M S. *Uma proposta de análise de um construto para a medição dos fatores críticos da gestão pela qualidade através da TRI*. Gestão & Produção. v.9, n.2, p.129-141, 2001.

ANDRADE, D. F.; TAVARES, H. R. e VALLE, R. C. *Teoria de Resposta ao Item: conceitos e aplicações*. ABE — Associação Brasileira de Estatística, 4º SINAPE, 2000.

BAKER, F. B. *The Basics of Item Response Theory*. 2 ed. USA: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BIRNBAUM, A. *Some Latent Trait Models and Their Use in Inferring an Examinee's Ability*. In F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*, MA: Addison-Wesley, 1968.

CHURCHILL, G. Jr., *A Paradigm for Developing Better Measures of Marketing Constructs*. Journal of Marketing Research, 26:64-73, February 1979.

- COSTA, M. B. F.** *Técnica derivada da Teoria de Resposta ao Item (TRI) aplicada ao setor de serviços*. Dissertação de Mestrado, UFPR, Curitiba, 2001.
- CYBIS, W.** *Ergonomia e Usabilidade : conhecimentos, métodos e aplicações* / Walter Cybis, Adriana Holtz Betiol, Richard Faust. São Paulo : Novatec Editora, 2007.
- DEVELLIS, R. F.** *Scale development: Theory and applications*, 2ed. Thousand Oaks, CA: Sage Publications 2003.
- DRUCKER, P. F.** The information executives truly need. *Harvard Business Review*, v.73, n.1, p.54-62, 1995.
- EDELEN, M. O.; REEVE, B. B.** *Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement*. *Qual Life Res* v.16 p.05–18. 2007.
- EMBRETSON, S.; REISE, S. P.** *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc. Publishers, 2000.
- FERGUSON, G. A.** *Item selection by the constant process*, *Psychometrika*, v.7 n.1, 1942.
- GIL, A. C.** *Como elaborar projetos de pesquisa*. 4^o ed São Paulo: Atlas, 2002.
- GUEWEHR, K.** *Teoria da resposta ao item na avaliação de qualidade de vida de idosos*. Dissertação de mestrado - UFRGS, Faculdade de Medicina, Programa de Pos-Graduação em Epidemiologia, RS, 2007.
- HAMBLETON, R.K., SWAMINATHAN, H.,** *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers, 1985.
- HAMBLETON, R.K., SWAMINATHAN, H., ROGERS, H.J.** *Fundamentals of item response theory*. Newbury Park, CA: Sage, 1991.
- HAYES, B. E.** *Measuring customer satisfaction: development and use of questionnaires*. Milwaukee, Wisconsin: ASQC Quality Press, 1992.
- KAPLAN, R. S.; NORTON, D. P.** *A Estratégia em Ação – Balanced Scorecard*. 7. ed. RJ: Campus, 1997.
- LAWLEY, D. N.** *On problems connected with item selection and test construction*. *Proceedings of the Royal Society of Edinburgh* 61, p. 273-287, 1943.
- LAZARFELD, P.F.** Chapter 10 and 11 in S.A. Stouffer et al. (eds.). *Studies in Social Psychology in World War II*: v. 4. Measurement and Prediction. Princeton, NJ: Princeton University Press. 1950.
- LORD, F. M.** *A theory of test scores*. *Psychometric Monograph No. 7*, Psychometric Society, 1952.
- MISLEVY, D. J., BOCK, R. D.** *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Chicago: Scientific Software, 1990.
- MOSIER, C. I.** *A modification of the method of successive intervals*. *Psychometrika*, v. 7 n. 1 p. 19-29, 1942.
- NIELSEN, J.; LORANGER, H.** *Prioritizing Web Usability*. California : New Riders, California, 2006.
- NIELSEN, J.; MACK, R.** *Usability Inspection Methods*. New York: John Wiley & Sons, 1994.
- NUNNALLY, J.C.** *Psychometric theory*. Nova York: McGraw-Hill, 1978.
- PASQUALI, L.** *Psicometria: teoria dos testes na psicologia e na educação*. Petropolis, RJ: Vozes, 2003.
- PASQUALI, L.** *Teoria da resposta ao item – IRT: uma introdução*. In L. Pasquali (Org.), *Teoria e métodos de medida em ciências de comportamento*. Brasília: INEP, 1996.
- PASQUALI, L.; PRIMI, R.** *Fundamentos da Teoria da Resposta ao Item – TRI*. *Avaliação Psicológica*, v.2 n.2, p. 99-110, 2003.
- REISE, S.P. AINSWORTH, A.T. HAVILAND, M.G.** *Item Response Theory Fundamentals, Applications, and Promise in Psychological Research*, *Current Directions in Psychological Science*, **Vol. 14 n. 2, p 95 – 101, 2005**.
- RICHARDSON M.W.** *The relationship between difficulty and the differential validity of a test*. *Psychometrika*. v.1 p.33-49. 1936.
- SCHMITT, N.** *Uses and abuses of coefficient alpha*. *Psychological Assessment*, v.8, p.350-353, 1996.

- SINGH, J.** *Tackling measurement problems with Item Response Theory: Principles, characteristics, and assessment, with an illustrative example.* Journal of Business Research, v. 57, p. 184-208, 2004.
- TEZZA, R.** *Proposta de um construto para medir usabilidade em site de e-commerce utilizando a Teoria da Resposta ao Item.* Dissertação de Mestrado - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis, 2009
- TUCKER, L.R.** *Maximum validity of a test with equivalent items.* Psychometrika, v.11, p.1-13. 1946.
- THURSTONE, L. L.** *A law of comparative judgments.* Psychological Review, v.34, p.278-286, 1928.
- THURSTONE, L. L.** *Attitudes can be measured.* American Journal of Sociology, v.33, p.529-554, 1927.
- TRUCKER, L.R.** *Maximum validity of a test with equivalent items.* Psychometrika v.11, p.1-13. 1946.
- VARGAS, V.C.C., SELIG, P.M., ANDRADE, D.F., RIBEIRO, J.L.D.** *Avaliação dos intangíveis: uma aplicação em capital humano.* Gestão e Produção, São Carlos v. 15, n.3, p.619-634, 2008.
- WEISS, D.J.** *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press, 1983.
- WILSON, M.** *Constructing Measures: An Item Response Modeling Approach,* Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey, London, 2005
- WINCKLER, M.** *Avaliação de usabilidade de sites Web.* In: IV Workshop sobre fatores humanos em sistemas de computação. Out. 2001, Florianópolis.