Produção, v. 23, n. 3, p. 525-536, jul./set. 2013
http://dx.doi.org/10.1590/S0103-65132012005000095

PRODUÇÃO

Algoritmo de um teste adaptativo informatizado com base na teoria da resposta ao item para a estimação da usabilidade de *sites* de *e-commerce*

Fernando de Jesus Moreira Junior[a]*, Rafael Tezza[b], Dalton Francisco de Andrade[c], Antonio Cezar Bornia[d]

[a]*fmjunior777@yahoo.com.br, UFSM, Brasil
[b]rafaeltezza@yahoo.com.br, UDESC, Brasil
[c]dandrade@inf.ufsc.br, UFSC, Brasil
[d]cezar@deps.ufsc.br, UFSC, Brasil

Resumo

O presente artigo propõe um algoritmo de um teste adaptativo informatizado baseado na teoria da resposta ao item, desenvolvido para estimar o grau de usabilidade de *sites* de *e-commerce*. Cinco algoritmos baseados no critério da máxima informação foram desenvolvidos e testados via simulação. O algoritmo com o melhor desempenho foi aplicado nos dados reais de 361 *sites* de *e-commerce*. Os resultados mostraram que o algoritmo desenvolvido consegue obter uma boa estimativa para o grau de usabilidade de *sites* de *e-commerce* com a aplicação de 13 itens.

Palavras-chave

Teste adaptativo informatizado. Teoria da resposta ao item. Usabilidade. *Sites* de *e-commerce*.

1. Introdução

A usabilidade em *websites* é um atributo de qualidade relacionado à facilidade de uso destes (NIELSEN; LORANGER, 2006). A medição de conceitos intangíveis, como qualidade, não é simples nem linear (STRAUB, 1989). Esse tipo de atributo que não pode ser medido diretamente – tais como a satisfação de usuários com determinada interface gráfica, o grau de aceitabilidade de novas tecnologias – é chamado de traço latente e necessita ser estimado de forma indireta utilizando-se instrumentos compostos por diversos itens secundários relacionados a ele (MAGOUTAS et al., 2010). Uma forma eficiente e rápida para estimar traços latentes é através de um teste adaptativo informatizado.

Testes adaptativos informatizados (TAI) são testes que procuram estimar a habilidade do respondente através da aplicação de itens que sejam adequados a ele (VAN DER LINDEN; GLAS, 2000). O resultado disso é um teste mais eficiente, rápido e confiável. Para implantar um TAI é necessário um banco de itens previamente calibrado, normalmente através da teoria da resposta ao item.

A teoria da resposta ao item (TRI) é um conjunto de modelos matemáticos que define uma maneira de estabelecer a correspondência entre variáveis latentes (tipo de variável que não pode ser medida diretamente) e suas manifestações (DE AYALA, 2009), possibilitando a criação de medidas padronizadas. Segundo De Ayala (2009), a TRI pode ser descrita como uma teoria de estimações estatísticas na qual características latentes de indivíduos ou sistemas são estimadas tendo como base as respostas deles a um determinado conjunto de itens. São exemplos de variáveis latentes: o grau de usabilidade de um *website*, o nível de proficiência em matemática de um aluno do ensino fundamental e o grau de depressão ou estresse. Atualmente, a teoria da resposta ao item vem sendo bastante difundida no mundo todo, principalmente na área de educação e testes psicológicos. Uma relação de trabalhos sobre TRI publicados no Brasil até o ano de 2009 encontra-se disponível em Moreira Junior (2010).

*UFSM, Santa Maria, RS, Brasil
Recebido 18/05/2011; Aceito 03/08/2012

O presente artigo tem como objetivo propor um algoritmo de um teste adaptativo informatizado baseado na teoria da resposta ao item desenvolvido para estimar o grau de usabilidade de *sites* de *e-commerce*. Em Tezza, Bornia e Andrade (2011), um pequeno banco de 32 itens foi analisado e calibrado e a usabilidade de *sites* reais foi estimada. A questão fundamental do presente trabalho é verificar se é possível estimar razoavelmente o grau de usabilidade de *sites* de *e-commerce* utilizando uma quantidade menor de itens (menos de 32), segundo alguns critérios de seleção de itens.

2. Estimação do grau de usabilidade de *sites* de *e-commerce*

Segundo Nielsen e Loranger (2006), a usabilidade em *websites* é um atributo de qualidade relacionado à facilidade de uso dele. Mais especificamente, (1) refere-se à rapidez com que os usuários podem aprender a usá-lo, (2) à eficiência deles ao serem usados, (3) o quanto os usuários lembram dele, (4) o grau de propensão a erros e (5) o quanto gostam de utilizá-lo.

Atualmente existem várias formas de avaliar e até medir usabilidade, entre elas destacam-se: inspeção cognitiva (KIERAS; POLSON, 1999), inspeção (IVORY; MEGRAW, 2005; CHEVALIER; BONNARDEL, 2007), grupo focal (CHOE et al., 2006; LARGE et al., 2006), avaliações heurísticas (NIELSEN, 1993; AGARWAL; VENKATESH, 2002), testes com usuários (SCHENKMAN; JÖNSSON, 2000; LAZAR; MEISELWITZ; NORCIO, 2004; FANG; HOLSAPPLE, 2007) e *card sorting* (RAU; LIANG, 2003; ROSSO, 2008). Essas medidas envolvem características objetivas e subjetivas, baseadas em critérios recomendados por especialistas ou em opiniões de usuários, gerando, muitas vezes, falta de sistematização e de precisão nos resultados (CYBIS, 2007), o que torna a maioria dessas medidas restritas a casos particulares de análise. Consequentemente, a subjetividade e a falta de sistematização nos resultados dificultam a comparabilidade entre os sistemas e a identificação das características mais importantes.

Para gerar um melhor entendimento das estruturas envolvidas em uma avaliação de usabilidade e sistematizar os resultados pode-se fazer uso de escalas de medida alicerçadas em conceitos matemáticos e de usabilidade. Desse ponto de vista, a teoria da resposta ao item representa uma poderosa ferramenta, uma vez que possibilita a criação de escalas a partir de um conjunto de itens que faz uso de conceitos aprofundados de usabilidade. Esse processo dá-se necessariamente por meio de estimações de parâmetros. Num primeiro momento, são estimados os parâmetros dos itens com a finalidade de mensurar a importância dos atributos de usabilidade e posicioná-los em uma escala (de grau de usabilidade, por exemplo), ordenados por ordem de importância (ou dificuldade). Depois, nessa mesma escala são realizadas as estimações do grau de usabilidade dos *websites* que "responderem" ao conjunto de itens já estimados. Ou seja, a aplicação da TRI configura um processo iterativo, no qual, a partir de um conjunto de itens, é possível estimar a "proficiência" (ou grau de usabilidade) de um respondente qualquer, independente de seu contexto, representando uma abordagem objetiva e geral.

3. Teoria da resposta ao item

A teoria da resposta ao item (TRI) é um conjunto de modelos estatísticos que procura medir traços latentes por meio de um conjunto de itens e da construção de uma escala na qual o traço latente do respondente e a dificuldade de um item podem ser comparados (HAMBLETON, 2000; HAMBLETON; SWAMINATHAN; ROGERS, 1991; EMBRETSON; REISE, 2000). Na TRI, a escolha do modelo matemático depende basicamente do tipo de item e representa a probabilidade de resposta a um item em função dos parâmetros do item e da proficiência do respondente (TAVARES; ANDRADE; PEREIRA, 2004; REISE; WIDAMAN; PUGH, 1993). O modelo mais utilizado para itens com resposta dicotômica e acumulativa, sem a possibilidade de acerto casual, que é o caso do objeto de análise neste trabalho, é o modelo logístico de dois parâmetros (ML2P) desenvolvido por Birnbaum (1968), representado pela seguinte equação:

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

onde:

- a_i é o parâmetro de discriminação do item *i*, proporcional à inclinação da função no ponto b_i;
- b_i é o parâmetro de dificuldade (ou de posição) do item *i*, medido na mesma escala do grau de usabilidade;
- U_{ij} é a resposta ao item *i* do site *j*, que pode ser positiva (1) ou negativa (0);
- θ_j representa o grau de usabilidade (traço latente) do *j*-ésimo *site* com distribuição normal N(0;1); e
- $P(U_{ij} = 1 / \theta_j)$ é a probabilidade de um *site j* com grau de usabilidade θ_j responder corretamente o item *i* e é chamada de função de resposta do item (FRI).

Dentro do contexto da aplicação desse trabalho, não são indivíduos que estão sendo avaliados mas "*sites* de *e-commerce*". Dessa forma, θ_j representa o grau de usabilidade do *j*-ésimo site, U_{ij} é a resposta

do *site j* para o item *i*, que pode ser "sim" (1) ou "não" (0), e $P(U_{ij} = 1 / \theta_j)$ é a probabilidade de um *site j* com grau de usabilidade θ_j responder possuir a característica descrita no item *i*.

A relação entre a resposta prevista ao item e o traço latente do indivíduo é definida pela curva característica do item (CCI) (RECKASE, 1997). A CCI, exemplificada na Figura 1, representa a regressão não linear de probabilidade de uma determinada resposta (eixo y) em função da habilidade (eixo x) (SANTOR; RAMSAY; ZUROFF, 1994), que, no presente trabalho, é o grau de usabilidade.

A maioria das aplicações da TRI assume unidimensionalidade do construto, o que significa que todos os itens estão medindo apenas uma dimensão ou traço latente, no caso do presente trabalho, o grau de usabilidade de *sites* de *e-commerce*. Todos os modelos da TRI assumem independência local, ou seja, as respostas dadas aos itens são independentes entre si (RECKASE, 1997).

O processo de estimação na TRI ocorre em duas etapas. Primeiro estimam-se os parâmetros dos itens, supondo uma distribuição de probabilidade para o traço latente (grau de usabilidade), e, depois, estima-se o grau de usabilidade supondo-se os parâmetros dos itens com valores conhecidos, obtidos na primeira etapa. É comum utilizar-se o método da máxima verossimilhança marginal para a estimação dos parâmetros dos itens. Para a estimação dos parâmetros θ_j, utiliza-se o método da máxima verossimilhança ou métodos bayesianos, tais como o da esperança a posteriori (EAP) ou da moda a posteriori (MAP). Esses métodos são discutidos com mais detalhes em Andrade, Tavares e Valle (2000).

4. Testes adaptativos informatizados

Os testes adaptativos informatizados (TAI) são testes adaptativos administrados via computador. Testes adaptativos (TA) são testes cuja aplicação dos itens se adapta ao indivíduo que está respondendo ao teste. Dessa forma, cada questão é apresentada isoladamente ao indivíduo e, de acordo com a sua resposta, uma próxima questão é selecionada no banco de itens do teste para ser administrada a ele. A lógica da seleção dos itens no teste, em geral, acontece da seguinte forma: se o indivíduo acerta o item atual, o próximo item deverá ser de um nível mais difícil; se o indivíduo erra o item atual, o próximo item deverá ser de um nível mais fácil. O objetivo do TAI é apresentar itens ao indivíduo que sejam adequados ao seu nível de habilidade, de tal forma que não seja necessário aplicar todos os itens existentes. A consequência disso é uma estimativa mais precisa da proficiência com menos itens aplicados e em menos tempo do que nos testes convencionais onde todos os indivíduos devem responder todas as questões de um mesmo teste (FETZER et al., 2008; SANDS; WATERS, 1997; VAN DER LINDEN; GLAS, 2000; WAINER, 2000b).

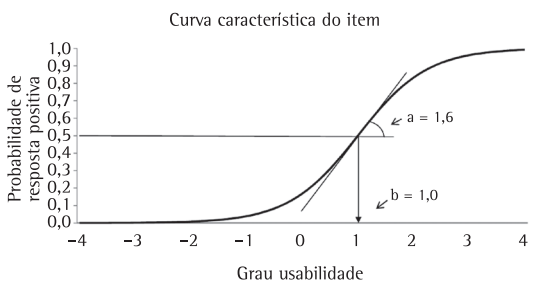

Figura 1. Curva característica de um item hipotético.

Segundo Olea et al. (2004), a maioria dos algoritmos de TAI utiliza uma estratégia que necessita estabelecer:

a) Um critério de partida, para determinar o primeiro item a ser apresentado. Geralmente se estabelece um valor para a habilidade inicial e administra-se um ou mais itens próximos a esse. Em bancos com muitos itens, também é comum selecionar itens mais informativos, onde essa informação está diretamente relacionada com o valor do parâmetro *a*, no ML2. Quando não existe uma informação prévia da habilidade da população, é comum adotar um valor mediano para a mesma, ou seja, valor zero no caso da escala (0,1);

b) Um método estatístico para estimar a proficiência do indivíduo e a precisão associada. Os métodos usualmente utilizados são o da máxima verossimilhança e os bayesianos (esperança ou moda a posteriori);

c) Um procedimento para selecionar o próximo item. Existem vários métodos para a seleção de itens, sendo que os mais utilizados são aqueles que combinam a proximidade da posição do item com a habilidade atual estimada (determinada pelo parâmetro *b*) e a informação do item (determinada pelo parâmetro *a*, no ML2). Em testes onde há a preocupação em não tornar os itens conhecidos para a população é comum utilizar um método para controlar a exposição do item. Nos testes cujos itens estão agrupados por conteúdos, onde há a necessidade de administrar itens relacionados a todos os conteúdos, utiliza-se o critério do balanceamento de conteúdo; e

d) Um critério de parada, para finalizar o teste. Geralmente é definido um número fixo de itens ou um valor mínimo para o erro padrão da habilidade ou uma combinação entre esses.

A lógica dos TAI é apresentada na Figura 2.

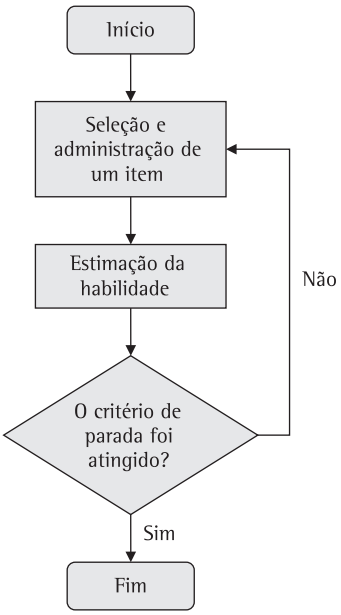

Figura 2. A lógica dos TAI.

O desenvolvimento de um TAI é um processo trabalhoso e exige conhecimentos e técnicas importantes. Em primeiro lugar é necessário haver um banco de itens devidamente calibrado na escala a ser utilizada. Em segundo lugar deve-se programar um conjunto de algoritmos para a seleção progressiva dos itens, para a estimação dos níveis de habilidade e sua respectiva precisão. Em terceiro lugar, um TAI deve submeter-se aos testes diagnósticos para garantir as propriedades desejáveis das estimações, assim como sua precisão e validade (OLEA et al., 2004).

A eficiência de um TAI pode ser verificada através de estudos empíricos ou simulação, em relação ao seguintes aspectos (MUÑIZ; HAMBLETON, 1999):

- Erro padrão médio (EPM): Trata-se da média do erro padrão das habilidades estimadas. Quanto menor for o valor, melhor é a precisão. O erro padrão da habilidade é a raiz quadrada do inverso da soma das informações fornecidas por cada item administrado.
- Raiz quadrada do erro quadrado médio (RQEQM): Calcula-se o quadrado da diferença entre a habilidade estimada e o valor do seu parâmetro (obtido por simulação), somam-se todos os resultados, divide-se pela quantidade de habilidades estimadas e extrai-se a raiz quadrada. Quanto menor for esse valor, melhor será a precisão;
- Desvio empírico (DE): Calcula-se a média dos valores absolutos das diferenças entre a habilidade estimada e o valor do seu parâmetro. Quanto menor for o valor, melhor é a precisão;
- Eficiência: Define-se pela quantidade média de itens necessários para alcançar um determinado erro padrão pré-determinado utilizado como critério de parada em testes de tamanho variável. Pelo teste, mais eficiente é aquele que consegue atingir o erro padrão pré-determinado com a menor quantidade de itens aplicados;
- Correlação linear: Calcula-se a correlação entre as estimativas das proficiências e o seu parâmetro. Quanto maior for o valor, maior será a precisão; e
- Procedimentos da TCM (teoria clássica da medida): Outros procedimentos da TCM podem ser utilizados; por exemplo, os coeficientes alfa de Crombach, teste-reteste, formas paralelas, entre outros (CROMBACH; GLESER, 1957; COHEN, 1960).

A implantação de um TAI pode trazer muitas vantagens para a avaliação (FAYERS; MACHIN, 2007; FETZER et al., 2008; OLEA et al., 2004; SANDS; WATERS, 1997; SUKAMOLSON, 2002; WAINER, 2000a, b). Entre elas, destacam-se:

- Redução do tempo de correção dos testes e consequente diminuição da ocorrência de erros nesse processo;
- Monitoramento do teste e controle do tempo de exposição do item;
- Redução do tempo de teste e do tamanho do teste;
- Redução do desgaste do indivíduo;
- Aplicação do teste nas ferramentas de educação à distância, através da internet;
- Melhoria na segurança do teste;
- Não necessita fazer impressão de provas e, consequentemente, prescinde de espaço físico e sigilo para o armazenamento delas;
- Pode fornecer o resultado imediatamente após o término do teste;
- Permite a manutenção do banco de itens, adição de novos itens e remoção de itens que se tornaram obsoletos;
- Pode produzir estimativas mais precisas das habilidades dos indivíduos;
- Permite a criação de itens em formatos multimídia, o que o torna mais atrativo do que os testes tradicionais; e
- Permite comparar os resultados entre indivíduos que respondem a diferentes itens.

Muitas avaliações internacionais conhecidas têm sido baseadas nos TAI (FETZER et al., 2008; OLEA et al., 2004; TEJADA, 2001). Entre elas, merecem destaque: o Test of English Foreign Language (TOEFL), o Graduate Record Exam (GRE), o Armed Services Vocational Aptitude Battery (ASVAB), o Scholastic Aptitude Tests I: Reasoning Test (SAT I), o Graduate Management Admission Test (GMAT), o National

Council of Architectural Registration Boards (NCARB), o National Council Licensure Examination for Registered Nurses (NCLEX), o Microsoft© Certified Professional Exams e o American Institute of Certified Public Accountants Exam (AICPA).

5. Procedimentos metodológicos

O banco de itens utilizado foi constituído de 32 itens testados e calibrados em Tezza, Bornia e Andrade (2011) através de um modelo logístico de 2 parâmetros (ML2), por meio do *software* BILOG-MG (TOIT, 2003), com uma amostra de 361 *sites* de *e-commerce*, estimados através do método da máxima verossimilhança marginal com distribuição normal (0;1) para o grau de usabilidade . Os itens foram elaborados de forma dicotômica, onde o *site* recebia o valor 1 se apresentasse o atributo relacionado à sua usabilidade, ou o valor 0 caso contrário. Nos casos em que o item não se aplicava, era atribuído o valor 9. O Tabela 1. apresenta os 32 itens que compõem o banco de itens com os respectivos parâmetros, os quais foram considerados como conhecidos neste trabalho. Os parâmetros estão representados na escala N (0;1) comumente utilizada pelos programas computacionais para calibração dos itens.

Observa-se que a maioria dos itens possui valor negativo para o parâmetro b, o que significa que esses itens são "fáceis". Dentro do contexto estudado, por itens fáceis entende-se que são atributos mais fáceis de o *site* possuir, enquanto difíceis são atributos mais difíceis de o *site* possuir. Os autores utilizaram como critério para manutenção do item no banco valores superiores a 0,70 para o parâmetro a.

Tabela 1. Itens do banco e respectivos parâmetros estimados.

Item	Descrição	Parâmetro	
		a	b
1	Homepage deixa claro o que o *site* faz sem precisar usar a rolagem?	0,76	-1,54
2	As imagens, botões ou palavras clicáveis apresentam uma forma diferenciada quando são selecionadas?	0,89	-1,46
3	As opções principais do *site* estão visíveis?	0,83	-2,93
4	A disposição dos objetos de interação das caixas de diálogo segue uma ordem lógica?	0,79	-0,90
5	Os rótulos de campos começam com letra maiúscula e as letras restantes são minúsculas?	0,80	-1,56
6	O *site* possui opção de acesso com outras línguas?	0,78	4,67
7	Títulos estão alinhados à esquerda?	1,05	-2,29
8	Parágrafos de texto são separados?	0,94	-2,91
9	As palavras aparentemente clicáveis são de fato clicáveis?	0,97	0,18
10	Os títulos de telas, janelas e caixas de diálogo estão no alto, centrados ou justificados à esquerda?	1,43	-2,49
11	Todas as páginas possuem um campo de busca?	0,77	-2,66
12	Os resultados de busca permitem classificação por outros critérios além de custo?	1,30	0,45
13	Listas longas apresentam indicadores de continuação, de quantidade de itens e de páginas?	0,75	-0,28
14	O preço de um produto consta ao lado da imagem ou do *link* do produto?	1,18	-2,54
15	Existe uma orientação ao usuário quanto ao restante do *site*?	0,92	0,18
16	A maioria dos produtos possui informações sobre eles?	1,16	-1,94
17	É possível ampliar as fotos dos produtos para visualizar detalhes?	0,98	-0,95
18	Em produtos em que existem mais de uma perspectiva, é possível visualizar todas as perspectivas?	1,09	1,68
19	Os grupos de botões de comando estão dispostos em coluna e à direita, ou em linha e abaixo dos objetos aos quais estão associados?	0,94	-2,84
20	Quando há rolagem, não existem elementos de *design* que pareçam com marcadores de final de página?	1,11	-2,86
21	Todos os campos e mostradores de dados possuem rótulos identificativos?	1,05	-2,14
22	O botão de finalização de compra está no final da lista?	0,75	-4,48
23	É possível saber o custo total antes de fazer cadastro?	0,87	-2,04
24	No preenchimento de um formulário é informado a forma de preenchimento?	1,09	-0,53
25	Os dados obrigatórios são diferenciados dos dados opcionais de forma visualmente clara?	0,76	-1,14
26	As mensagens de erro estão isentas de abreviaturas e/ou códigos gerados pelo sistema operacional?	1,04	-2,95
27	Todas as páginas possuem o mesmo *layout* e exibem ao usuário as mesmas características?	1,28	-1,63
28	O logotipo da empresa está no canto superior esquerdo em todas as páginas do *site*?	1,55	-1,74
29	Existe um *link* de um único clique que conduz a *homepage*?	1,28	-2,07
30	Qualquer ação do usuário pode ser revertida através da opção DESFAZER?	1,44	-1,20
31	O *site* permite navegação em suas páginas em apenas uma janela, ou seja, não há abertura de novas janelas em meio a navegação?	1,11	-1,69
32	Os *links* já visitados mudam de cor?	0,69	4,41

Os algoritmos para o teste adaptativo informatizado foram elaborados no Excel®. Utilizou-se o método da máxima verossimilhança para a estimação do grau de usabilidade, o qual consiste em encontrar o valor máximo da função verossimilhança, ou seja, a cada item respondido era calculado o valor da função verossimilhança para vários pontos do domínio da variável latente (considerado entre –4 e 4) e identificado o ponto de máximo da função. O método MV possui um problema de indeterminação no caso de respostas constantes, ou seja, enquanto as respostas aos itens forem todas as mesmas (todas positivas ou todas negativas), o método MV não chega a um valor máximo de função. Para contornar esse problema nesse TAI foi utilizado o critério de Herrando (1989), o qual sugere que se as respostas forem todas positivas seja atribuído o valor 4 para a habilidade e que se as respostas forem todas negativas esse valor atribuído seja –4.

A habilidade inicial estimada foi considerada média, no caso da escala utilizada (0,1), igual a zero. Esse procedimento é usual nos TAI quando não existe uma informação prévia sobre a habilidade inicial.

Foram elaborados cinco algoritmos para o teste, denominados TAI1, TAI2, TAI3, TAI4 e TAI5. Esses algoritmos diferem entre si segundo o critério de seleção do(s) primeiro(s) item(ns) do teste. Esses critérios não estão descritos na literatura mas foram desenvolvidos baseados no critério de máxima informação (LORD, 1977), o qual consiste em selecionar o item mais próximo e com a maior informação. Para a seleção do primeiro item, foram testados (via simulação) os seguintes critérios:

TAI1 - O item mais próximo da habilidade inicial. Quanto mais próximo o item for da habilidade, mais adequado ele é. O problema desse critério é que a próxima estimativa da habilidade será 4, se o *site* tiver a característica, ou -4, se o *site* não tiver;

TAI2 - Os três itens mais próximos da habilidade inicial. Esse critério é uma extensão do critério do TAI1 que tenta resolver o problema da estimação por MV causada pelo padrão de respostas constante;

TAI3 - O item mais próximo da habilidade inicial, o item mais fácil e o item mais difícil. Essa é outra alternativa para tentar resolver o problema da estimação, de tal forma que se espera que a grande maioria dos respondentes deva responder positivamente para o item mais fácil e negativamente para o item mais difícil, evitando padrão de respostas constante;

TAI4 - O item com a maior informação. Quanto maior for a informação do item, maior será o seu poder de discriminação, o que pode contribuir para fornecer boas estimativas da habilidade no início do TAI. Por outro lado, esse critério apresenta o mesmo problema do critério do TAI1; e

TAI5 - Os três itens com maior informação. Esse critério é uma extensão do critério do TAI4 que tenta resolver o problema da estimação por MV causada pelo padrão de respostas constante.

Para a administração dos demais itens, utilizou-se o critério que seleciona o item mais próximo da habilidade atual estimada (LORD, 1977). Não foram utilizados critérios mais sofisticados devido à pouca quantidade de itens no banco de itens. Também não houve a necessidade de controlar a exposição dos itens pelo fato de eles não serem sigilosos e também devido à pouca quantidade de itens no banco. Neste construto não foi utilizado o critério de balanceamento de conteúdo porque os itens não estavam divididos em conteúdos.

O critério de parada utilizado foi a aplicação de um número fixo de itens. Entretanto, esse número fixo foi definido por meio de análise das simulações, segundo a estabilização do erro padrão médio das habilidades. Não foi utilizado um número variável de itens como critério de parada pelo fato de que alguns testes poderiam ficar com muitos itens, o que perderia o sentido uma vez que o objetivo do teste adaptativo é diminuir a quantidade de itens aplicados.

Os algoritmos foram avaliados através de simulações. Foram simuladas as respostas de mil *sites* com habilidade simulada na escala (0,1) para cada um dos 32 itens. Os critérios utilizados para a avaliação foram: o erro padrão médio das habilidades (EPM), a raiz quadrada do erro quadrado médio (RQEQM) e a correlação linear (CL). As simulações foram realizadas no Excel®.

O algoritmo selecionado, segundo os resultados das simulações, foi aplicado aos dados reais de respostas de 361 *sites*.

6. Resultados das simulações

O primeiro teste analisado (TAI1) utilizou o primeiro item que era mais próximo da habilidade média (zero), que é o item 9, cujo parâmetro b é igual a 0,18. Nesse teste, todos os *sites* foram primeiramente submetidos a esse item: "As palavras aparentemente clicáveis são de fato clicáveis?" O *site* que fosse avaliado positivamente nesse item tinha a sua habilidade estimada igual a 4 e o próximo item a ser respondido (mais próximo do valor 4) era o 32: "Os *links* já visitados mudam de cor?" Por outro lado, o *site* que fosse avaliado negativamente nesse item tinha a sua habilidade estimada igual a –4 e o próximo item a ser respondido (mais próximo do valor –4) era o 22: "O botão de finalização de compra está no final da lista?"

No segundo teste (TAI2), todos os *sites* foram submetidos aos três itens mais próximos da habilidade inicial, no caso os itens 9 ("As palavras aparentemente clicáveis são de fato clicáveis?"), 13 ("Listas longas apresentam indicadores de continuação, de quantidade de itens e de páginas?") e 15 ("Existe uma orientação ao usuário quanto ao restante do *site*?").

No terceiro teste (TAI3), todos os *sites* foram submetidos ao item mais próximo da habilidade inicial (9, "As palavras aparentemente clicáveis são de fato clicáveis?"), o item mais fácil (22, "O botão de finalização de compra está no final da lista?") e o item mais difícil (6, "O *site* possui opção de acesso com outras línguas?").

No quarto teste (TAI4), todos os *sites* foram submetidos ao item com maior informação, ou seja, o item 28: "O logotipo da empresa está no canto superior esquerdo em todas as páginas do site?" Na administração do segundo item ocorre o mesmo que aconteceu no TAI1.

No quinto teste (TAI5), todos os *sites* foram submetidos aos três itens com maior informação, no caso os itens 28 ("O logotipo da empresa está no canto superior esquerdo em todas as páginas do site?"), 30 ("Qualquer ação do usuário pode ser revertida através da opção DESFAZER?") e 10 ("Os títulos de telas, janelas e caixas de diálogo estão no alto, centrados ou justificados à esquerda?").

A Figura 3 apresenta o erro padrão médio das habilidades estimadas (EPM) segundo a quantidade de itens administrados para cada TAI.

Observa-se que o TAI4 foi o teste que apresentou o menor EPM nos primeiros itens administrados (a partir do terceiro) até o sexto. A partir do sétimo item, o menor EPM foi apresentado pelo TAI2, entretanto, a partir do 14º item administrado, o EPM é praticamente igual para todos os TAI, sendo que a diferença entre eles não passa de 0,03. Os maiores EPM registrados foram o do TAI5 entre terceiro e quinto itens e o do TAI3 a partir do sexto item.

O gráfico da Figura 4 apresenta a raiz quadrada do erro quadrado médio das habilidades estimadas (RQEQM) segundo a quantidade de itens administrados para cada TAI.

Observa-se que os menores RQEQM são obtidos por TAI4 entre quarto e nono itens, TAI1 no décimo item, TAI4 novamente no 11º item, e por TAI5 de 12º a 17º item, com exceção do 14º, obtido por TAI3. A partir da aplicação do 14º, os valores da RQEQM são praticamente iguais para todos os TAI, sendo que a diferença entre eles não passa de 0,02. Os maiores valores foram obtidos por TAI2 no quarto, 11º, 12º e 14º itens, por TAI5 no quinto e no sexto item, por TAI3 do sétimo ao décimo item e a partir do 15º item, e por TAI4 no 13º item.

Considerando os resultados obtidos nos gráficos das Figuras 3 e 4, decidiu-se limitar o teste à quantidade fixa de 13 itens uma vez que as diferenças

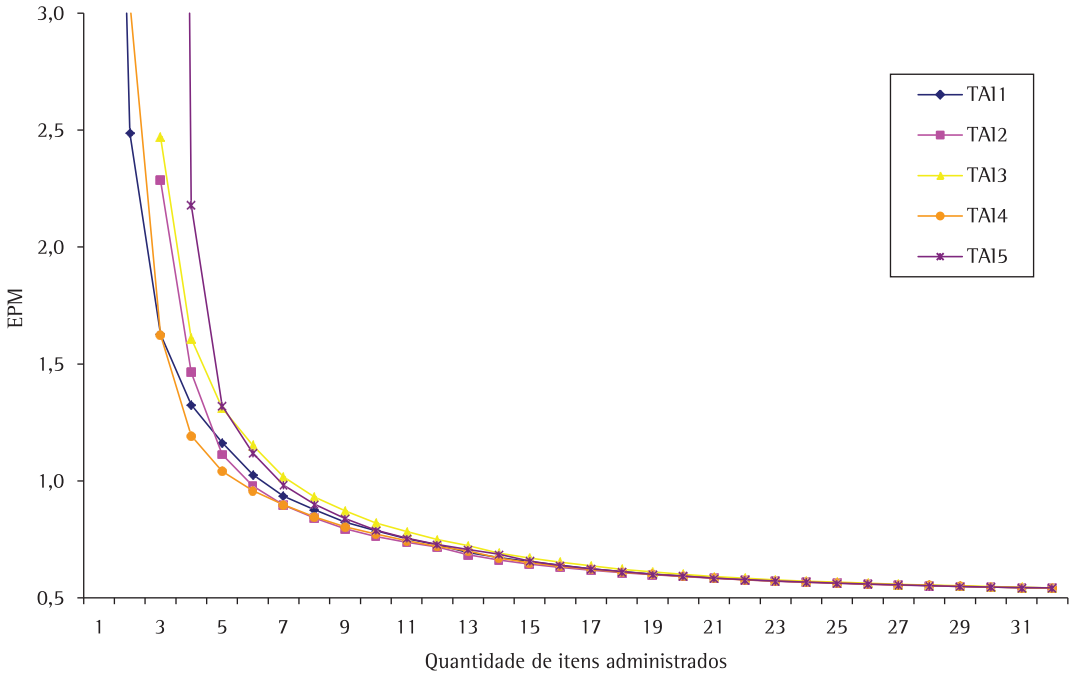

Figura 3. Erro padrão médio das habilidades estimadas.

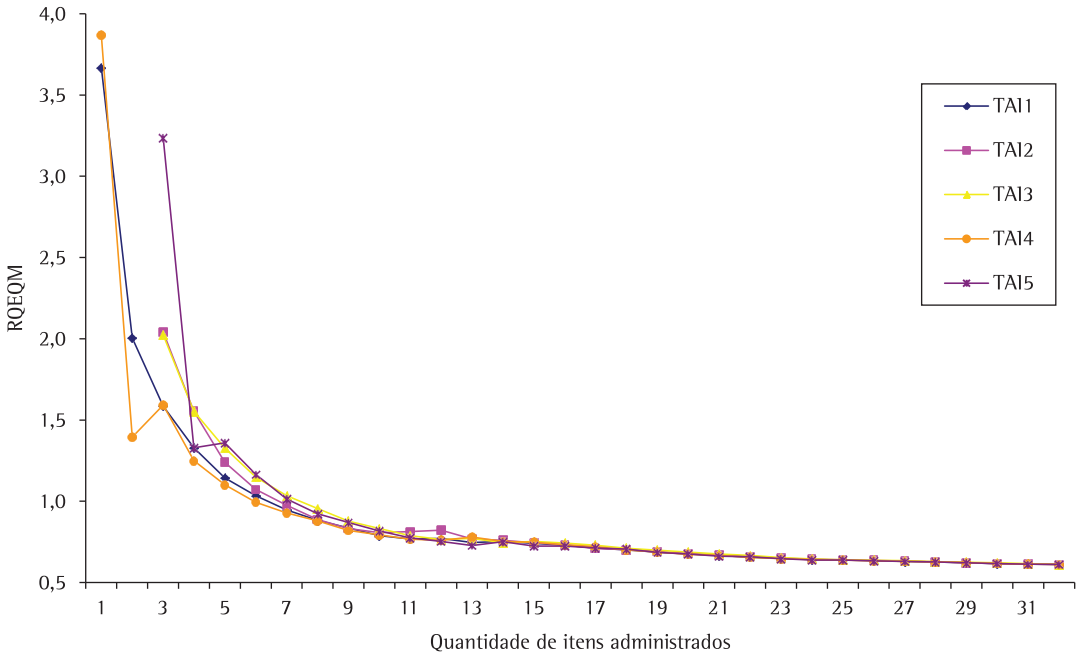

Figura 4. Raiz quadrada do erro quadrado médio.

das estimativas do EPM e do RQEQM passam a ser pequenas a partir da aplicação do 14º item. Dessa forma, analisando-se conjuntamente o EPM e o RQEQM, optou-se por escolher o TAI5 que, além de apresentar o menor valor de RQEQM em relação aos demais testes, administra itens mais próximos da habilidade atual estimada até o 13º item. Portanto, o algoritmo para o teste adaptativo para medir o grau de usabilidade foi definido conforme mostra a Figura 5.

7. Aplicação do algoritmo escolhido

Definido o teste, ele foi aplicado às respostas reais de 361 *sites*, da seguinte forma: (1) o *site* é selecionado; (2) observa-se as respostas do *site* aos itens 10, 28 e 30; (3) estima-se a habilidade do *site*; (4) verifica-se se o critério de parada foi atingido (aplicação de 13 itens); se sim, finaliza-se o teste, caso contrário seleciona-se o item seguinte mais próximo da habilidade estimada. Repetem-se os passos (3) e (4) até o final do teste.

A Figura 6 apresenta o gráfico de dispersão entre as habilidades estimadas pelo TAI e as habilidades estimadas dos 361 *sites* obtidas pela aplicação dos 32 itens, bem como a reta ajustada aos dados. Observa-se que os valores estimados pelo TAI são ligeiramente inferiores às estimativas considerando-se todos ou itens do banco. O coeficiente de determinação foi 0,87 e o coeficiente de correlação foi 0,93. Pode-se observar ainda nesta figura uma linha tracejada que representa a correlação teoricamente perfeita, para efeitos de comparação.

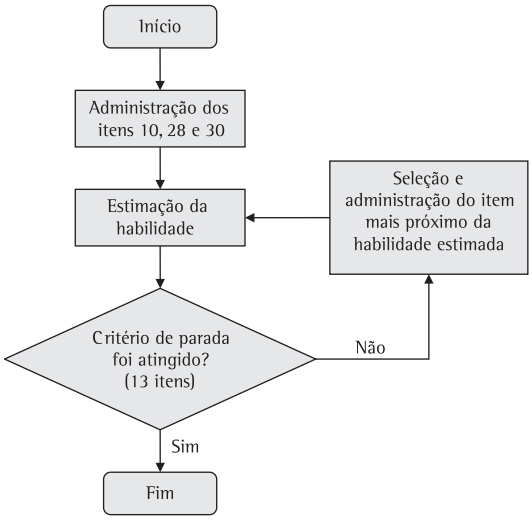

Figura 5. Algoritmo do TAI para medir o grau de usabilidade.

A Tabela 2 apresenta a habilidade estimada pelo TAI com seu erro padrão (EP) e a habilidade estimada com os 32 itens com seu erro padrão para 10 *sites* analisados selecionados aleatoriamente.

Na Tabela 2 observa-se que os EP das habilidades estimadas pelo TAI são maiores do que os EP das habilidades estimadas com os 32 itens, como esperado, já que o TAI utiliza um subconjunto do banco de itens. Entretanto, os valores não são tão diferentes, em geral. Considerando todos os 361 *sites*, o EPM do TAI foi 0,71, enquanto que o EPM com a aplicação de todos os *itens* foi 0,54.

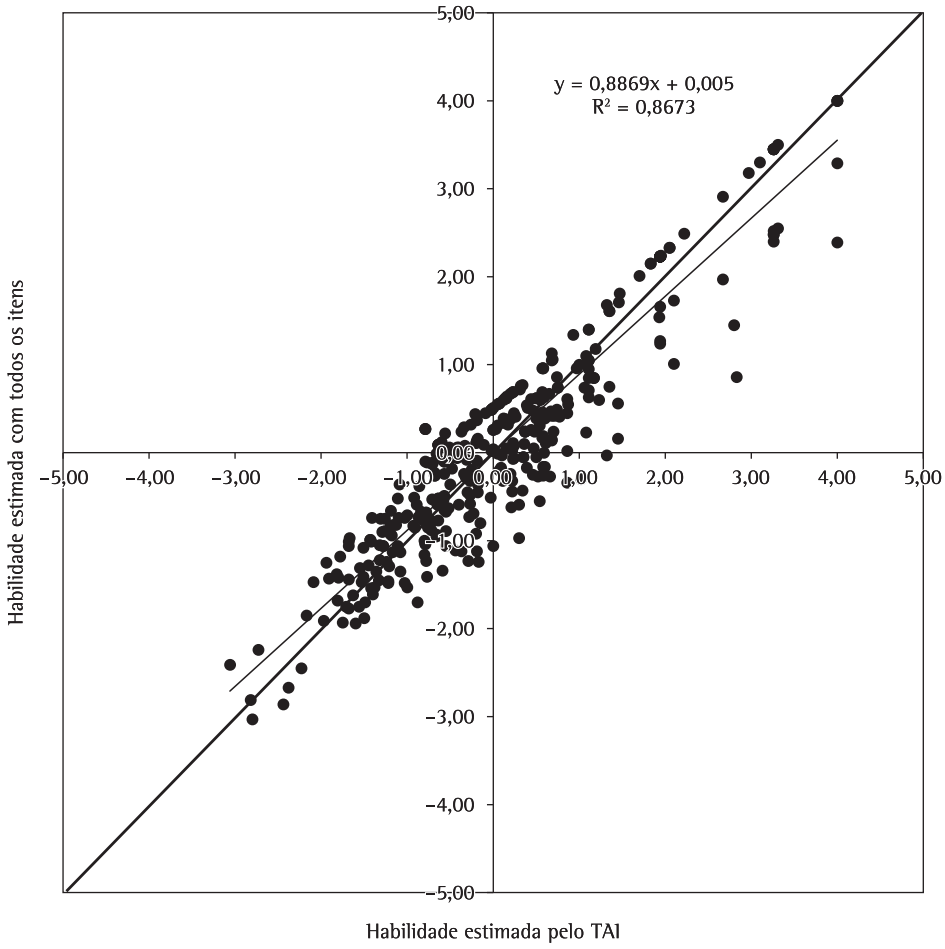

Figura 6. Gráfico de dispersão.

Tabela 2. Grau de usabilidade estimado, EP.

	TAI		Aplicação com os 32 itens	
Site	Habilidade	EP	Habilidade	EP
1	−0,40	0,63	−0,59	0,42
2	−1,67	0,49	−0,97	0,42
3	−0,81	0,63	−0,77	0,44
4	−1,40	0,61	−1,61	0,40
5	−1,80	0,49	−1,42	0,40
6	−0,34	0,65	−0,24	0,49
7	1,08	0,78	0,23	0,53
8	−2,80	0,55	−3,03	0,45
9	2,83	1,18	0,86	0,62
10	−0,79	0,59	−1,11	0,50

Na Tabela 2 destaca-se o resultado obtido para o *site* 9, muito diferente do obtido com aplicação do banco completo., Esse grande erro deve-se possivelmente ao fato de o grau de usabilidade do *site* situar-se na região superior da escala, na qual existem poucos itens, conforme pode ser observado no Tabela 1.

A Tabela 3 acompanha passo a passo como foi a aplicação do TAI para o primeiro site avaliado, apresentando a diferença entre a habilidade estimada e a posição do item, o item selecionado, a resposta dada, a habilidade atual estimada e o seu erro padrão.

Através da Tabela 3 pode-se acompanhar o andamento do teste. Nos três primeiros passos são administrados os itens 10, 28 e 30. Embora eles sejam aplicados um de cada vez, a habilidade e seu erro padrão são estimados somente após a aplicação de todos eles, já que esses três são considerados itens iniciais. Com base nas respostas, utilizando-se o método da máxima verossimilhança, a habilidade é estimada em −1,25. Então observa-se o item mais próximo, no caso o 25 (b = −1,14), que é aplicado no teste. O indivíduo responde e a habilidade e o erro padrão são novamente estimados, até que seja aplicado o 13º item. Nota-se que toda vez que a resposta é "sim" o valor da habilidade estimada aumenta em relação à anterior, e toda vez que a resposta é "não"

Tabela 3. O TAI passo a passo (*site* 1).

Passo	Diferença	Item selecionado	Resposta	Habilidade	EP
1	-	10	sim	-	-
2	-	28	sim	-	-
3	-	30	não	−1,25	0,88
4	0,11	25	não	−1,49	0,81
5	0,03	2	sim	−1,22	0,79
6	0,27	17	não	−1,45	0,72
7	0,09	1	sim	−1,27	0,70
8	0,29	5	não	−1,47	0,67
9	0,16	27	sim	−1,25	0,63
10	0,35	4	sim	−1,08	0,62
11	0,55	24	sim	−0,82	0,62
12	0,54	13	sim	−0,65	0,62
13	0,83	9	sim	−0,40	0,63

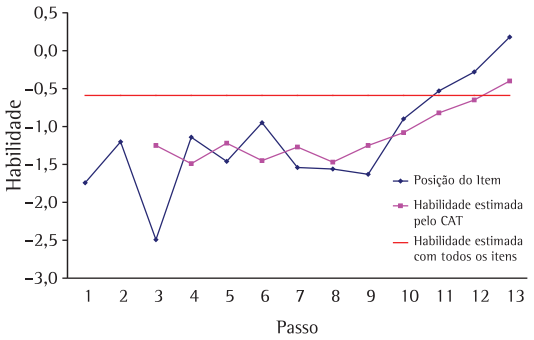

Figura 7. Posição do item aplicado e grau de usabilidade estimado a cada passo.

o valor diminui. A Figura 7 apresenta a habilidade estimada e a posição (parâmetro b) na escala dos itens aplicados em cada passo do TAI.

Nessa aplicação do TAI, além dos itens 10, 28 e 30, que foram aplicados obrigatoriamente em todos os testes, os itens mais aplicados foram: 17, em 90,3% dos testes, 4, em 89,8% e 24, em 88,1%. Esses itens podem ser considerados levemente fáceis ou mesmo medianos, já que se encontram na faixa entre −0,5 e −1 na escala. Por outro lado, os menos aplicados foram os itens 22, com 0,6% de aplicações, o 3, com 1,1%, e o 26, com 1,4%, que são os itens mais fáceis do banco.

8. Considerações finais

Neste trabalho foi desenvolvido um algoritmo para um teste adaptativo informatizado baseado na teoria da resposta ao item para estimar o grau de usabilidade de *sites* de *e-commerce*. O banco de itens utilizado continha um total de 32 itens e foi calibrado em estudo anterior, segundo o modelo logístico de 2 parâmetros (ML2) da TRI.

Cinco diferentes algoritmos foram analisados para estabelecer qual seria o mais adequado para estimar a usabilidade. Os testes foram analisados através da simulação de mil respondentes e os critérios utilizados para a seleção do teste foram o erro padrão médio das estimativas das habilidades (EPM) e a raiz quadrada do erro quadrado médio (RQEQM). Todas as análises, simulações e aplicações foram feitas no *software* Excel®.

O teste selecionado (TAI5) foi composto com a seguinte configuração:

- Item inicial: foram selecionados os três itens com a maior informação como itens iniciais;
- Método de estimação da habilidade: máxima verossimilhança;
- Método de seleção de itens: o mais próximo da habilidade atual estimada; e
- Critério de parada do teste: número fixo de itens igual a 13.

A pequena quantidade de itens no banco, a ausência do balanceamento de conteúdo e a não necessidade de controlar a exposição do item tornaram o teste mais simples de ser desenvolvido e aplicado, uma vez que não foi necessária a elaboração de um algoritmo mais complexo. O algoritmo escolhido, segundo os resultados obtidos na simulação, foi aplicado aos dados reais de 361 *sites*. O EPM do TAI foi 0,71, enquanto que o EPM com a aplicação de todos os 32 itens foi 0,54. A correlação linear entre as habilidades estimadas pelo algoritmo do TAI com 13 itens e as habilidades reais dos 361 *sites* obtidas pela aplicação dos 32 itens foi de 0,93. Esse é um indicativo de que o algoritmo TAI possui um bom desempenho para estimar o grau de usabilidade dos *sites* de *e-commerce*.

Ressalta-se que os resultados apresentados devem ser vistos como uma demonstração da potencialidade do uso de um TAI. Para uma aplicação real faz-se necessário um número bastante grande de itens e que cubram uma amplitude maior dos níveis da escala de grau de usabilidade, fato esse que não ocorreu no presente trabalho. O critério de parada deve ser escolhido para cada aplicação, segundo as especificidades de cada situação e das características dos bancos de itens. O critério fixo de 13 itens utilizado neste trabalho, juntamente com a pequena quantidade de itens no banco, resultou em grandes erros padrões das estimativas.

Para trabalhos futuros recomenda-se a inclusão de mais itens para compor o banco de itens, principalmente itens com maior nível de dificuldade, para se obter um melhor desempenho do teste, uma vez que no banco atual existem mais itens fáceis (com parâmetro b negativo) do que difíceis (com parâmetro b positivo).

Referências

AGARWAL, R.; VENKATESH, V. Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability. *Information Systems Research*, v. 13, n. 2, p. 168-186, June 2002. http://dx.doi.org/10.1287/isre.13.2.168.84

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da resposta ao item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística - ABE, 2000.

DE AYALA, R. J. *The Theory and Practice of Item Response Theory*. New York: The Guilford Press, Wiley, 2009.

BIRNBAUM, A. Some Latent Trait Models and Their Use in Infering an Examinee's Ability. In: LORD, F. M.; NOVICK, M. R. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley, 1968.

CHEVALIER, A.; BONNARDEL, N. Articulation of web site design constraints: Effects of the task and designers' expertise. *Computers in Human Behavior*, v. 23, n. 5, p. 2455-2472, 2007. http://dx.doi.org/10.1016/j.chb.2006.04.001

CHOE, P. et al. Evaluating and improving a self-help technical support Web site: Use of focus group interviews. *International Journal of Human-Computer Interaction*, v. 21, n. 3, p. 333-354, 2006. http://dx.doi.org/10.1207/s15327590ijhc2103_4

COHEN, J. A coefficient of agreement for nominal scales. *Educacional and Psychological Measurement*, v. 20, n. 1, p. 37-46, 1960. http://dx.doi.org/10.1177/001316446002000104

CROMBACH, L. J.; GLESER, G. C. *Psychological Tests and Personal Decisions*. Urbana: University of Illinois Press, 1957.

CYBIS, W. *Ergonomia e Usabilidade*: conhecimentos, métodos e aplicações. São Paulo: Novatec Editora, 2007.

EMBRETSON, S.; REISE, S. P. *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc. Publishers, 2000.

FANG, X.; HOLSAPPLE, C. W. An empirical study of web site navigation structures' impacts on web site usability. *Decision Support Systems*, v. 43, n. 2, p. 476-491, 2007. http://dx.doi.org/10.1016/j.dss.2006.11.004

FAYERS, P. M.; MACHIN, D. *Quality of Life:* The Assessment, Analysis and Interpretation of Patient-reported Outcomes. 2nd ed. Wiley, 2007.

FETZER, M. et al. *Computer Adaptive Testing* (*CAT*) *in an Employment Context*. Roswell: PreVisor, 2008. White paper.

HAMBLETON, R. K. Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care*, v. 38, n. 9, p. 60-65, 2000. Supplement II.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamentals of item response theory*. Newbury Park: Sage, 1991.

HERRANDO, S. Tests adaptativos computerizados: una sencilla solución al problema de la estimación con puntuaciones perfectas y cero. In: CONFERENCIA ESPAÑOLA DE BIOMETRÍA, 2., 1989, Segovia, Espanha. *Anales*... Segovia: Biometric Society, 1989.

IVORY, M. Y.; MEGRAW, R. Evolution of web site design patterns. *ACM Transactions on Information Systems*, v. 23, n. 4, p. 463-497, 2005. http://dx.doi.org/10.1145/1095872.1095876

KIERAS, D. E.; POLSON, P. G. An approach to the formal analysis of user complexity. *International Journal of Human-Computer Studies*, v. 51, n. 2 p. 405-434, 1999. http://dx.doi.org/10.1006/ijhc.1983.0317

LARGE, A. et al. Web Portal Design Guidelines as Identified by Children through the Processes of Design and Evaluation. *Proceedings of the American Society for Information Science and Technology*, v. 43, n. 1, p. 1-23, 2006. http://dx.doi.org/10.1002/meet.1450430120

LAZAR, J.; MEISELWITZ, G.; NORCIO, A., A taxonomy of novice user perception of error on the web. *Universal Access in the Information Society Journal*, v. 3, n. 3-4, p. 202-208, 2004. http://dx.doi.org/10.1007/s10209-004-0095-9

LORD, F. M. A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, v. 1, n. 1, p. 95-100, 1977. http://dx.doi.org/10.1177/014662167700100115

MAGOUTAS, B. et al. An adaptive e-questionnaire for measuring user perceived portal quality. *International Journal of Human-Computer Studies*, v. 68, n. 10, p. 729-745, 2010. http://dx.doi.org/10.1016/j.ijhcs.2010.06.003

MOREIRA JUNIOR, F. J. Aplicações da Teoria da Resposta ao Item (TRI) no Brasil. *Revista Brasileira de Biometria*, v. 28, n. 4, p. 137-170, 2010.

MUÑIZ, J.; HAMBLETON, R. Evaluación psicométrica de los tests informatizados. In: NIELSEN, J. *Usability Engineering*. California: Morgan Kaufmann, 1993.

NIELSEN, J.; LORANGER, H. *Prioritizing Web Usability*. California: New Riders, 2006.

OLEA, J. et al. Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. *Psicothema*, v. 16, n. 3, p. 519-525, 2004.

RAU, P.; LIANG, S. F. Internationalization and localization: evaluating and testing a website for Asian users. *Ergonomics*, v. 46, n. 1, p. 255-270, 2003. http://dx.doi.org/10.1080/00140130303527

RECKASE, M. D. A linear logistic multidimensional model for dichotomous item response data. In: VAN DER LINDEN, W. J.; HAMBLETON, R. K. (Eds.). *Handbook of modern item response theory*. New York: Springer-Verlag, 1997. *p.* 271-286.

REISE, S. P.; WIDAMAN, K. F.; PUGH, R. H. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, v. 114, n. 3, p. 552-566, 1993. http://dx.doi.org/10.1037/0033-2909.114.3.552

ROSSO, M. User-Based Identification of Web Genres. *Journal of the American Society for Information Science and Technology*, v. 59, n. 5 p. 1-20, 2008.

SANDS, W. A.; WATERS, B. K. Introduction to ASVAB and CAT. In: SANDS, W. A; WATERS, B. K.; MCBRIDE, J. R. *Computerized Adaptive Testing*: from inquiry to operation. Washington: American Psychological Association, 1997. http://dx.doi.org/10.1037/10244-001

SANTOR, D. A.; RAMSAY, J. O.; ZUROFF, D. C. Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, v. 6, n. 3, p. 255-70, 1994. http://dx.doi.org/10.1037/1040-3590.6.3.255

SCHENKMAN, B. N.; JÖNSSON, F. U. Aesthetics and preferences of web pages. *Behaviour & Information Technology*, v. 19, n. 5, p. 367-377, 2000. http://dx.doi.org/10.1080/014492900750000063

SUKAMOLSON, S. *Computerized Test/Item Banking and Computerized Adaptive Testing for Teachers and Lecturers*. Information Technology and Universities in Asia – ITUA, 2002.

STRAUB, D. W. Validating instruments in MIS research. *MIS Quarterly*, v. 13, n. 2, p. 147-169, 1989. http://dx.doi.org/10.2307/248922

TAVARES, H. R.; ANDRADE, D. F.; PEREIRA, C. A. Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology*, v. 27, n. 4, p. 679-685, 2004. http://dx.doi.org/10.1590/S1415-47572004000400033

TEJADA, A. J. R. Pasado, presente y futuro de los Tests Adaptativos Informatizados: entrevista con Isaac I. Bejar. *Psicothema*, v. 13, n. 4, p. 685-690, 2001.

TEZZA, R.; BORNIA, A. C.; ANDRADE, D. F. Measuring web usability using item response theory: Principles, features and opportunities. *Interacting with Computers*, v. 23, n. 2, p. 167-175, 2011. http://dx.doi.org/10.1016/j.intcom.2011.02.004

TOIT, M. *IRT from SSI*: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Scientific Software International, 2003.

VAN DER LINDEN, W. J.; GLAS, C. A. W. *Computerized Adaptive Testing:* Theory and Practice. Dordrecht: Kluwer Academic, 2000.

WAINER, H. CATs: Whither and whence. *Psicológica*: revista de metodología y psicología experimental, v. 21, n. 1, p. 121-133, 2000a.

WAINER, H. *Computerized Adaptive Testing:* A Primer. New Jersey: Lawrence Erlbaum Associates, 2000b.

Algorithm of computerized adaptive testing to estimate the usability of *e-commerce sites*

Abstract

This paper proposes an algorithm of a computerized adaptive testing based on Item Response Theory, designed to estimate the degree of usability of *e-commerce sites*. Five algorithms were tested by simulation. The algorithm with the best performance was applied to real data from 361 *e-commerce sites*. The results showed that the algorithm could obtain good estimates for the degree of usability of *e-commerce sites* with the application of 13 items.

Keywords

Computerized adaptive testing. Item response theory. Usability. *e-commerce sites*.