Estimação da usabilidade de sites *e-commerce* pelo Método da Máxima Verossimilhança

Fernando de Jesus Moreira Junior[1], Rafael Tezza[2], Antonio Cezar Bornia[2]

[1]Departamento de Estatística/CCNE - UFSM; Santa Maria, RS
[2]Engenharia de Proução - UFSC; Florianópolis, SC
e-mail: fmjunior777@yahoo.com.br

Resumo

O presente artigo investiga o desempenho do método da Máxima Verossimilhança (MV) na estimação da usabilidade de sites *e-commerce* e compara o desempenho entre os Softwares BILOG-MG® e Excel® na estimação dessas usabilidades. Para isso, foram utilizados dados reais de um estudo sobre o grau de usabilidade de 361 sites de *e-commerce*, no qual foram aplicados 32 itens calibrados por meio do modelo logístico unidimensional de dois parâmetros (MLU2) da Teoria da Resposta ao Item (TRI). O processo de estimação da usabilidade por MV foi feito nos softwares BILOG-MG® e Excel®. Os resultados mostraram que o método de MV apresenta deficiências quando existe um padrão de resposta constante, o que pode ocorrer durante a aplicação dos primeiros itens do questionário. Entretanto, o método apresenta um bom desempenho quando o padrão de respostas não é constante. Além disso, o desempenho do processo elaborado no Excel® foi melhor do que no software convencional BILOG-MG®. Os parâmetros dos itens também influenciam na estimação por MV.

Abstract

This article investigates the performance of Maximum Likelihood (ML) method to estimate the usability of e-commerce sites and compares the performance between the software Bilog-MG ® and Excel ® in the estimation of these usability. For this, we used real data from a study on the degree of usability of 361 e-commerce sites, which were applied 32 items calibrated by the unidimensional logistic model of two parameters (MLU2) of Item Response Theory (IRT). The estimation process by ML was developed in BILOG® and Excel® softwares. The results showed that the ML method is flawed when there is a constant pattern of response, which may occur during application of the first items on the questionnaire. However, the method performs well when the pattern of responses is not constant. Moreover, process performance prepared in Excel® was better than in conventional software BILOG-MG®. The parameters of the items also influence the estimation of ML.

1. Introdução

A Teoria da Resposta ao Item (TRI) é um sistema de modelos matemáticos que define uma maneira de estabelecer a correspondência entre variáveis latentes e suas manifestações (AYALA, 2009) possibilitando a criação de medidas padronizadas. Variável latente é uma característica que não pode ser medida diretamente por meio de aparelho ou de uma única pergunta, mas por um conjunto de itens que estejam relacionados a essa variável. Alguns exemplos de variáveis latentes são: o nível de proficiência em matemática de um aluno do ensino fundamental, o grau de depressão de um paciente, o nível de satisfação de um cliente, o grau de usabilidade de um *website*. Atualmente, tanto no Brasil quanto no exterior, a Teoria da Resposta ao Item vem sendo bastante difundida, principalmente na área de educação e testes psicológicos. Uma relação dos trabalhos sobre TRI publicados no Brasil até o ano de 2009 se encontra disponível em Moreira Junior (2010).

Segundo 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 destes a um deter-

minado conjunto de itens previamente calibrados.

Sendo assim, o processo de estimação, tanto dos itens quanto da habilidade dos respondentes é uma etapa crucial no processo de aplicação da Teoria da Resposta ao Item. Vários estudos têm abordado os diversos métodos de estimação nessa teoria e ressaltado características peculiares de cada um deles por meio de simulações (AZEVEDO, 2003; 2008). Os principais métodos de estimação utilizados são o método da Máxima Verossimilhança (MV) e os métodos Bayesianos da Esperança a Posteriori (EAP) e da Moda a Posteriori (MAP).

O presente artigo tem por objetivo investigar o desempenho do método de Máxima Verossimilhança (MV) na estimação da usabilidade de sites *e-commerce*, utilizando dados reais do trabalho de Tezza, Bornia e Andrade (2011) e comparar o desempenho entre os Softwares BILOG-MG® e Excel® na estimação dessas usabilidades.

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 seu uso. Mais especificamente, (1) refere-se à rapidez com que os usuários podem aprender a usá-lo, (2) a eficiência deles ao usá-lo, (3) o quanto lembram deste, (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, pode-se destacar: 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), *card sorting* (RAU; LIANG, 2003; ROSSO, 2008) entre outras. Estas medidas envolvem características objetivas e subjetivas, baseadas em critérios recomendados por especialistas ou em opiniões de usuários, gerando muitas das vezes falta de sistematização e de precisão nos resultados (CYBIS, 2007) o que torna a maioria destas 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 para sistematizar os resultados, pode-se fazer uso de escalas de medidas alicerçadas em conceitos matemáticos e de usabilidade. Desse ponto de vista, a Teoria da Resposta ao Item representa uma poderosa ferramenta, uma vez que esta possibilita a criação de escalas a partir de um conjunto de itens que faz uso de conceitos aprofundados de usabilidade. Este processo se dá 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) em ordem de importância (ou dificuldade). Após isso, nesta 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 interativo, no qual, a partir de um conjunto de itens previamente estimados, é possível estimar a "proficiência" (ou grau de usabilidade) de um respondente qualquer, independente de seu contexto, representando, portanto, uma abordagem objetiva e geral.

3. Teoria da Resposta ao Item

3.1 Aspectos básicos

A TRI é um conjunto de modelos matemáticos 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 é o modelo logístico de dois parâmetros (ML2P) desenvolvido por

Birnbaum (1968), representado pela Equação 1:

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

em que:

a_i é o parâmetro de discriminação do item *i*, proporcional à inclinação da função no ponto b_i;

b_i representa o valor da variável latente *è* (grau de usabilidade), para o qual há 0,50 de probabilidade do indivíduo *j* (no caso, *website j*) responder positivamente ao item, representada por U=1;

U é a resposta ao item, que pode ser positiva (U=1), no caso, se o *website* possui a característica descrita no item, ou negativa (U=0), caso contrário; e

θ representa a proficiência (ou grau de usabilidade) do respondente.

A relação entre a resposta prevista ao item e o traço latente do indivíduo é conhecido através da curva característica do item (*item characteristic curve - ICC*) (RECKASE, 1997). A ICC, exemplificada na Figura 1, representa a regressão não linear de probabilidade de uma determinada resposta (eixo y) em função do nível de habilidade (*trait level*) (eixo x) numa escala (0,1), ou seja, construída com média igual a zero e desvio padrão igual a um (SANTOR; RAMSAY; ZUROFF, 1994).

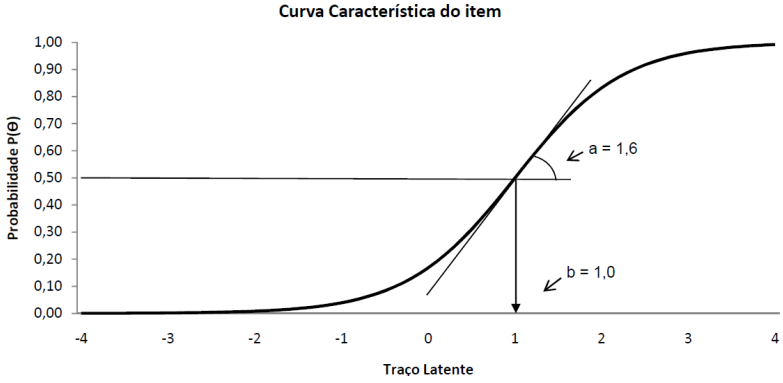

Figura 1. Curva Característica de um item hipotético.

No caso de medição do grau de usabilidade em sites da web, o eixo x da figura 1 representa o grau de usabilidade no qual é possível posicionar os itens e os sites, o que possibilita comparar os desempenhos dos sites em relação a sua usabilidade e a qualidade dos itens. Na ICC, também é possível visualizar os parâmetros do item, nota-se que o parâmetro de dificuldade (*b*) de um item representa a posição na escala em que a probabilidade de acerto é de 0,5. Quanto mais deslocado para a direita, maior é a dificuldade do item e consequentemente maior terá que ser o grau de usabilidade de um site para ter 0,5 de probabilidade de responder positivamente ao item. O parâmetro *a* é o ponto de inflexão no ponto *b* e representa o poder de discriminação do item (ANDRADE; TAVARES; VALLE, 2000).

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

A estimação dos parâmetros dos itens e, posteriormente, das habilidades, se dá por meio de complexos métodos estatísticos que necessitam de recursos computacionais para serem utilizados. Em geral, utiliza-se o método da máxima verossimilhança (MV) ou algum método bayesiano, por exemplo, da esperança a posteriori (EAP) ou da moda a posteriori (MAP). Esses métodos são discutidos com mais detalhes em Andrade, Tavares e Valle (2000).

3.2 Estimação das habilidades pelo Método da Máxima Verossimilhança

A estimação das habilidades, assim como a estimação dos parâmetros dos itens, também envolve complexos métodos estatísticos que necessitam de recursos computacionais para serem utilizados. Existem vários métodos de estimação, sendo que os principais métodos utilizados são o método da Máxima Verossimilhança (MV) e os métodos Bayesianos da Esperança a Posteriori (EAP) e da Moda a Posteriori (MAP). Esse artigo aborda o processo de estimação pelo método da Máxima Verossimilhança (MV), o qual será explicitado nesta seção.

Levando em consideração a suposição de independência local e a independência entre as respostas dos diferentes indivíduos, pode-se obter a

função de verossimilhança apresentada na Equação 2:

$$L(u/\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j} \qquad (2)$$

Considerado o ML2P, a Equação 2 pode ser expressa conforme a Equação 3:

$$L(u/\theta) = \prod_{j=1}^{n} \left(\frac{1}{1+e^{-a_i(\theta_j - b_i)}} \right)^{u_j} \left(\frac{e^{-a_i(\theta_j - b_i)}}{1+e^{-a_i(\theta_j - b_i)}} \right)^{1-u_j} \qquad (3)$$

O método MV consiste em encontrar o valor máximo dessa função. Uma forma usualmente utilizada para encontrar esse valor máximo consiste em aplicar o logaritmo na Equação 3, obtendo-se a Equação 4:

$$\ln L(u/\theta) = \sum_{j=1}^{n} \left\{ u_j \left(\frac{1}{1+e^{-a_i(\theta_j - b_i)}} \right) + (1-u_j) \left(\frac{e^{-a_i(\theta_j - b_i)}}{1+e^{-a_i(\theta_j - b_i)}} \right)_j \right\}, \qquad (4)$$

que deve ser derivada e igualada a zero. O Estimador de Máxima Verossimilhança de θ_j (EMV) é o valor que maximiza a função de verossimilhança, ou seja, é a solução da Equação 5:

$$\frac{\partial \ln L(u/\theta)}{\partial \theta_j} = 0. \qquad (5)$$

O resultado obtido por essa derivada não apresenta solução explícita para θ_j e, por isso, necessita-se de algum método iterativo para obter as estimativas desejadas. Dois processos iterativos utilizados são o Newton-Raphson (ISSAC; KELLER, 1966) e "Scoring" de Fisher (RAO, 1973), cujas expressões matemáticas podem ser encontradas em Andrade, Tavares e Valle (2000).

Na TRI, o erro padrão (EP) do EMV é dado pela Equação 6:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}}, \qquad (6)$$

em que $I(\theta)$ é a Função de Informação do Teste (FIT), que é simples-

mente a soma das informações fornecidas em cada item respondido pelo respondente, apresentada na Equação 7:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta). \quad (7)$$

No caso no ML2P, abordado neste trabalho, a informação do item é obtida pela Equação 8:

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta). \quad (8)$$

Dessa forma, o erro padrão da habilidade pode ser obtido pela Equação 9:

$$EP(\theta) = \frac{1}{\sqrt{\sum_{i=1}^{I} a_i^2 P_i(\theta) Q_i(\theta)}}. \quad (9)$$

O método MV apresenta problemas na estimação do valor máximo quando existem padrões de respostas constantes (ANDRADE; TAVARES; VALLE, 2000), ou seja, quando as respostas são sempre "sim" ou sempre "não". Quando isso ocorre, geralmente os softwares adotam algum procedimento alternativo para estimar a proficiência, que é o caso, por exemplo, do BILOG-MG® (TOIT, 2003), em que os indivíduos que erraram todos os itens ganham um meio certo no item mais fácil respondido até então, e os indivíduos que acertaram todos os itens perdem um meio certo no item mais difícil, conforme mencionam Andrade, Tavares e Valle (2000). Entretanto, na maioria dos casos práticos, essa situação não ocorre se o conjunto de itens for composto por itens com diferentes dificuldades (parâmetro *b*), ou seja, um instrumento de medida que possua itens fáceis e difíceis, a fim de que ninguém responda "sim" para tudo ou "não" para tudo.

4. Materiais e métodos

Foram utilizados dados coletados referentes à usabilidade de 361 sites e-commerce (TEZZA; BORNIA; ANDRADE, 2011). Quarenta e quatro

itens foram elaborados de forma dicotômica, em que 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. Os itens foram analisados e calibrados através do software BILOG-MG®, com o Modelo Logístico Unidimensional de 2 Parâmetros da TRI. Por meio da análise dos parâmetros dos itens, os itens inadequados foram removidos e o questionário final ficou constituído de 32 itens.

Um dos objetivos deste trabalho é comparar o desempenho entre os Softwares BILOG-MG® e Excel® na estimação da usabilidade de sites pelo Método MV. Normalmente, os softwares que utilizam o método da máxima verossimilhança, como o BILOG-MG® (TOIT, 2003), utilizam a metodologia descrita na Seção 3.2, que compreende, na aplicação do logaritmo na função de verossimilhança, a derivação e métodos numéricos iterativos. A estimação da usabilidade dos sites pelo Método MV feita no BILOG-MG® utilizou Newton-Raphson (ISSAC; KELLER, 1966). No Excel®, foi utilizada uma abordagem diferente para a maximização da MV, sem nenhum processo iterativo, a qual é descrita a seguir.

A partir da própria função de verossimilhança (Equação 2), definiu-se um intervalo entre -4 e 4, em que se acredita que as habilidades irão situar-se, considerando a escala (0, 1). Em seguida, dividiu-se esse intervalo em 800 subintervalos consecutivos de amplitude 0,01. Dessa forma, obteve-se 801 pontos de limites inferiores e superiores de intervalos. Para cada site, para cada um desses pontos e para cada item respondido, calculava-se a quantidade $P_j^{u_j} Q_j^{1-u_j}$, ou seja, se o site possuía a característica do item j ($u_j = 1$), calculava-se a probabilidade de um site situado nesse ponto ter essa característica, e, se o site não possuía a característica do item j ($u_j = 0$), calculava-se a probabilidade de um site situado nesse ponto não ter essa característica.

Finalmente, para cada um dos 801 pontos e para cada site, obteve-se o produtório de todas essas probabilidades, que é o uso da própria definição de verossimilhança (Equação 2). Dessa forma, criou-se uma forma discretizada da função de verossimilhança. O máximo dessa função discretizada será aquele ponto em que o produtório das probabilidades for o maior entre todos os 801 pontos determinados.

Nas situações em que o item não se aplicava a um determinado site, utilizou-se um artifício, atribuindo-se o valor 1 à quantidade para todos

os pontos do intervalo. Assim, o valor do produtório não era afetado nessas situações. Nos casos em que o valor da estimativa fosse maior do que 4, o valor atribuído pelo presente procedimento foi igual a 4, e nos casos em que o valor da estimativa fosse menor do que -4, o valor atribuído pelo presente procedimento foi igual a -4.

Nas situações em que o padrão de respostas se manteve constante, foi atribuído o valor 4 caso o padrão tenha sido "sim" (possui o atributo) ou o valor -4 caso o padrão tenha sido "não" (não possui o atributo).

Os cálculos foram realizados com uma precisão de duas casas decimais para a estimação da habilidade. O software Excel® foi escolhido pela facilidade em editar equações e gráficos e para verificar o seu desempenho em comparação com o Software BILOG-MG®.

5. Resultados e discussões

O primeiro resultado observado foi que os valores das habilidades estimadas no Excel® foram os mesmos das habilidades estimadas no BILOG-MG®, com a precisão utilizada no Excel® de duas casas decimais, mostrando a consistência do processo elaborado no Excel®. As estimativas dos erros padrão também foram muito parecidas, apresentando diferença apenas na terceira ou quarta casa decimal. Ressalta-se ainda que o procedimento realizado no Excel® conseguiu estimar o erro padrão dos sites 241, 247 e 257 (que responderam "sim" para 31 dos 32 itens), o que o BILOG-MG® não conseguiu fazer. A Figura 2 apresenta as habilidades estimadas com o respectivo erro padrão (EP) e a posição dos 32 itens na escala. Observa-se que o erro padrão das estimativas é menor onde a maioria dos itens está concentrada e aumenta, à medida que os itens vão ficando mais distantes um do outro. Na TRI, esse resultado tem coerência, uma vez que, quanto maior for a quantidade de itens numa região, maior será a informação do teste nessa região e menor será o erro padrão, o qual é inversamente proporcional à informação do teste.

A Figura 3 apresenta a estimativa da habilidade do site 1 pelo método da máxima verossimilhança, à medida que os itens vão sendo respondidos, e seu respectivo erro padrão (EP).

Observa-se que nos 5 primeiros itens o valor estimado para a habilidade é igual a 4. No procedimento utilizado neste trabalho, enquanto o padrão de respostas se manteve constante ou se o valor estimado da

habilidade esteve fora do intervalo entre -4 e 4, foi atribuído o valor 4 caso o padrão fosse "sim" (possui o atributo) ou o valor -4 caso o padrão fosse "não" (não possui o atributo). Enquanto esse padrão se manteve, o erro padrão da estimativa foi tão alto que optou-se por não mostrá-lo para não comprometer a escala do gráfico da Figura 3.

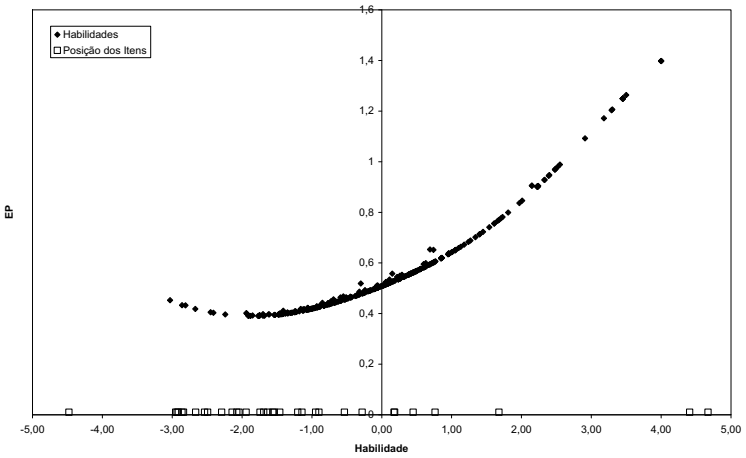

Figura 2. Habilidade e EP dos sites.

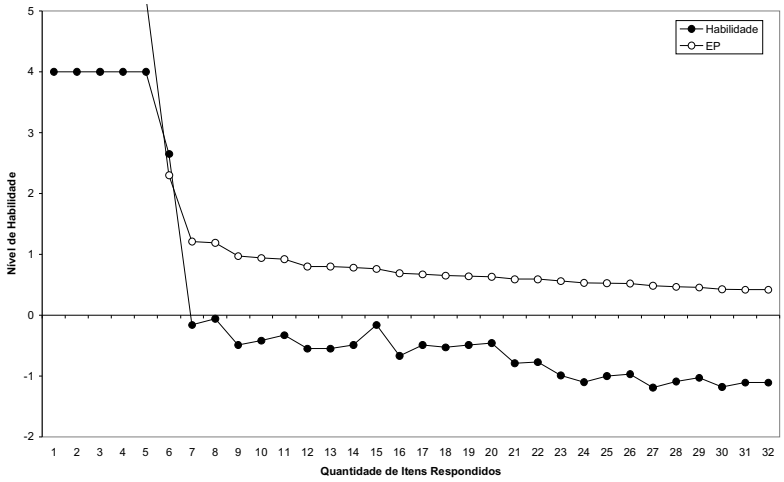

Figura 3. Desempenho do site 1.

Observa-se que, a partir do sétimo item, tanto a habilidade quanto o seu erro padrão caem consideravelmente. À medida que os itens vão sendo respondidos, o EP vai diminuindo, como é de se esperar. A partir do 23° item, a estimativa da habilidade se aproxima de -1 e, ao final dos 33 itens respondidos, atinge seu valor definitivo; -1,11. Ambas as estimativas da habilidade e do EP mantiveram os mesmos valores no 13° item em relação ao 12° item, uma vez que o 13° item era um tipo "não aplicado" a esse site.

A Figura 4 apresenta a estimativa da habilidade do site 1, pelo método MV, à medida que os itens iam sendo respondidos, e o respectivo intervalo de confiança (95%), com o limite superior (ICS) e o limite inferior (ICI). Foram omitidos os intervalos de confiança (IC) das primeiras 6 estimativas, por serem muito grandes. Observa-se evolução da estimativa da habilidade e o estreitamento do IC à medida que os itens vão sendo respondidos. Ao final do questionário, obteve-se o intervalo de 95% de confiança de -0,20 a -1,84. Isso significa que o intervalo [-0,29; -1,93] possui 0,95 de probabilidade de conter a verdadeira habilidade do site 1.

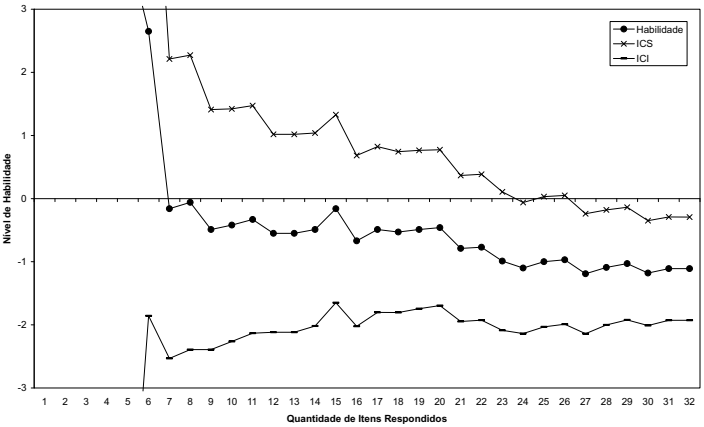

Figura 4. Desempenho do site 1 e Intervalo de Confiança.

As Figuras 5 e 6 apresentam a evolução da curva da função de verossimilhança do site 1 à medida que os itens vão sendo respondidos. Os números entre parênteses indicam a quantidade de itens aplicados à respectiva curva. À medida que os itens vão sendo respondidos, a curva vai tomando uma forma mais estreita, o que torna mais precisa a estimação do valor máximo da função.

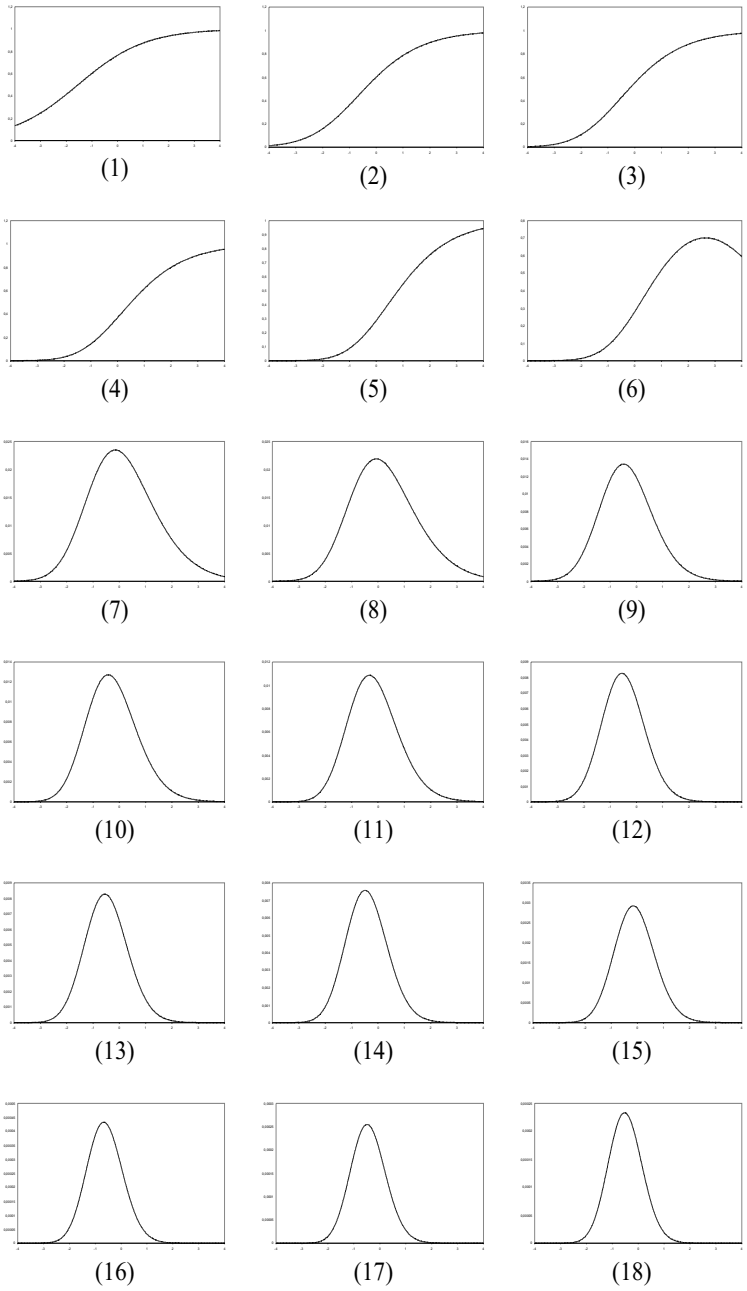

Figura 5. Evolução da função de Verossimilhança do site 1.

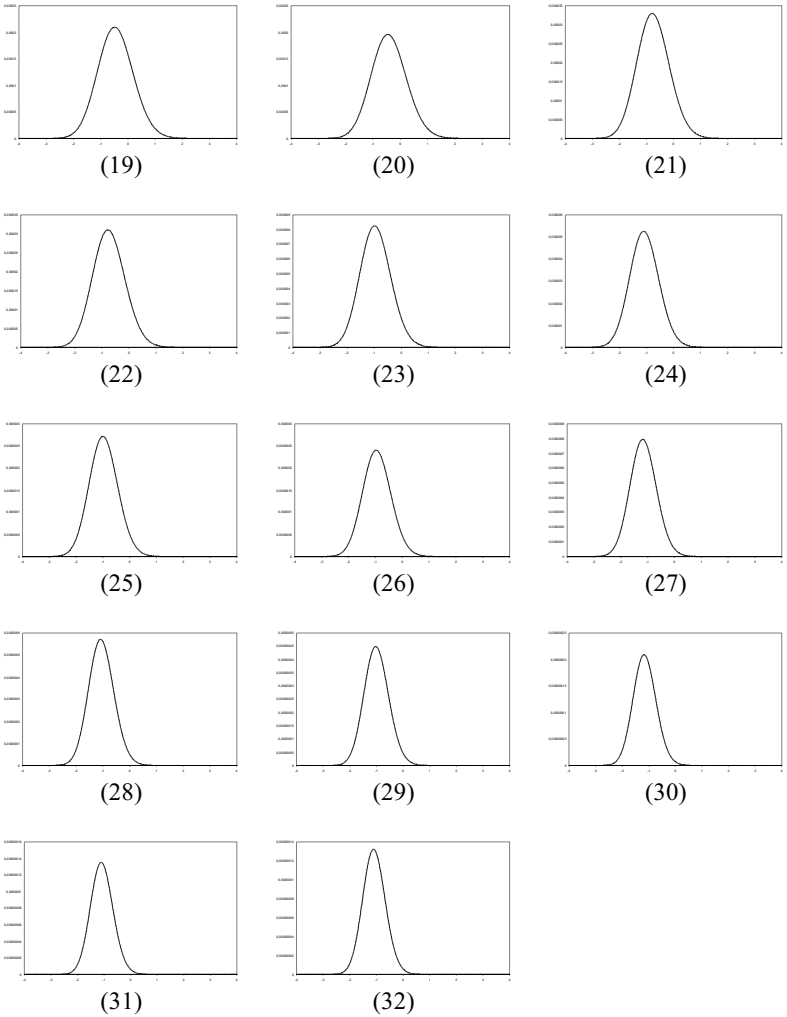

Figura 6. Evolução da Função de Verossimilhança do site 1.

A Figura 7 apresenta a curva da função de verossimilhança do site 1 após a aplicação de todos os 32 itens. O valor máximo (-1,11) é a estimativa da habilidade pelo método MV.

A Tabela 1 apresenta os 11 sites que responderam "sim" em 30 dos 32 itens aplicados, a habilidade estimada de cada um, os itens que obtiveram resposta "não" e o valor dos parâmetros de discriminação e de dificuldade, respectivamente, entre parênteses.

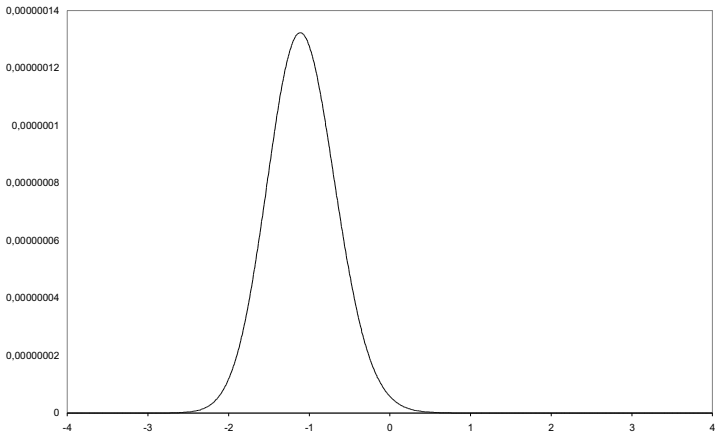

Figura 7. Função de Verossimilhança final do site 1.

Tabela 1. Sites que responderam "sim" em 30 dos 32 itens.

Site	Habilidade	Itens que receberam "não"
51	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
173	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
193	3,30	4 (0,79; -0,90) e 6 (0,78; 4,67)
211	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
235	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
255	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
264	3,45	6 (0,78; 4,67) e 32 (0,69; 4,41)
325	3,50	13 (0,75; -0,28) e 32 (0,69; 4,41)
329	3,18	9 (0,97; 0,18) e 32 (0,69; 4,41)
331	2,91	6 (0,78; 4,67) e 18 (1,09; 1,68)
337	3,29	5 (0,80; -1,56) e 6 (0,78; 4,67)

Observa-se que, apesar de todos terem respondido "sim" em 30 dos 32 itens aplicados, nem todos tiveram a mesma habilidade estimada já que não foram os mesmos itens respondidos afirmativamente por todos. Diferentemente da Teoria Clássica dos Testes (TCT), em que todos os sites teriam o mesmo escore, a TRI leva em consideração cada item para atribuir a habilidade. Portanto, se os itens respondidos positiva ou nega-

tivamente não forem os mesmos, a habilidade estimada não será a mesma, seja por meio do método MV ou por algum método Bayesiano.

A maior habilidade foi de 3,50 para o site 325 e a menor foi de 2,91 para o site 331, proporcionando uma diferença de 0,59 entre as habilidades. Esses dois sites responderam "não" para itens diferentes, o que proporcionou diferentes valores de habilidade para eles. Ambos responderam "não" para o item de dificuldade maior que a sua habilidade e para o item com dificuldade menor que a sua habilidade. O site 331 obteve a menor habilidade porque respondeu "não" (o que não era esperado) para um item com dificuldade menor que a sua habilidade e que possuía um parâmetro de discriminação alto (1,09), o que "penalizou" a estimativa da sua habilidade, diminuindo-a bastante em relação aos demais sites que também responderam "sim" a 30 itens.

A segunda maior habilidade estimada (3,45) foi apresentada por 6 sites (sites 51, 173, 211, 235, 255 e 264) dentre os 11 sites da Tabela 1. Eles tiveram resposta "não" para dois itens de dificuldade maior que a sua habilidade estimada, o que é coerente dentro do contexto da TRI.

O site 325 teve resposta negativa para os itens 13 e 32, e obteve habilidade estimada igual a 3,50. Já o site 329 teve resposta negativa para os itens 9 e 32, e obteve habilidade estimada igual a 3,18. Nota-se que os sites 325 e 329 tiveram resposta "não" a um mesmo item (32) e a outro item diferente. Ambos os itens 9 e 13 possuem uma dificuldade menor do que a habilidade estimada desses sites, o que significa que, teoricamente, os sites 325 e 329 deveriam ter resposta positiva para esses itens. Entretanto, o item 9 possui uma discriminação maior do que o item 13, o que contribuiu para estimar uma habilidade menor para o site 329. Conclui-se, portanto, que os parâmetros de discriminação (a) e de dificuldade (b) têm influência na estimativa da habilidade utilizando o método da máxima verossimilhança.

6. Conclusão

Este artigo analisou o desempenho do método de Máxima Verossimilhança (MV) na estimação da usabilidade de sites *e-commerce*.

Os resultados mostraram que o método de Máxima Verossimilhança apresenta deficiências quando existe um padrão de resposta constante, o

que pode ocorrer durante a aplicação dos primeiros itens do questionário. Entretanto, o método apresenta um bom desempenho quando o padrão de respostas não é constante.

Além disso, o desempenho do processo elaborado no Excel® foi melhor do que no software convencional BILOG-MG® utilizado nas análises da TRI, pois, além de obter as mesmas estimativas que o BILOG-MG® para as habilidades, também conseguiu estimar o erro padrão de alguns sites que o BILOG-MG® não conseguiu.

Também foi possível verificar que os parâmetros de discriminação (*a*) e de dificuldade (*b*) exercem influência na estimativa da habilidade utilizando o método da máxima verossimilhança. Esses parâmetros são capazes de determinar se um site receberá uma estimativa maior ou menor para a sua habilidade. Para futuros trabalhos, sugere-se estudar como funciona essa influência, ou seja, o quanto os valores dos parâmetros de discriminação (*a*) e de dificuldade (*b*) influenciam na estimativa da habilidade por MV.

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.

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

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

AZEVEDO, C. L. N. **Métodos de Estimação na Teoria da Resposta ao Item.** 2003. 133 f. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2003.

AZEVEDO, C. L. N. **Modelos Longitudinais de Grupos Múltiplos Multiníveis na Teoria da Resposta ao Item: Métodos de Estimação e Seleção Estrutural sob uma Perspectiva Bayesiana.** 2008. 265 f. Tese

(Doutorado em Ciências) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2008.

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, MA: Addison-Wesley, 1968.

CHEVALIER, A.; BONNARDEL, N. Articulation of web site design constraints: Effects of the task and designers' expertise. **Computers in Human Behavior**, 23, p. 2455-2472, 2007.

CHOE, P.; KIM, C.; LEHTO, M. R.; LEHTO, X.; ALLEBACH, J. Evaluating and improving a self-help technical support Web site: Use of focus group interviews. **International Journal of Human-Computer Interaction** 21(3), p. 333–354, 2006.

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** (43:2), p 476-491, 2007.

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

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

ISSAC, E.; KELLER, H. B. **Analysis of Numerical Methods.** New York: Wiley & Sons, 1966.

IVORY M. Y., R. MEGRAW, Evolution of web site design patterns, **ACM Transactions on Information Systems (TOIS)**, v.23 n.4, p.463-497, October 2005

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.

LARGE, A.; BEHESHTI, J.; NESSET, VALERIE; BOWLER, Leanne *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**. Vol. 43, Issue 1, p. 1-23, 2006.

LAZAR J.; MEISELWITZ, G.; NORCIO, A., A taxonomy of novice user perception of error on the web, **Universal Access in the Information Society Journal 3** (3/4), p. 202–208, 2004.

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

NIELSEN, J. **Usability Engineering**. California : Morgan Kaufmann , 1993.

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

RAO, C. R. **Linear Statistical Inference and Its Applications**. New York: Wiley & Sons, 1973.

RAU, P.; LIANG, S. F. Internationalization and localization: evaluating and testing a website for Asian users. **Ergonomics**, 46, 1-3, p. 255–270, 2003.

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** (p. 271–286). New York: Springer- Verlag, 1997.

REISE, S. P.; WIDAMAN, K. F.; PUGH, R. H. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. **Psychological Bulletin,** 114(3), 552–566, 1993.

ROSSO M. User-Based Identification of Web Genres. **Journal of the American Society for Information Science and Technology,** 59(5):1–20, 2008.

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**, 6 (3), p. 255-70, 1994.

SCHENKMAN, B. N.; JÖNSSON, F. U. Aesthetics and preferences of web pages. **Behav. Inf. Technol**. 19, p. 367-377, 2000.

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.

TEZZA, R.; BORNIA, A. C.; ANDRADE, D. F. Measuring web usability using item response theory: Principles, features and opportunities. **Interacting with Computers.** Volume 23, Issue 2, p. 167-175, 2011.

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

Submetido: 09/05/2011
Aceito: 26/10/2011
Revisão do *Abstract*: CCAA-Sul (Santa Maria, RS)