Interacting with Computers 23 (2011) 167–175

Contents lists available at ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Measuring web usability using item response theory: Principles, features and opportunities ☆

Rafael Tezza [a,*], Antonio Cezar Bornia [a], Dalton Francisco de Andrade [b]

[a] Production Engineering Department, Federal University of Santa Catarina, SC, Brazil
[b] Informatics and Statistics Department, Federal University of Santa Catarina, SC, Brazil

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form 27 January 2011
Accepted 20 February 2011
Available online 26 February 2011

Keywords:
Usability checklist
Item response theory
E-commerce

ABSTRACT

Usability is considered a critical issue on the web that determines either the success or the failure of a company. Thus, the evaluation of usability has gained substantial attention. However, most current tools for usability evaluation have some limitations, such as excessive generality and a lack of reliability and validity. The present work proposes the construction of a tool to measure usability in e-commerce websites using item response theory (IRT). While usability issues have only been considered in theoretical or empirical contexts, in this study, we discuss them from a mathematical point of view using IRT. In particular, we develop a standardised scale to measure usability in e-commerce websites. This study opens a new field of research in the ergonomics of interfaces with respect to the development of scales using IRT.

1. Introduction

1.1. Context

The advancement and popularisation of the Internet has significantly changed the way people do business. Such popularisation has broadened e-commerce and has made it more dynamic. Companies have generally tried to use e-commerce features to attract more clients and to be successful in a more competitive market (Fink and Nyaga, 2009). In this context, some issues arise regarding e-commerce management and service such as the performance evaluation of e-commerce companies, the identification of the most important factors regarding customer interaction, the facilitation of customer interaction, and so on.

Studies on e-commerce performance evaluation have highlighted user interaction with the website as a key issue in the purchase decision as well as in the development of consumer fidelity (Janda et al., 2002; Kim and Stoel, 2004; Lim and Dubinsky, 2004; Song and Zinkhan, 2003; Wang et al., 2001). A website is the first point of contact for a potential client to interact with the company through the Internet (Zhang and Dran, 2002).

Following this line of thought, a good ergonomic interface is so important that it can even influence whether a sale occurs, as users will not want to read web pages that are not user-friendly. In addition, users will not want to navigate through difficult pages in order to explore or understand them, and they will not interact with pages that respond differently from their expectations (Goto and Cotler, 2005). To develop user-friendly websites, it is necessary to perform a usability evaluation (Dumas and Redish, 1993). Generally, usability evaluations are conducted by applying user testing, heuristic evaluation, or specific questionnaires (Alexander and Tate, 1999; Flanders and Willis, 1998; Nielsen, 1992; Okada, 2001). Each one of the methods has its own benefits and drawbacks, and as such, they are generally applied in tandem in different phases of an interface design (Desurvire et al., 1992; Jeffries et al., 1991; Tan et al., 2009). However, the application of these methods has some disadvantages; for example, user testing and heuristics require financial resources and specialised personnel for implementation. Popular questionnaires used to assess usability or user attitudes such as SUS (System Usability Scale) (Brooke, 1996), WAMI (Website Analysis and Measurement Inventory) (Kirakowski and Cierlik, 1998), PUTQ (Purdue Usability Testing Questionnaire) (Lin et al., 1997), IsoMetrics (Gediga et al., 1999), and UFO (Usability Questionnaire for Online Shops) (Konradt et al., 2003) are not always directly applicable depending on the application domain, and they also may not cover all aspects in need of evaluation (Díaz et al., 2002; Konradt et al., 2003). Moreover, theirs scales cannot demonstrate reliability and validity (Barriocanal et al., 2005).

The lack of good evaluation for usability and the difficulties in comparing usability performance restrict website development as well as company positioning among competitors. The difficulties involved in clearly identifying the desired features can reduce the allocation of resources to less important features, thereby potentially neglecting otherwise important features (Nielsen and Loranger, 2006).

☆ This research was sponsored by CNPq.
* Corresponding author. Tel.: +55 48 9608 9690.
E-mail address: rafaeltezza@yahoo.com.br (R. Tezza).

0953-5438/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.intcom.2011.02.004

To promote a better understanding of the structures involved in usability evaluations and to systematise results, it is possible to use measurement scales based on mathematical and usability concepts. From this point of view, item response theory (IRT) is a powerful tool that enables the construction of standardised scales from a set of items via mathematical models (Embretson and Reise, 2000; Hambelton et al., 1991). In the context of usability, Schmettow and Vietze (2008) discuss the use of IRT in measuring usability inspection processes.

The objective of this study is to build a set of items and a standardised scale in order to measure usability in e-commerce website interfaces based on IRT. This work is structured as follows. First, the context of study is presented, followed by an explanation of scale construction, IRT and web usability. In Section 2, the methodology is described. In Section 3, the findings are presented and then discussed further in Section 4. In Section 5, the conclusion is presented.

1.2. Scale creation: item response theory and web usability

IRT describes a set of mathematical models aimed at measuring latent traits (that is, individual profile characteristics that cannot be measured directly). The models use a set of items to the construction a scale such that the latent trait of the respondent and item difficulty can be compared (De Ayala, 2009; Embretson and Reise, 2000; Hambleton, 2000).

The construction of measurement scales facilitates the understanding of complex concepts. First, the look for several aspects of a variable provides the creation of knowledge about it. Second, various perspectives highlight the range, which allows finer distinctions, especially if measures are ordinal. Third, the construction of measures allows for the efficient reduction of data such that a numerical score can represent an ordinal position for certain characteristics of items or elements of the population; this ranking thus allows comparability (Babbie, 2005).

Variables must be theoretically or conceptually connected with the intended object of measurement (Baker, 2009) to develop a scale based on an item pool. Therefore, it is necessary to develop a concept for the latent variable under study, which in this case is web usability.

Nielsen and Loranger (2006, p. 16) defines usability as "a quality attribute relating to how easy something is to use". More specifically, it refers to the ease with which a system or component can be learned, how efficiently it can be used, how memorable it is, how error-prone it is, and how much users like using it (IEEE, 1998; Nielsen and Loranger, 2006). For software developers, usability can be described in terms of the internal attributes of a system that affect user performance and productivity (Seffah and Metzker, 2004).

The present study focuses on the scope of usability measures and, more specifically, the internal attributes of a system that affect user performance and productivity in e-commerce systems.

According to Courville (2004), Embretson and Reise (2000), Fan (1998), Uttaro and Lehman (1999), IRT methods have distinct advantages over classical methods in that (a) item parameters and subject latent trait levels are independent, while in the classical test theory the estimators have circular dependency and are thus dependent on the sample. (b) The model is expressed at the level of the observed item response rather than at the level of the observed test score. (c) In IRT, the item is the unit of focus, while in classical methods, the respondent's observed score for an entire instrument is the unit of focus. (d) The standard error of measurement in IRT differs across scores (or response patterns), while in classical methods the standard error is the same across scores. (e) The shorter tests under IRT can be more reliable than longer tests, and (f) subject scores under IRT can be equated even if respondents answer different questions. Finally, (g) IRT allows the use of computerised adaptive testing (details, see van der Linden and Glas (2000)).

Traditionally, IRT is used to develop instruments for assessing individual differences based on the responses of prior participants. For example, in a mathematic assessment, individuals respond to items as correct or incorrect depending on their proficiency in mathematics. In this paper, we apply IRT to the usability field, where the focus is less on individuals and more on systems, and there are no correct versus incorrect answers to be scored but rather degrees of ease of use, in other words, if the website has or has no specific feature that facilitates its use.

In IRT, the choice of mathematical model basically depends on the type of item. The model represents the probability of a certain response to a determined item in accordance with the parameters of the item and the respondent's latent traits (Andrade et al., 2000; Reise et al., 1993; Tavares et al., 2004). One of the most widely used IRT models for items with dichotomous and cumulative responses is the two-parameter logistic model (2PLM) developed by Birnbaum (1968) and based on Lord (1952), which is represented by the following equation.

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

b_i is the difficulty parameter of item i represented on the same scale as the latent trait θ_j, and a_i is the discrimination (or inclination) parameter of item i. Usually, the quantity of b is represented on a scale with mean zero and standard deviation one. The Item Characteristic Curve (ICC) represents the $P(U_{ij} = 1/\theta_j, a_i, b_i)$ relationship with respect to the probability of a certain response to an item, the respondent's latent traits and the item's parameters. As shown in Fig. 1, the ICC represents a non-linear regression of the probability of a certain response according to the trait level (Santor et al., 1994).

To measure the usability level of websites, axis x in Fig. 1 represents the usability level at which it is possible to rank items and websites, thereby allowing the comparison of websites based on usability and item characteristics.

This IRT application assumes unidimensionality, which means that all items measure only one dimension, namely, the usability level of e-commerce websites. With the assumption of unidimensionality, local independence can be obtained; in other words, the response to a certain item does not depend on responses to other items given the latent traits of respondents (Lin and Yao, 2008; Reckase, 1997).

To achieve a given level of website usability using IRT, a practitioner should follow the following steps: (a) create or adapt items correlated with the construct to be measured (i.e., usability), (b) evaluate a group of websites with the purpose of gathering data for parameter estimation, (c) calibrate the items to estimate their *alpha* and *beta* parameters, and (d) estimate the website usability levels. Steps (c) and (d) need the support of a software application because the calculations are complex.

2. Method

2.1. Population

The data used in this analysis are from a sample of 361 Brazilian e-commerce websites. The websites present different products and have various characteristics; they were randomly chosen using a search tool from websites as well as an Internet search. The reason for observing simple web page design as well as more elaborate design is that the latter does not necessarily imply better usability but rather represents diversity in the latent trait. This diversity is required by IRT for better estimations of the item parameters.