Is there a Universal Instrument for Measuring Interactive Information Retrieval? The Case of the User Engagement Scale[1]

Heather L. O'Brien
University of British Columbia
Vancouver, BC, Canada
hlobrien@interchange.ubc.ca

Elaine G. Toms
Dalhousie University
Halifax, NS, Canada
elaine.toms@dal.ca

ABSTRACT
This paper examines the validity of the User Engagement Scale (UES). Originally developed and tested in e-shopping, the scale was administered to users of a multimedia webcast system in an experimental setting. Factor analysis examined the structure and loadings of 31 items. As in previous research, a six-factor solution was found. However, the number of items was reduced and one of the original sub-scales (Felt Involvement) was eliminated. These results are examined contextually by comparing the current study with previous research. The findings discuss the feasibility of a universal measure of user engagement in Interactive Information Retrieval (IIR).

Categories and Subject Descriptors:
H.5 [Information Interfaces and Presentation]
H.5.2 User Interfaces – evaluation/methodology, user-centred design.

General Terms:
Measurement, Design, Reliability

Keywords:
Measurement, user experience, validity, context

1. INTRODUCTION

Interactive Information Retrieval (IIR) combines system-centered and human-centered approaches to investigate information seeking, retrieval and use [6] in dynamic contexts. As such, there are a variety of methods employed in the conduct of IIR studies, ranging from system-based precision and recall to more user- centered measures pertaining to users' thoughts and feelings while interacting with IIR systems. Users' perceptions of experience are challenging to quantify, since users themselves may have difficulty articulating their cognitive and affective responses to systems, and these reactions may not be readily observable to researchers. However, perceptual variables are essential components of IIR research. They may determine if users will use a system to learn, purchase a product, carry out research or work-related tasks, etc. in future, and they may influence users' performance with a system [1].

[1] This is the author's post-print version of a published work. Full citation is as follows:
O'Brien, H.L. & Toms, E.G. (2010). Is there a Universal Instrument for Measuring Interactive Information Retrieval? The Case of the User Engagement Scale. In *Proceedings of Information Interaction in Context (IIiX), Rutgers, NJ,* pp. 335-340, ACM Digital Library, DOI: 10.1145/1840784.1840835.

One type of measurement used to quantify and make sense of users' experience with IIR systems are psychometric scales. There is a history of such instruments in the disciplines of psychology, education, and business to address, for instance, technology adoption in the workplace (Technology Acceptance Model) [3], and to make abstract constructs concrete through "a surrogate set of behaviorally relevant measures" [14]. Scales have been developed to explore users' interactions with technology. Some of these are focused on aspects of system usability, such as satisfaction [5] or disorientation in navigation [1], while others examine pleasurable states of interaction with systems, such as playfulness [15], flow [10], and engagement [8].

The User Engagement Scale (UES) [8] is a multidimensional scale that contains six sub-scales: Aesthetics, Novelty, Felt Involvement, Focused Attention, Perceived Usability, and Endurability. Its purpose is to assist researchers and designers of IIR systems in reaching a holistic understanding of users' encounters with technology, tapping into cognitive, affective and behavioural perceptions of interactions, and gauging future intentions to use a system. The scale seeks to simultaneously measure multiple aspects of engaging experiences and understand their relationships to one another. Thus far, the scale is limited to the e-shopping environment; it must demonstrate its generalizability to other contexts. We report here on the administration of the scale in an experimental setting in which participants were asked to perform search and summarizing tasks using a multimedia webcast system.

2. Prior Research

2.1 Scale Evaluation and Use

In order for measurement scales to become established, practical tools in IIR research and system design, they must be reliable, valid, and generalizable. Reliability pertains to whether or not the items that make up the overall scale or its sub-scales demonstrate internal consistency [11]. The purpose of validity is to demonstrate the ability of an instrument to capture the phenomena of interest to the researcher. Of particular significance to the current study is external validity, or the pertinence of research findings to "the real world" [6]. Another associated term is generalizability, the "administrative viability and interpretation [of a scale] in different research situations" [11, p. 79-80], or the "larger universe" [13, p. 288].

Scales and questionnaires are utilized in IIR, though few have established their reliability or validity [6]. The repeated use of some instruments across studies and over time has made them "core by default" [6, p. 179]; however, repeated use must not be equated with statistical precision. There are several benefits of developing reliable, valid, and generalizable research instruments. The first and most obvious reason is to demonstrate rigor [13]. Efforts to construct and evaluate measures result in improved instruments, and in the compilation of a body of research that permits observations of phenomena over time. Second, such instruments enable researchers to define and measure variables consistently; this may facilitate communication and collaborative research efforts in IIR. Valid instruments may serve to scrutinize the degree of fit between research questions and real-world problems. Lastly, solid instruments give us confidence in the design of research studies, and in our results [13].

2.2 Background: The User Engagement Scale

Scale development is a longitudinal process that involves instrument construction and evaluation. The UES is the product of several years of research. It was rooted in a strong conceptual foundation [9] and involved a methodical process to construct and assess potential items. This culminated in testing over 120 items with 440 online shoppers. The results of this study resulted in a parsimonious, reliable scale and informed a subsequent study with 802 e-

shoppers in which the factor structure of the items and the relationships among factors were examined [8]. The outcome of this work was a 31-item instrument with six factors: Aesthetics, Novelty, Felt Involvement, Focused Attention, Perceived Usability, and Endurability.

2.3 The Current Study

As [6] points out, measures developed in the fields of human-computer interaction and information systems may not be plug-and-play in IIR studies. IIR encompasses differentdevices, domains, populations, tasks, settings, and numbers of simultaneous users. In the current study, data were collected to assess the external validity of the UES in a different context from the one in which it was originally developed and tested. An experimental set up, rather than an online survey, was the setting in which we examined participants' interactions with a webcast system. We evaluated the reliability of the sub-scales and the factor structure of the UES to explore its external validity.

3. METHOD

The current study was designed to test users' engagement with and performance using a Webcast system (Figure 1). Webcasts contain multimedia content (video, audio, images, and text) and are used to broadcast presentations, meetings, and lectures, and record and archive them for later retrieval.

3.1 Participants

Participants were 53 male and 37 female university students (58% undergraduates) who were mostly under the age of 27 (78%).

3.2 Interactive System

Participants interacted with one of two versions of the webcast system. One version (Figure 1) contained a basic timeline ("Slide-based") similar to RealMedia Player, for example. Participants could mouse over the tick marks, which represented presentation slides, to see the slide number, title, and time it was featured during the presentation. The other version of the interface was almost identical, except that its timeline ("Webcast") had enhanced functionality in the form of a scrollable filmstrip and zoom lens that displayed more detail on a particular slide.

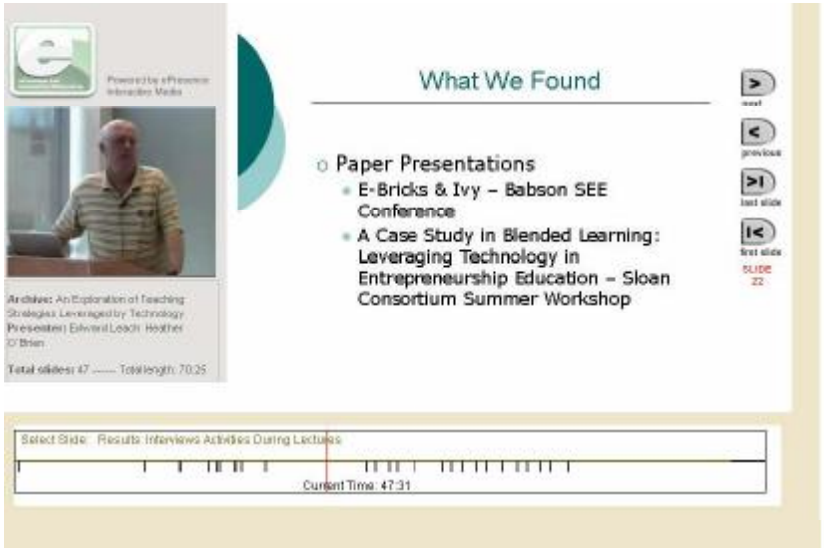

Figure 1. Screenshot of Webcast System

3.3 Procedure

This was a between-subjects design where participants interacted with the Slide-based or Webcast interface. All participants followed the same experimental procedure: they completed a brief tutorial of the system they were randomly assigned to, and then performed two tasks (presented in counterbalanced order). Participants viewed two presentations, also counterbalanced to prevent order effects. One presentation was on the topic of management and comprised two lectures on leadership and health (total slides: 48; duration: 72 minutes). The other presentation pertained to folksonomies, and culture and heritage (total slides: 46; duration: 57 minutes). Participants completed demographic, two post-task, and a post-session questionnaire. The latter contained the UES with items rated on a 7-point Likert scale ranging from strongly disagree (1) to strongly agree (7).

For one task, participants were asked to write a "gist" or summary of the presentation's content; the other task was to locate specific information within the presentation ("fact finding"). Experimental sessions (tutorial, two tasks, and questionnaires) lasted an average of 70 minutes. The total time allotted for the experiment meant that participants were required to use searching and browsing strategies to complete the tasks; watching a presentation in its entirety was not possible. Strategies observed included using slide titles and content, and the familiar structure of a presentation (beginning, middle, end) to "jump" into the presentation at certain points. Unlike typical IIR system, the multimodal nature of the webcast system required users to retrieve information from not only text, but from audio and visual content.

3.4 Data Preparation

To prepare the data for analysis, some items were reverse coded. An initial examination of the data showed that there were no missing variables for any of the items. The item means ranged from 4.32 to 5.34 on the 7-point Likert scale. None of the means resided near the extremes of the scale, indicating they should have good variability and yet correlate with other items [4].

4. RESULTS

4.1 Reliability Analysis

Previous research indicated that the UES was comprised of 31 items that loaded on six distinct factors: Aesthetics, Novelty, Focused Attention, Felt Involvement, Perceived Usability, and Endurability. Data from the current study was examined to determine if the items associated with the six factors in [8] would form reliable sub-scales. Table 1 shows that the six sub-scales were reliable with Cronbach's alpha values in the respectable (0.7) to very good range (0.9) [4].

Table 1. Reliability Analysis of User Engagement Sub-scales

Sub-scale	Mean	St. Dev	No. Items	α
Aesthetics (AE)	4.4	1.19	5	.92
Perceived Usability (PU)	4.68	.96	8	.8
Focused Attention (FA)	4.26	1.11	9	.9
Endurability (EN)	4.58	1.05	5	.85
Novelty (NO)	3.82	1.37	3	.79
Felt Involvement (FI)	4.42	.96	3	.74

4.2 Factor Structure

4.2.1 Correlation Analysis

An examination of the factor structure of the UES items began with correlation analysis of the items within each of the six sub-scales. As demonstrated in Table 2, most of the sub-scales were significantly correlated with correlation coefficients in the moderate range (0.23 to 0.63). However, the Perceived Usability (PU) showed negative relationships with Novelty (NO) and Focused Attention (FA), a low correlation with Felt Involvement (FI), and a significant relationship with Endurability (EN).

Table 2. Correlation Coefficients of Sub-scales

	AE	PU	FA	EN	(NO)
PU	.3*				
FA	.13	-.13			
EN	.4**	.52**	.28*		
NO	.23*	-.04	.45**	.48**	
FI	.34**	.06	.56**	.59**	.63**

*p<0.05; ** p<0.001

The results of the correlation analysis indicated that the factor structure of the scale might not correspond to previous findings (specifically PU) [8]. The strong relationship between some subscales (FA, NO, and EN) indicated that some items may load on multiple factors, or analysis may result in fewer factors/items.

4.2.2 Factor Analysis

Maximum Likelihood Factor Analysis with oblique rotation [2] was used to examine the factorstructure of the UES. The guidelines for interpreting item loadings were derived from [15]. Items were discarded if they did not load with a minimum value of 3.2 (10% overlapping variance) or crossloaded with a value of 3.2 on more than one factor. Over five iterations, 12 items were eliminated: 3 PU, 3 EN, 1 NO, 2 FA, and all 3 FI items. The result was six factors. Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO = 0.72) and Bartlett's Test of Sphericity ($x2=1042.6$, $df=171$, $p<0.001$) were significant. The final factor analysis is displayed in Table 3.

Cronbach's alpha values of the resulting factors are shown in row 2. Items are listed in column one; the label in brackets corresponds to the original results [8]. Factor 1 consisted of 5 AE items, while FA (5) items loaded on factor 2. The PU items loaded on two factors: 3 items pertaining to affective responses formed factor 3, while 2 challenge items loaded on factor 4. Factor 5 consisted of 2 NO items and factor 6 contained 2 EN items.

Table 3: Factor Analysis of Engagement Scale Items

	1	2	3	4	5	6
Cronbach's alpha (α)	.91	.83	.85	.89	.83	.85
The webcast systems was aesthetically appealing (AE)	.83					
This webcast system appealed to my senses (AE).	.82					
I found the screen layout of this system to be visually pleasing (AE).	.81					
This webcast system is attractive (AE).	.81					
I liked the graphics and images used of this webcast system (AE).	.77					
I blocked out things around me when I was using this system (FA).		.89				
When I was using the system, I lost track of the world around me (FA).		.89				
I was absorbed in my task (FA).		.69				
I was so involved in my task that I lost track of time (FA).		.50				
I lost myself in this experience (FA)		.49				
I felt frustrated while using this webcast system (PU).			.88			
I felt annoyed while using this webcast system (PU).			.75			
I felt discouraged while using this webcast system (PU).			.74			
Using this system was taxing (PU).				.98		
This task was stimulating (PU).				.79		
I continued to use this webcast system out of curiosity (NO).					.97	
The content of the webcast incited my curiosity (NO).					.68	
Using this webcast system was worthwhile (EN).						.91
I would recommend that others use this webcast system (EN).						.62

5. DISCUSSION

The findings of the current study are not consistent with previous research [8]. Although a six-factor solution resulted, one of the original sub-scales, FI, was eliminated, and PU items were split across two factors. Aside from these differences, the AE, FA, NO, and EN sub-scales retained their integrity, although they did not, with the exception of AE, retain the same number of items. It is also interesting that AE was not correlated with FA or PU because [8] found a predictive relationship between these factors.

The reasons for these discrepancies may be an inherent problem with the composition of the scale. However, a strategic and rigorous process was followed in the development and construction of the original instrument [8,9] as prescribed by [4, 12] and as exemplified by [1, 3, 15] . Another explanation may be the timeline system itself. We investigated interactions with a novel interface, the webcast, to explore strategies used to browse and search multimodal information. Participants' lack of familiarity with the system may have made the interaction novel and graphically interesting, but not involving. The FI items related to "losing oneself" in the interaction. This may not have been possible if users were challenged by the mechanisms of the interface. This may also be the reason why PU items relating to negative affect and challenge loaded on different factors.

Users' interactions with IIR systems are "dynamic, complex, situated [and] temporally bound," and contingent upon the internal states of users [7]. Context is another area of interest when comparing the findings of the current study to previous results. The UES was developed and administered to online shoppers (70% female) who ranged in age and occupation. They completed the scale based on experiences with online shopping in general (study 1) and with a specific retailer (study 2) [8]. The shopping tasks evolved in the course of their everyday lives.

Regardless of how often they visited the e-commerce site they reported on, there was likely some degree of familiarity with the look and feel of a shopping website, and its associated tasks of browsing and searching for products, saving items to a shopping cart, and completing billing and shipping information forms. They completed the UES based on a shopping experience that had occurred within the past six months. The current study had a different demographic make-up: they were undergraduate university students (~ 60% male) under the age of 27. They completed the UES in the context of a lab experiment with researcher-generated tasks immediately following the interaction. Although participants may have had experience with video, slide, and timeline applications, the webcast system presented these media together in what was likely a novel application.

Given the diversity of IIR systems and users, should we conclude that the UES and psychometric instruments in general are not useful in IIR evaluation? In short, no. Measuring user perception is essential in IIR evaluation, and these tools give us parameters for gauging these perceptions. In the case of the UES, the number of items in the current study did not match previous research. However, with the exception of FI and the loading of PU items on two factors, the structure of the factor analysis was maintained. This may indicate that components of engagement are consistent across systems, but the manifestation and salience of these elements is what varies. It is not practical to develop a new measure of engagement for every IIR system/interaction. The UES gives designers and researchers a set of factors for defining experience and focusing measurement efforts. For example, if researchers wished to examine interface presentation, then they might hone in on AE and NO items. The UES may also be used as a comparison for other IIR metrics. For instance, do the PU items indicate that users found the system difficult to use, even though the performance metrics indicate an efficient interaction?

6. CONCLUSION

This study tested the external validity of the UES in an experimental setting with users of a webcast system. Findings were not consistent with previous research [8], as 12 items and one sub-scale were not retained in the final factor structure and perceived usability items loaded on two factors. However, with the exception of Felt Involvement, each sub-scale based on the original study remained with that sub-scale after factor analysis. We identified a number of contextual differences between the current and original studies. The cognitive perspective that informs IIR highlights the role that users and context play in shaping interactions with systems [6]. Given the versatility in IIR environments and the need for instruments that are meaningful in naturalistic and experimental settings, it is important to address the applicability of the UES across contexts. While the results of this study question the feasibility of a universal instrument to evaluate user engagement in IIR systems, we maintain that the UES identifies factors that inform experience and provides scope in the measurement of user perceptions in IIR research. We also encourage further research in the area of measurement, specifically the reliability, validity, and generalizability of user perception metrics in IIR. It is imperative to rigorously evaluate our instruments as well as the data we collect. This will ensure confidence in our analysis of the relationships between measures and conclusions, and will establish a body of work that can be used to assess longitudinal trends and outcomes in IIR.

7. REFERENCES

[1] Ahuja, J. and Webster, J. Perceived disorientation: An examination of a new measure to assess web design effectiveness. Interact Comput 14, 5-29.

[2] Costello, A.B. and Osborne, J.W. 2005. Best practices in factor analysis. PARE 10, 9 pp. http://pareonline.net/pdf/v10n7.pdf

[3] Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. User acceptance of computer technology: A comparison of two theoretical models. Manage Sci 35, 982-1003.

[4] DeVellis, R. F. 2003. Scale Development, 2nd ed. Sage.

[5] Doll, W. J. and Torkzadeh, G. 1988. The measurement of end-user computing satisfaction. MISQ 12, 259-274.

[6] Kelly, D. 2009. Methods for evaluating information retrieval systems with users. Foundations and Trends in IR 3, 224 pp.

[7] Hassanzahl, M. and Tractinsky, N. 2006. User experience: A research agenda. Behav Info Tech 25, 91-97.

[8] O'Brien, H.L. and Toms, E.G. 2010. The development and evaluation of a survey to measure user engagement in e-commerce environments. JASIS&T 61(1), 50-69.

[9] O'Brien, H.L. and Toms, E.G. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. JASIS&T 59(6), 938-955.

[10] Pace, S. 2004. A grounded theory of the flow experiences of Web users. Int J Hum-Comput St 60, 327-363.

[11] Peterson, R.A. 2000. Constructing Effective Questionnaires. Sage.

[12] Sharma, S. and Weathers, D. 2003. Assessing generalizability of scales used in cross-national research. Int J Res Mark 20, 287-295.

[13] Straub, D.W. and Carlson, C.L. 1989. Validating instruments in MIS research. MISQ 13(2), 147-169.

[14] Tabachnick, B. G., and Fidell, L. S. 2007. Using Multivariate Statistics, 5th Ed. Allyn & Bacon.

[15] Webster, J. and Martocchio, J.J. (1992). Microcomputer playfulness: Development of a measure with workplace implications. MISQ 16, 201–226.

Is there a Universal Instrument for Measuring Interactive Information Retrieval? The Case of the User Engagement Scale

Heather L. O'Brien, University of British Columbia, Vancouver, British Columbia, Canada
Elaine G. Toms, Dalhousie University, Halifax, Nova Scotia, Canada
hlobrien@interchange.ubc.ca; elaine.toms@dal.ca

Motivation

Users' perceptions of experience are challenging to measure, yet essential to paint a holistic picture of interactive Information Retrieval (IIR) evaluation. Psychometric scales, which try to make abstract constructs concrete through "a surrogate set of behaviourally relevant measures" [3] offer a means through which users may articulate their cognitive and affective responses to systems.

Although scales and questionnaires are utilized in IIR, few have been tested for reliability or validity [1]. We need to demonstrate statistical rigor of these tools:

- To further collective research efforts and create a body of work for meta-analyses and observing phenomena over time.
- To facilitate communication and collaborative research efforts in IIR through consistent definitions and measurements.
- To scrutinize the degree of fit between research questions and real-world problems.
- To be confident research study designs and results [3].

User Engagement Scale

We developed and evaluated a multidimensional instrument to measure user engagement with IIR systems in the e-shopping domain [2] designed to tap into the cognitive, affective and behavioural perceptions of and future intentions to use systems. It has six sub-scales: *Aesthetics, Novelty, Felt Involvement, Focused Attention, Perceived Usability, and Endurability*. Here we examine the external validity of the User Engagement Scale (UES) by comparing the results of factor analysis (FA) in the original e-shopping context to an experimental study with a multimedia webcast system.

Contextual Comparison of the User Engagement Scale (UES)

802 shoppers
Major online book retailer [2]
Purchasing an item through the website

E-Shopping	Attribute	Webcast
18-70 yrs	– Age –	78% <27 yrs
70% F; 30% M	– Gender –	41% F, 59% M
Variety	– Occupation –	University students
Online survey	– Context –	Lab experiment
Self-generated	– Task source –	Researcher
Familiar	– Interface –	Novel
Within past 6 mos	– Self-report mechanism –	Immediately

90 students
Two versions of a webcast system
Two tasks: fact finding and summarizing a presentation

Are UES/psychometric scales useful in IIR evaluation?

YES. The UES:

- Gives us parameters for gauging holistic experiences with/ users' perceptions of IIR systems.
- Provides designers and researchers with a set of factors that can be used to define experience and focus measurement efforts.
- May be used as a comparison with other IIR metrics.

Results: Factor Analysis

UES Scale Item	Original Sub-scale	Factor	Loading
The webcast system was aesthetically appealing.	Aesthetics (AE)	1	.83
The webcast system appealed to my senses.	AE	1	.82
I found the screen layout visually appealing.	AE	1	.81
This webcast system was attractive	AE	1	.81
I liked the graphics and images used in this system.	AE	1	.77
I blocked out things around me when using this system.	Focused Attention (FA)	2	.89
When using this system, I lost track of the world around me.	FA	2	.89
I was absorbed in my task.	FA	2	.65
I was so involved in my task that I lost track of time.	FA	2	.5
I lost myself in this experience.	FA	2	.49
I felt frustrated using this system.	Perceived Usability (PUE +Affect)	3	.88
I felt annoyed using this system.	PU-Affect	3	.75
I felt discouraged using this system.	PU-Affect	3	.745
Using this system was taxing.	PU-challenge	4	.68
This task was stimulating.	PU-challenge	4	.79
I continued to use the system out of curiosity.	Novelty (NO)	5	.67
The content of the webcast incited my curiosity.	NO	5	.68
Using the system was worthwhile.	Endurability (EN)	6	.91
I would recommend others to use this system.	EN	6	.62

Comparison of Results of Two Studies

E-Shopping	Webcast
6 Factor solution	6 Factor Solution
31 items	19 items
Factors: AE, FA, PU, NO, Felt Involvement (FII), EN	Factors: AE, FA, NO, EN, FI eliminated. PU split into two factors

Explaining Discrepancies

- **A problem with the scale?**
  A strategic and rigorous process was followed in the development and construction of the original instrument [2].
- **Context?**
  Need further testing
- **The timeline system itself?**
  The AI items related to "being unaware," which may not have been possible if users were challenged by the mechanisms of the interface.
  May also be why AU items relating to negative affect and challenge loaded on different factors.

Conclusion

This study tested the external validity of the UES in an experimental setting with users of a webcast system. Findings were not consistent with previous research [2]; however, with the exception of FI, the composition of the sub-scales were consistent with [2].

While the results of this study question the feasibility of a universal instrument to evaluate user engagement in IIR systems, we maintain that the UES identifies factors that inform experience and provides scope in the measurement of user perceptions in IIR research. Given the versatility in IIR environments and the need for instruments that are meaningful in naturalistic and experimental settings, it is important to address the applicability of the UES across contexts. We also encourage further research in the area of measurement, specifically the reliability, validity, and generalizability of user perception metrics.

References

(1) Kelly, D. 2009. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in IR 3, 1-224.

(2) O'Brien, H.L. and Toms, E.G. 2010. The development and evaluation of a survey to measure user engagement in e-commerce environments. JASIST 61(8), 50-69.

(3) Straub, D.W. and Gefen, D.V. 2004. Validating instruments in MIS research. MISQ 13(2), 147-169.

(4) Tabachnick, B. G., and Fidell, L. S. 2007. Using Multivariate Statistics, 5th Ed. Allyn & Bacon.

Shopping cart image: http://www.clker.com/clipart-24761.html