Advance Access publication on 27 March 2013
doi:10.1093/iwc/iwt018

A Single-Item Measure of Website Usability: Comments on Christophersen and Konradt (2011)

JEFF SAURO*

*Corresponding author: jeff@measuringusability.com

Christophersen and Konradt show that a single-item measure of usability can retain most of the information garnered in multi-item scale. While there is some loss of information and reliability, the authors show that when a single item is needed, it can be reliable, sensitive (discriminate between good and bad usability) and valid (correlates with other known measures of usability).

Keywords: Usability testing

Special Issue Editors: Gitte Lindgaard & Jurek Kirakowski

The authors (Christophersen and Konradt, 2011) provide an excellent review of relevant research on multi-item and single-item scales across HCI. They remind us what many learned about scale construction, that multiple items are expected to have higher reliability than single items. When measuring a construct, it has become good practice to include multiple items because it can be difficult to capture the essence of a construct with just a single item. What is more, errors in responses are expected to be distributed across multiple responses (Nunnally, 1978).

The authors collected items from the marketing and usability literature and use one item (slightly adapted) from the Post Study System Usability Questionnaire PSSUQ for the single measure of online usability. They generated a questionnaire with 18 items that encompass four constructs: usability, trust, aesthetics and intention to buy. All subscales had high reliability ($\alpha = 0.91$) (Nunnally, 1978).

In their validation study, 378 participants were asked to visit two online stores. In total, 5 product groups were selected and 7 stores per product group were selected for each product group, for a total of 35 stores (most of them familiar to the participants).

The participants were asked to go to the online store and perform tasks such as browsing for products.

The authors assessed the reliability in three ways: using the commonality of the factor from a factor analysis, item-total correlation and a method for correction for attenuation as suggested by Nunnally (1978). All three methods suggested that the single item had high reliability, >0.8.

They found the single item correlated with the aesthetics, trust and intention to buy factors ($r = 0.61$, 0.53 and 0.62), respectively. They did find the multiple usability measure correlated more strongly than the single item; however, the difference was modest, considering that seven more items were used to measure the usability construct. The higher correlations were 0.63, 0.56 and 0.63, respectively, meaning the multi-item measure explained on average about 2% points more in variability (from using r^2).

Overall, the study was solid, the literature review helpful and comprehensive, the results interesting and relevant (at least to one who's created similar instruments). In short, the authors show that multi-items scales probably do have more reliability but not that much. In applied settings (not undergraduates answering surveys for extra credit) having users answer surveys can be difficult. Reducing the number of items can certainly help increase both the number of responses and the number of completed responses—which in many cases would offset the modest loss in reliability.

Internal reliability cannot be assessed using a single item; however, the authors used what seemed to me to be reasonable approach. A couple minor notes, the authors state that the PSSUQ is unidimensional (p. 270); however, the PSSUQ has three dimensions such as SysUse, InfoQual and IntQual. More recent data also show that SUS probably has two dimensions (see Lewis and Sauro, 2009; Lewis, 2002). Finally, it was also not clear whether the store pairings in the experiment were paired up with good and bad stores or just random set of two stores.

In conclusion, there is a strong need to reduce the length of questionnaires in applied research due to the limited time researchers have with test participants in a task-based usability

test. While the authors have shown that multi-item measures of usability are more reliable than single measures, they have provided a compelling case that the extra reliability might not be worth the burden of additional items.

REFERENCES

Christophersen, T. and Konradt, U. (2011) Reliability, validity, and sensitivity of single-item measure of online store usability. Int. J. Hum.-Comput. Studies, 69, 269–280.

Lewis, J.R. (2002) Psychometric evaluation of the PSSUQ using data from five years of usability studies. Int. J. Hum. Comput. Interact. 14, 463–488.

Nunnally, J.C. (1978) Psychometric Theory (2nd edn). McGraw-Hill, New York.

Lewis, J.R. and Sauro, J. (2009) The Factor Structure of the System Usability Scale. In Proc. Human Computer Interaction International Conf. (HCII 2009), San Diego, CA, USA.