Interacting with Computers Advance Access published March 27, 2013

doi:10.1093/iwc/iwt019

A Commentary on Short Questionnaires for Assessing Usability

PAUL CAIRNS

Department of Computer Science, University of York, York, UK YO10 5GH
**Corresponding author: paul.cairns@york.ac.uk*

Usability is an important aspect of any interactive system, but only one aspect leading to its success. Thus, when evaluating a system, it is useful to have short questionnaires that can lead to reliable, sensitive and valid measures of usability. In this commentary, I discuss two papers that aim to develop two such measures: one is for a four-item usability metric and the other a single-item usability. Whilst care is taken in both cases to produce suitable metrics, the processes have flaws that influence the value of these metrics. The underlying problem seems to lie with the application of psychometric methods when there are better methods available in this context. Even allowing for this, in the end there remains the problem of construct validity that undermines any such efforts.

Keywords: User studies, usability testing, exmpirical studies in interaction design

Special Issue Editors: Gitte Lindgaard and Jurek Kirakowski

1. SHORT MEASURES OF USABILITY

This paper is intended as a commentary on two papers, Christophersen and Konradt (2011) and Finstad (2010), that aim to develop short questionnaires to measure the usability of interactive systems. As such both come from a similar perspective, which is that usability is just one aspect of interactive systems that lead to their success. When evaluating a product, there is a need to incorporate measures of usability as part of a larger evaluation to look at other aspects such as trust, aesthetics, installation, product support and so on. Typically, these things are measured based on customer surveys, that is, questionnaires and if each aspect had a lengthy set of questions, then very quickly the survey would become unwieldy and adversely affect response rates. Better to have something short that gives acceptable results rather than something that produces better results but which nobody uses. Within this context, short questionnaires that produce reliable measures of usability are a desirable goal. Finstad (2010) aims to produce a four-item usability questionnaire whereas Christophersen and Konradt (2011) go even further and look to produce a single-item measure.

Questionnaire design and validation has a long history drawing strongly on the tradition of psychometrics within psychology but of course more recently being adopted in marketing and other business contexts. The process of questionnaire development can draw on a wide variety of techniques and corresponding statistical measures but all with the same aim of ensuring the validity of the questionnaire. Regardless of context, there are some key considerations that underpin the value of questionnaires and these are used to structure the critique that follows (Kline, 2000). These might be characterized in this context with questions:

1. Reliability: to what extent does the questionnaire produce consistent results?
2. Construct validity: does the questionnaire really measure usability? This may also consider:
   (a) Face validity: do the questions look like sensible questions for measuring usability?
   (b) Convergent or concurrent validity: to what extent does the questionnaire agree with other measures of usability?
   (c) Predictive validity: building on convergent validity, does the questionnaire accurately predict the usability of systems?
   (d) Discriminant validity: to what extent does the questionnaire differentiate from concepts that are not usability, e.g. trust, product support etc?
3. Sensitivity: to what extent does the measure pick up on differences in usability between systems?

Each of these issues is considered in turn for both questionnaires. For convenience, we refer to Finstad's

questionnaire by his acronym, UMUX, the Usability Metric for User eXperience and Christophersen and Kondradt's as SIUM, the Single-Item Usability Metric.

It should be noted that the starting point for UMUX and SIUM in each case is an existing questionnaire. In the case of UMUX, this was the System Usability Scale, SUS (Brooke, 1996) and for the SIUM, this was the Usability Questionnaire for Online Stores, UFOS (Konradt *et al.*, 2003), with an extra question which is the SIUM item drawn from the Post Study System Usability Questionnaire, PSSUQ (Lewis, 2002). Of course, the value of the new questionnaires can to some extent piggy-back on the value of the existing questionnaires though as will be seen, this can also cause some problems.

2. RELIABILITY

Reliability is the degree to which an instrument, like these questionnaires, produces consistent results. This is not to say that the results it produces are accurate (that would be validity) merely that the numbers produced by the questionnaire produce some sort of consistent result, indicating that the questionnaire is measuring something robust and persistent. With regards to questionnaires, there are usually two distinct types of reliability that need to be considered. The first is statistical reliability which is a measure of the internal consistency of the items of the questionnaire with each other. The second is test–retest reliability which is the ability of the questionnaire to reproduce the same results at different points in time. Neither questionnaire took into consideration test–retest reliability so on that basis, the measures of reliability at best aim for internal consistency between items of the questionnaire not consistency in terms of repeatable results over time.

There are various ways of measuring internal consistency. One straightforward way is to split the items of the questionnaire into two sets and see whether the scores due to one half correlate reasonably with the scores due to the other. This split-half method shows that the two halves of the questionnaire are measuring the same thing, namely the variable underlying the questionnaire. There are of course many ways of dividing the questionnaire in two and the Cronbach α is a measure of the whole questionnaire which is effectively the mean of all the split-half correlations. Thus a high α indicates good consistency between all of the items of the questionnaire. Typically $\alpha > 0.7$ is required for good internal consistency.

There is a drawback, although, which is if α is too high, then it indicates that the items are redundant. That is, they are all measuring the same thing identically and not different aspects of some underlying latent variable. Christophersen and Konradt (2011) point out the problem of questionnaire redundancy and that it can actually lead to reduced respondent engagement and hence completion rates and the quality of answers. Exactly what constitutes too high is unclear. Kline (2000) suggests $\alpha > 0.8$ whereas other guidelines suggest $\alpha > 0.9$ (de Vellis, 2003).

High internal consistency is found in UMUX, where $\alpha = 0.94$ (Finstad, 2010). One argument would be that α should be so high as all four questions relate to usability; however, all four questions are targeting different aspects of usability which have previously been found not to correlate very strongly (Hornbaek and Law, 2007). This suggests that the UMUX is perhaps too specific and not measuring usability in a broad sense but perhaps in some narrow sense.

Item-total correlations are also a good measure of internal consistency showing that each item contributes usefully to the total score. It should be noted though that there will always be some correlation between each item and the total because each item contributes to the total. For four questions, each item adds to a quarter of the total variance, that is $r^2 = 0.25$ and hence you would expect $r = 0.5$ for each item-total correlation even when items are scored randomly.[1] The item-total correlations for each varies between 0.69 and 0.89. All correlations are reported as significantly different from zero, but it is not possible to infer if they are significantly different from 0.5.

Overall then, reliability of UMUX is only really asserted through the Cronbach α but this is much higher than would be expected both for a good scale and given what is already known about the correlations between different aspects of usability.

With the SIUM, there is no possibility of using Cronbach α or item-total correlations because there is only one item! Instead, reliability of SIUM is evaluated using an estimation procedure, first based on correlations between the SIUM and the UFOS, its parent questionnaire, and secondly on the communality of the item within a factor analysis of the UFOS with the extra single item. Both methods give a high value of statistical reliability, $\alpha = 0.82$, which give confidence in the reliability of the SIUM.

However, the reliability of the SIUM is entirely in terms of its fit with the UFOS. There can be no measure of internal reliability and it may be that when people are answering the fuller UFOS they are seeing the SIUM in that context which leads them to answer it in some way consistent with the other answers. Thus, in many ways, this supports the reliability of SIUM as having consistency with UFOS but does not support it as having its own consistency if used independently of UFOS. Really, the only available test of consistency available to SIUM is the test-retest method and this was not done.

3. CONSTRUCT VALIDITY

Construct validity is essentially concerned with answering the question 'To what extent does this questionnaire measure what it is meant to measure?' For SIUM and UMUX, construct validity is therefore: to what extent do these questionnaires measure usability? Whereas these different types of validity can be quite thorny in psychometrics because of the challenge in accessing internal states of people, in the case of usability, there

[1] It is relatively straightforward to set up an Excel spreadsheet to see this in practice.

are some very good external objective measures. Efficiency is well measured by task completion times and effectiveness by number of errors or backtracking in navigation. Satisfaction is some what more subjective to measure and often measured in an ad hoc manner (Hornbaek and Law, 2007) but even then there are some standard questionnaires that can measure satisfaction. However, neither questionnaire attempts to directly position itself in relation to objective measures of usability.

The validity of SIUM and UMUX must therefore come from the more usual arguments and statistical analysis used in psychometrics. This is where the other forms of validity related to construct validity come in. Even here, it is not possible to consider predictive validity because that would require predicting the objective usability from the outcomes of the questionnaires. Whilst both questionnaires are developed with reference to real systems, to say that these would be sources of predictive validity would be circular: the systems used to develop the questionnaires would of course confirm the validity of the questionnaires. Predictive validity could only be confirmed through a subsequent study. We therefore turn to each of the other forms of validity to evaluate the two questionnaires.

3.1. Face validity

Face validity is a useful check in that items in a questionnaire should at least appear to measure the underlying concept. In this case, the brevity of UMUX and SIUM do make it easy to see that there is high face validity, each question clearly being concerned with some aspect of usability. However, a threat due to high face validity is that people recognize immediately what the underlying concept is and so answer questions in order to produce the socially desirable answers (Kline, 2000). In the processes of developing both questionnaires, participants were required to use two systems. The shortness and high face validity means that the participants would have been able to see any perceived differences between the systems and answer the questions 'correctly', that is, to provide the socially desirable answers that they feel are requested of them. This may be independent of their perceptions of the systems had they seen each without the other.

The SIUM has the additional problem that the item asks whether the participant is satisfied with the usability of the (online) store. Clearly, this has high face validity but are participants sufficiently knowledgeable to answer this question? There are well-understood psychological mechanisms whereby when faced with a difficult question that would tax even experts (would this person make a good president?), people user fast and frugal heuristics that are effectively answering a simpler question (does this person look dominant and speak well?) (Hardman, 2009). Participants that have only a passing knowledge of what is meant by usability are likely to make such substitutions and so whilst there is high face validity to the SIUM, there is no guarantee that participants are answering the question that is being asked. Without secure knowledge of what usability is, people may in fact be answering the easier question of how much they enjoyed the task.

This is one of the arguments for multi-item questionnaires. Finding out what a person thinks about a complex, latent (that is, unconscious or hard to access) concept is inferred from people's answers to more accessible but imperfect reflections of that concept (Cox, 1980). To compensate for the imperfections, multiple items should address the different facets of the latent concept in order to reduce systematic error in the questionnaire. The UMUX, even though brief, does exactly this by asking about the three core components of usability: efficiency, effectiveness and satisfaction.

3.2. Concurrent (or convergent) validity

Concurrent validity is reached when one questionnaire gives results that tally with the results of a different questionnaire or method of measurement of the same concept. With UMUX, it was natural to consider the concurrent validity of UMUX with the parent SUS questionnaire and as hoped, there is a high correlation between the scores, $r = 0.96$. That is, scores on the UMUX are strong indications of the overall scores on SUS. So, whatever the SUS is measuring, UMUX is measuring something very similar. However, it should be noted that whilst there has been some attempt at validating the SUS (e.g. Lewis and Sauro, 2009; Bangor *et al.*, 2008), there seems to be no large-scale concurrent validation of the SUS in relation to other usability questionnaires nor in direct relation to objective measures of usability relating to efficiency and effectiveness. Thus even if SUS and UMUX are measuring the same thing, it merely pushes the question of UMUX's validity as a usability metric to the validity of the SUS as a usability metric.

SIUM is compared for concurrent validity with the UFOS questions but additionally it is correlated with three further short scales of aesthetics, trust and intention to buy. UFOS correlates strongly with the SIUM and, being a measure of usability, this provides support for the concurrent validity of the SIUM. However, the good correlations between SIUM and the other measures do not as such support convergent validity. What can be inferred is that previous patterns of correlation between usability measures and these other measures are also seen in the SIUM. Despite what Christophersen and Konradt (2011) claim, these correlations do not support convergent validity, merely that SIUM enters into the same relationship with these concepts as other measures of usability. This is not evidence that SIUM is measuring the same as those other usability measures. This leads directly on to discussions of discriminant validity.

3.3. Discriminant validity

Usability is currently understood to sit as only one aspect of many possible perceptions that people may have of interactive systems. It is closely related to people's perception of systems in terms of pleasure (Jordan, 2000), user experience (McCarthy

and Wright, 2007), aesthetics (Hassenzahl, 2004) and so on. When measuring usability, there is therefore a risk that what is being measured is not usability as such but some other concept that relates to usability but also is bleeding into these neighbouring but distinct concepts. Question substitution, as discussed with reference to face validity, also plays a role here.

Thus, to claim to have a good measure of usability, it is important to make it clear that the questionnaires have discriminant validity and dissociate from other closely related concepts. In neither Finstad (2010) nor Christophersen and Konradt (2011) is there an explicit consideration of the discriminant validity of the measures. Indeed, Christophersen and Konradt show that the SIUM correlates with trust, aesthetics and intention to buy, none of which are aspects of usability per se. Moreover, they point out that precisely the fast and frugal heuristics, such as halo effects, might be at play but that this supports convergent validity. In fact, this would reduce discriminant and hence construct validity: if halo effects are in play, how can we know that SIUM is measuring usability and not something else?

4. SENSITIVITY

Both the UMUX and SIUM are acknowledged to be possibly inferior measures of usability because of the brevity of the scales. Christophersen and Konradt (2011) even go so far as to test how much less-sensitive SIUM is than UFOS. The value comes from such short scales if, despite their limitations, they are able to discern differences in usability between systems.

To see the sensitivity of SIUM, it was developed over seven distinct online stores with each participant seeing two of the stores. It was found that there was a significant difference between the seven online stores. Whilst this suggests that the SIUM does distinguish between the stores, it does not say what it is distinguishing and without a prior expectation of what differences ought to have been detected, there is no indication whether SIUM is sensitive to usability differences or even over-sensitive to differences where there ought to be none. Unfortunately, the reporting of this aspect of the analysis is insufficient for the reader to make their own interpretations.

UMUX fares a little better in that it was developed over two systems one of which was generally highly regarded for usability and one of which was not. The UMUX clearly reflects these differences. However, this introduces its own problems. The participants in the study were selected from the company where these systems were in use. It may be that they were answering the UMUX in the way they knew they 'should' answer it in relation to the clear face value of the questions. There is also the risk that the use of two such widely differing systems is inflating all correlations used in the reliability and convergent validity analysis (due to clustering of data for each system). Thus, whilst the study design suggests sensitivity, it may be introducing more problems than it solves. Again the paper does not have sufficient detail to confirm or refute this.

5. DISCUSSION AND CONCLUSIONS

Overall then, there are good grounds to doubt the reliability, validity and sensitivity of both the UMUX and SIUM questionnaires. At the very least, better reporting is needed to allow the reader to judge their value for themselves. At the root of these concerns though is what exactly do these measures mean?

Meaning is a concern for all questionnaires in psychometrics where the inner workings of people's psyches are not only closed to external observers but can be largely closed to people themselves (Kline, 2000). It is only through subtle and ingenious methods that psychologists are able to peer through the complexity of self-reports, behaviour and experimental manipulations to begin to guess at what really goes on in people's heads.

For usability at least though, we need not suffer so. There are (broadly) agreed measures of usability that exist in the objective domain: measures of performance, of errors and even of satisfaction. Yet neither UMUX or SIUM are positioned in relation to these measures but rather tested against other questionnaires in the pscyhometric tradition. It seems not to step back from the psychometric process to see it for what it is—a best effort in the face of insurmountable challenges in psychology but challenges that simply do not apply in this context.

It may be that more careful studies may give greater confidence in UMUX, SIUM or short questionnaires like them. Yet, there remain fundamental theoretical problems. Both scales are uni-dimensional and indeed any short scale almost of necessity must be too. Yet, usability is clearly not uni-dimensional, not only in definition, but also across studies where the three dimensions of efficiency, effectiveness and satisfaction emerge as independent components of usability (Hornbaek and Law, 2007). Any collapsing of this three-dimensional construct to a single dimension must make simplifying compromises. At no point is the nature of these compromises discussed for either UMUX or SIUM. Even discounting the criticisms of reliability and validity, there remains the issue of what exactly these measures can be measures of.

A fallback position is that the questionnaires are in fact measures of perceived usability though neither paper offers this position as an interpretation. There are measures of perceived usability of which the subscale of the Technology Acceptance Model (Davis, 1989) probably has the widest and most extensive validation. But without measures of concurrent validity with these scales, this is merely a matter of conjecture.

HCI in general and usability in particular suffers from the problem of what exactly do things mean. And perhaps the issue

of meaning here could run endlessly (and fruitlessly). Being pragmatic, if researchers, usability consultants, marketers etc. were to use these scales, what could they get from them? They provide some summative, quantitative assessment of something that might be related to usability. So what? If one system has a high usability as measured by UMUX or SIUM, there is no indication of how to repeat this success in future products. Conversely, low usability scores given no indication of how to fix it. What exactly is the context where these quantifications of usability might have value? A problem of aesthetics, of installation or product support at least suggests where to start looking for solutions. But as raised by Christophersen and Konradt (2011), a problem raised by SIUM may be an indication of a lack of trust or aesthetics and not usability at all. And even if it were usability, where would be the place to start looking? UMUX does a little better perhaps indicating whether the problem is one of the effectiveness, efficiency or satisfaction but there is no further indication of what might be causing these problems.

It seems then that these scale development papers are technical exercises where substantial experimental and statistical muscles are flexed but they result in scales that have questionable reliability and validity. The methods of psychometrics are mimicked, mostly unnecessarily, to produce questionnaires whose value is possibly more about showing numbers to clients than improving the usability of systems that all of us use, and suffer from, every day.

ACKNOWLEDGEMENTS

I am very grateful to Prof. Harold Thimbleby and Prof. Heather O'Brien for their careful and constructive comments on earlier versions of this paper.

REFERENCES

Bangor, A., Miller, P.T. and Miller, J.T. (2008) An empirical evaluation of the system usability scale. Int. J. Hum.-Comput. Interact., 24, 574–594.

Brooke, J. (1996) Sus: A "Quick and Dirty" Usability Scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds), Usability Evaluation in Industry, pp. 189–206. Taylor and Francis, London, UK.

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

Cox, E.P. (1980) The optimal number of response alternatives for a scale: a review. J. Market. Res., 18, 407–22.

Davis, F. (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q., 13, 319–339.

de Vellis, R.F., 2003. Scale Development: Theory and Applications (2nd edn). Sage, Thousand Oaks, CA.

Finstad, K. (2010) The usability metric for user experience. Interact. Comput., 22, 323–327.

Hardman, D. (2009) Judgment and Decision Making. John Wiley and Sons, Chichester, UK.

Hassenzahl, M. (2004) The interplay of beauty, goodness, and usability in interactive products. Hum.-Comput. Interact., 19, 319–349.

Hornbaek, K. and Law, E.L.C. (2007) Meta-analysis of correlations among usability measures. In: CHI'07: Proc. SIGCHI Conf. Human Factors in Computing Systems. pp. 617–626. ACM Press, New York.

Jordan, P. (2000) Designing Pleasurable Products. Taylor and Francis, London.

Kline, P. (2000) A Psychometrics Primer. Free Association Books, London, UK.

Konradt, U., Wandke, H., Balazs, B. and Christophersen, T. (2003) Usability in online shops: scale construction, validation and the influence on the buyers' intention and decision. Behav. Inform. Technol., 22, 165–174.

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.

Lewis, J.R. and Sauro, J. (2009) The Factor Structure of the System Usability Scale. In: Kurosu, M. (ed.), Human-Centered Design, pp. 94–103. Springer, Berlin, Germany.

McCarthy, J. and Wright, P. (2007) Technology as Experience. MIT Press, Boston, MA.