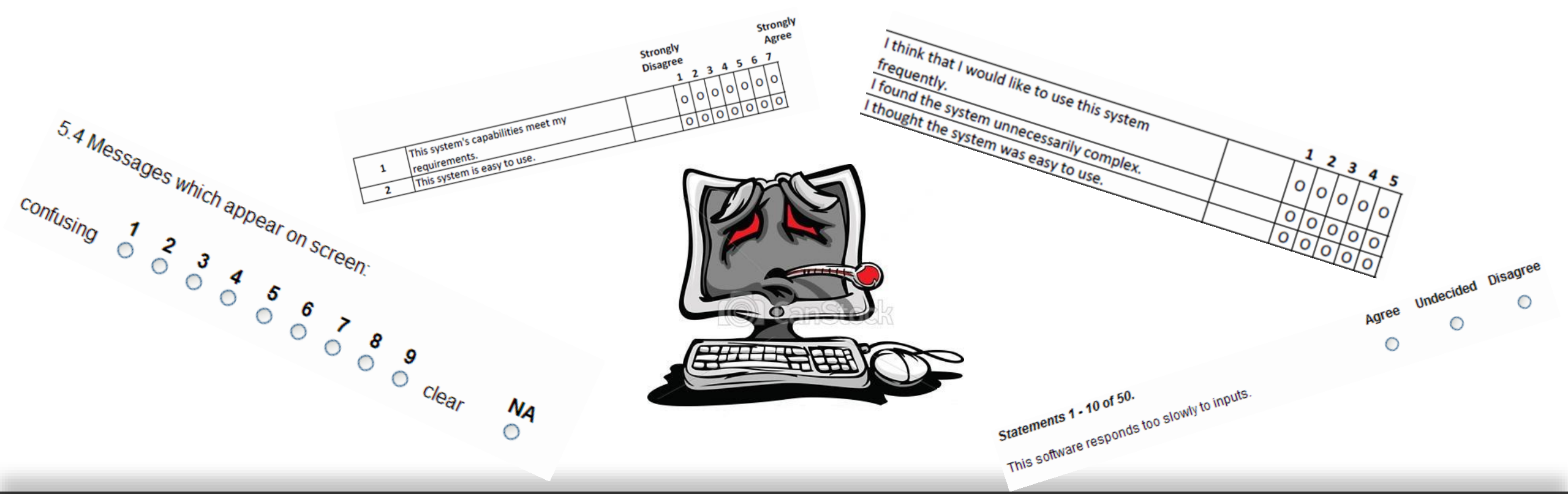


Perceived Usability

Usefulness and measurement

James R. Lewis, PhD, CHFP
Distinguished User Experience Researcher
jim@measuringu.com

Measuring U

What is Usability?

SEE THE PROOF!

PROOF 1 LOWER OPERATING COST

An accurate electric meter measures the amount of current used by Frigidaire and *proves* how little current the Meter-Miser uses. *Compare it with others.*

PROOF 2 SAFER FOOD PROTECTION

Frigidaire dares to place a Food-Safety Indicator in the cabinet to *prove* that Safety-Zone Temperature is maintained — at lowest cost. *Compare it with others.*

PROOF 3 FASTER FREEZING—MORE ICE

The Thermo-Gauge *proves* Frigidaire's ability to freeze larger quantities of ice—faster. *Compare it with others.*

PROOF 4 MORE USABILITY

Wider, roomier, handier to use. Automatic Reset Defroster. Full-Width Sliding Shelves. Portable Utility Shelf. Double-Range Cold Control. Saves steps, Saves work. *Compare it with others.*

PROOF 5 FIVE-YEAR PROTECTION PLAN

Frigidaire's sealed-in mechanical unit—a marvel of outstanding design and engineering—comes to you protected for Five Years against service expense for only Five Dollars included in the purchase price. *Compare it with others.*

Ask us for Proof!

FRIGIDAIRE
MADE ONLY BY GENERAL MOTORS

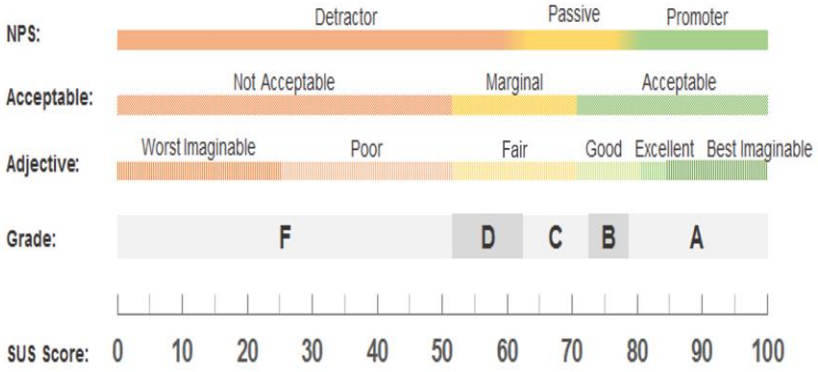
Look For This Name-Plate

- Earliest known (so far) modern use of term “usability”
- Refrigerator ad from Palm Beach Post, March 8, 1936
- Note “handier to use”
- “Saves steps, Saves work”
- tinyurl.com/yjn3caa
- Courtesy of Rich Cordes

What is Usability?

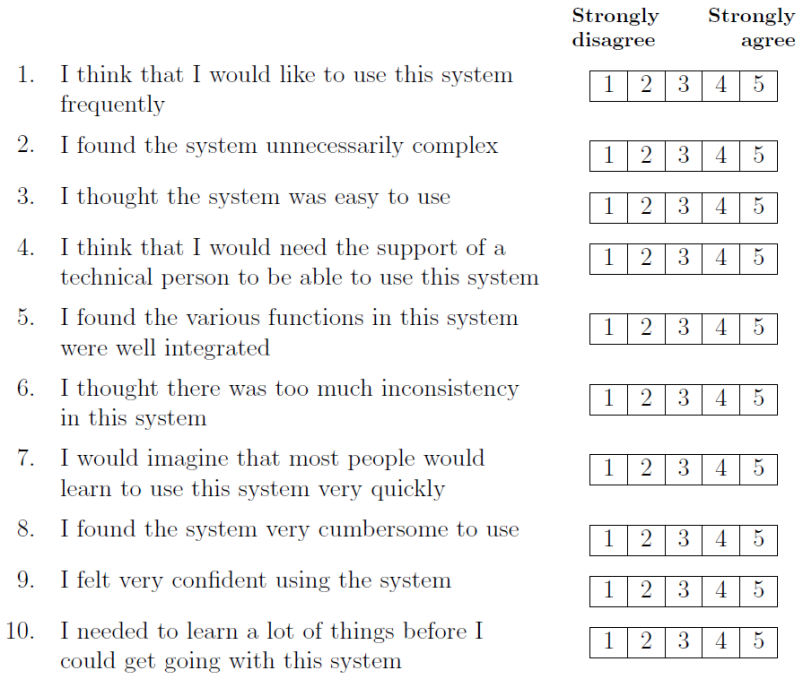
- Usability is hard to define because:
- It is not a property of a person or thing
- There is no thermometer-like way to measure it
- It is an emergent property that depends on interactions among users, products, tasks and environments
- Typical metrics include effectiveness, efficiency, and satisfaction

Introduction to Standardized Usability Measurement



- What is a standardized questionnaire?
- Advantages of standardized usability questionnaires
- What standardized usability questionnaires are available?
- Assessing the quality of standardized questionnaires

What Is a Standardized Questionnaire?



		Strongly disagree				Strongly agree
1.	I think that I would like to use this system frequently	1	2	3	4	5
2.	I found the system unnecessarily complex	1	2	3	4	5
3.	I thought the system was easy to use	1	2	3	4	5
4.	I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5.	I found the various functions in this system were well integrated	1	2	3	4	5
6.	I thought there was too much inconsistency in this system	1	2	3	4	5
7.	I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8.	I found the system very cumbersome to use	1	2	3	4	5
9.	I felt very confident using the system	1	2	3	4	5
10.	I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

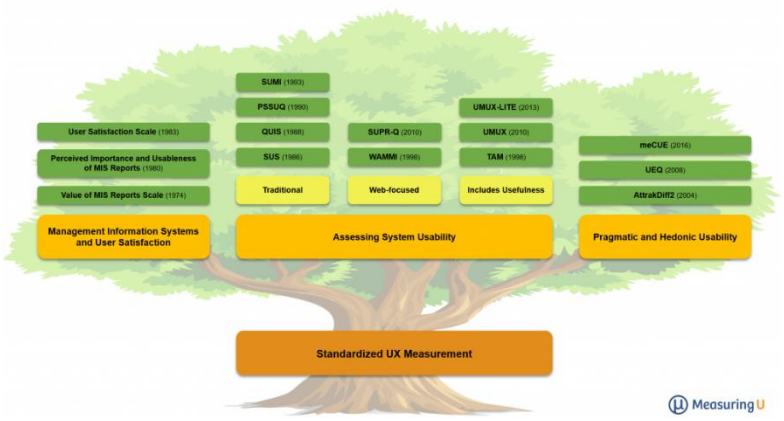
- Designed for repeated use
- Specific set of questions presented in a specified order using a specified format
- Specific rules for producing metrics
- Customary to report measurements of reliability, validity, and sensitivity (psychometric qualification)
- Standardized usability questionnaires assess participants' satisfaction with the perceived usability of products or systems

Advantages of Standardized Questionnaires

Key disadvantage: Lack of diagnostic specificity

- **Objectivity**: Independent verification of measurement
- **Replicability**: Easier to replicate
- **Quantification**: Standard reporting of results and use of standard statistical analyses
- **Economy**: Difficult to develop, but easy to reuse
- **Communication**: Enhances practitioner communication
- **Scientific generalization**: Essential for assessing the generalization of results

What Standardized UX Questionnaires Are Available?



- Historical measurement of satisfaction with computers
- Gallagher Value of MIS Reports Scale, Computer Acceptance Scale
- Post-study questionnaires
- QUIS, SUMI, USE, PSSUQ, SUS,UMUX, UMUX-LITE
- Post-task questionnaires
- ASQ, Expectation Ratings, Usability Magnitude Estimation, SEQ, SMEQ
- Website usability
- WAMMI, SUPR-Q, PWQ, WEBQUAL, PWU, WIS, ISQ
- Other questionnaires
- CSUQ, AttrakDiff, UEQ, meCUE, EMO, ACSI, NPS, CxPi, TAM

Assessing Standardized Questionnaire Quality

Possible: High reliability with low validity

Not possible: High validity with low reliability

- Reliability
  - Typically measured with coefficient alpha (0 to 1)
  - For research/evaluation, goal > .70
- Validity
  - Content validity (where do items come from?)
  - Concurrent or predictive correlation (-1 to 1)
  - Factor analysis (construct validity, subscale development)
- Sensitivity
  - t- or F-test with significant outcome(s), either main effects or interactions
  - Minimum sample size needed to achieve significance

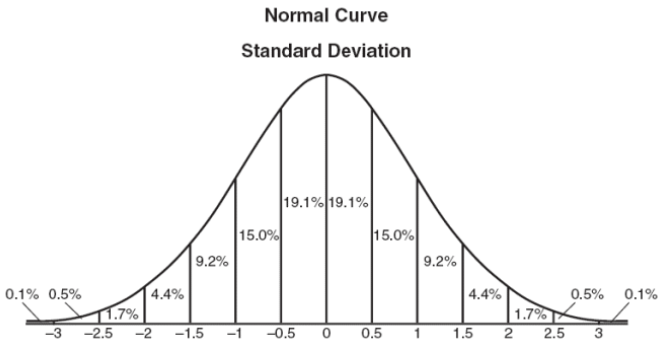
Scale Items

In general, any common item design is OK

But scale designers have to make a choice for standardization

- Number of scale steps
  - More steps increases reliability with diminishing returns
  - No practical difference for 7-, 11- and 101-point items
  - Very important for single-item instruments, less important for multi-item
- Forced choice
  - Odd number of steps or providing NA choice provides neutral point
  - Even number forces choice
  - Most standardized usability questionnaires do not force choice
- Item types
  - Likert (most common) – agree/disagree with statement
  - Item-specific – endpoints have opposing labels (e.g., “confusing” vs. “clear”)

Norms



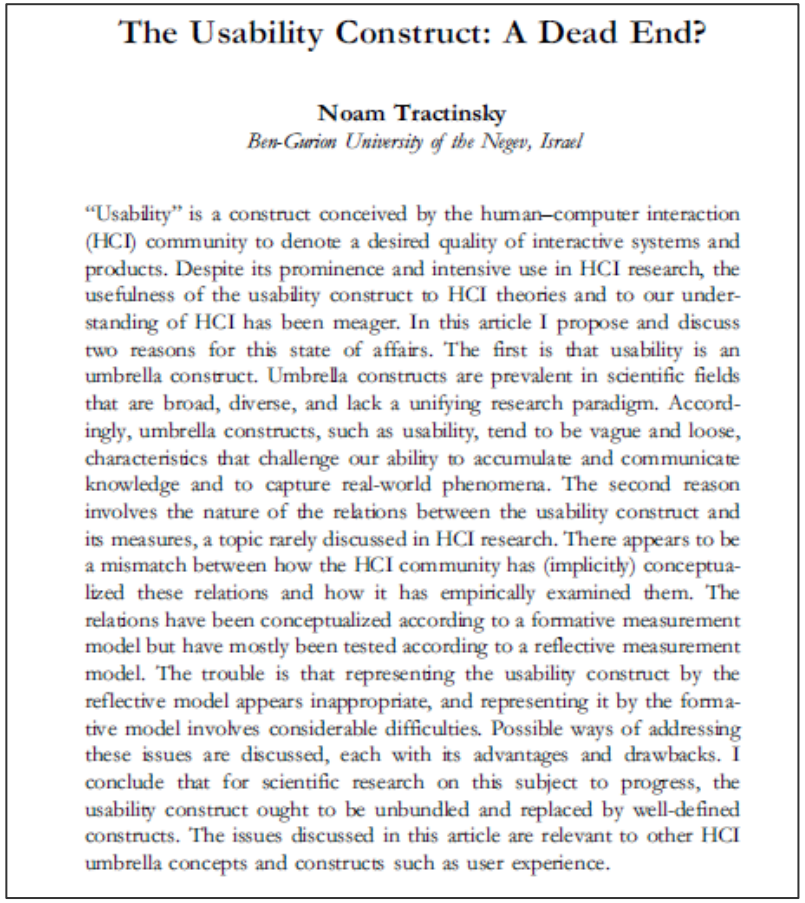
- By itself, a score (individual or average) has no meaning
- One way to provide meaning is through comparison (t- or F-test)
  - Comparison against a benchmark
  - Comparison of two sets of data (different products, different user groups, etc.)
- Another is comparison with norms
  - Normative data is collected from a representative group
  - Comparison with norms allows assessment of how good or bad a score is
  - Always a risk that the new sample doesn't match the normative sample – be sure you understand where the norms came from

Post-Study Questionnaires: Perceived Usability

Which one(s) (if any) do you use?

- QUIS: Questionnaire for User Interaction Satisfaction
- SUMI: Software Usability Measurement Inventory
- PSSUQ: Post-Study/Computer System Usability Questionnaire
- CSUQ: Computer System Usability Questionnaire
- SUS: System Usability Scale
- UMUX(-LITE): Usability Metric for User Experience
- SUPR-Q: Standardized UX Percentile Rank Questionnaire
- AttrakDiff: AttrakDiff
- UEQ: User Experience Questionnaire

Criticism of the Construct of Perceived Usability



The Usability Construct: A Dead End?

Noam Tractinsky
Ben-Gurion University of the Negev, Israel

"Usability" is a construct conceived by the human–computer interaction (HCI) community to denote a desired quality of interactive systems and products. Despite its prominence and intensive use in HCI research, the usefulness of the usability construct to HCI theories and to our understanding of HCI has been meager. In this article I propose and discuss two reasons for this state of affairs. The first is that usability is an umbrella construct. Umbrella constructs are prevalent in scientific fields that are broad, diverse, and lack a unifying research paradigm. Accordingly, umbrella constructs, such as usability, tend to be vague and loose, characteristics that challenge our ability to accumulate and communicate knowledge and to capture real-world phenomena. The second reason involves the nature of the relations between the usability construct and its measures, a topic rarely discussed in HCI research. There appears to be a mismatch between how the HCI community has (implicitly) conceptualized these relations and how it has empirically examined them. The relations have been conceptualized according to a formative measurement model but have mostly been tested according to a reflective measurement model. The trouble is that representing the usability construct by the reflective model appears inappropriate, and representing it by the formative model involves considerable difficulties. Possible ways of addressing these issues are discussed, each with its advantages and drawbacks. I conclude that for scientific research on this subject to progress, the usability construct ought to be unbundled and replaced by well-defined constructs. The issues discussed in this article are relevant to other HCI umbrella concepts and constructs such as user experience.

- Tractinsky (2018) argued against usefulness of construct of usability in general – reaction to the paper was mixed
- It offered valuable arguments regarding difficulty of measuring usability and UX
- The arguments were not accepted as the final word on the topic – e.g., see 11/2018 JUS essay
- Tractinsky cited the Technology Acceptance Model (TAM) as a good example of the use of constructs in science and practice
- This led to investigation of the relationship between perceived usability and TAM

The UMUX-LITE: History and Research

Perceived Usability and the Modified Technology Acceptance Model

Urška Lah[a,b], James R. Lewis [c], and Boštjan Šumak[b]

[a]EMO - Orodjarna Proizvodna družba, d. o. o, Celje, Slovenia; [b]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia; [c]MeasuringU, Denver, Colorado, USA

ABSTRACT

In response to recent criticism of the usefulness of the construct of usability, we investigated the relationships between measures of perceived usability and the components of a modified version of the Technology Acceptance Model (mTAM) – Perceived Usefulness (PU) and Perceived Ease-of-Use (PEU). In three surveys, respondents used SUS, UMUX-LITE and mTAM to rate their actual (as opposed to expected) experience with three software products. As expected, the correlations between PEU and other measures of perceived usability tended to be significantly stronger than those with PU. Additional findings support the use of the UMUX-LITE as a compact measure of perceived usability that has a strong relationship to the mTAM and strong correspondence with concurrently collected SUS scores. The main theoretical result of this research were regression results providing evidence that the PEU component of the mTAM appears to be another measure of the construct of perceived usability, connecting the TAM to the construct of perceived usability through the mTAM and providing evidence against the claim that the construct of usability is a theoretical dead end.

1. Introduction

1.1. Perceived usability and technology acceptance

In addition to the objective components of efficiency and effectiveness, perceived usability is an important component of the classical conception of usability (Brooke, 2013; ISO, 1998; Lewis, Utesch, & Maher, 2013, 2015; Sauro & Lewis, 2009, 2016), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). The first standardized usability questionnaires intended for application in usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Around the same time that usability researchers were producing the first standardized questionnaires to assess perceived usability, market researchers who studied the adoption of information systems were addressing similar issues. Of these, one of the most influential has been the Technology Acceptance Model (TAM, Davis, 1989). According to TAM, the primary factors that affect a user's intention to use a technology are its perceived usefulness (PU) and perceived ease of use (PEU). This model addressed early criticism of focusing only on usability without consideration of whether a product or system was useful (Pearson & Bailey, 1980).

A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007). In the TAM, PU is the extent to which a person believes a technology will enhance job performance, and PEU is the extent to which a person believes that using the technology will be effortless. The more someone holds these beliefs before use, the greater their intention to use, and the more likely they are to try the technology. Figure 1 illustrates this model, and shows the wording of the items that Davis (1989) used to measure its constructs.

Recently, Tractinsky (2018), in a paper entitled "The Usability Construct: A Dead End?", argued against the usefulness of the construct of usability as a part of theory construction in human-computer interaction, in part due to "the inadequate modeling of the relations between the construct and its measures" (p. 133). In the same paper, he later cited the TAM as a good example of the use of constructs in scientific and practical model building, writing:

> Constructs contribute to a theory if they add to our understanding of the phenomenon under study. For example, the general domain that serves as the background for the emergence of the usability construct, namely, the use (often termed adoption or acceptance) of information technology, is a point of contact with various other theories. A notable such theory is the technology acceptance model (TAM; Davis, Bagozzi, & Warshaw, 1989). In TAM, the constructs "perceived ease-of-use" (a close relative of the intuitive meaning of "usability") and "perceived usefulness" are instrumental in explaining variations in the construct "behavioral intention" (people's intention to use a certain information system or product). Unfortunately, as previously mentioned, it is hard to name any influential theory in which the construct of "usability" plays a similarly useful role. (Tractinsky, 2018, p. 141)

The general reaction to the Tractinsky paper was that it offered valuable arguments regarding the difficulty of measuring usability and user experience, but those arguments were not universally accepted as the final word on the topic, especially with regard to the usefulness of usability as a construct

- Need to know research on related measures
- System Usability Scale (SUS) – well-known measure of perceived usability
- Technology Adoption Model (TAM) – information systems research
- Net Promoter Score (NPS) – market research measure based on likelihood-to-recommend
- Usability Metric for User Experience (UMUX) – short measure designed as alternative to SUS
- Need to know UMUX-LITE research
- Origin
- Psychometric properties
- Correspondence with SUS
- Relationship to TAM
- UMUX-LITE vs. NPS

The System Usability Scale (SUS)

No license required for use – cite the source

Brooke (1996) – as of 4/2/20 had 8,736 Google Scholar citations

- Developed in mid-80s by John Brooke at DEC
- Probably most popular post-study questionnaire (PSQ)
- Accounts for about 43% of PSQ usage (Sauro & Lewis, 2009)
- Self-described “quick and dirty”
- Fairly quick, but apparently not that dirty
- Psychometric quality
- Initial publication – n = 20 – now there are >10,000
- Unidimensional measure of perceived usability
- Good reliability – coefficient alpha usually around .92
- Good concurrent validity – e.g., high correlations with concurrently collected ratings of likelihood to recommend (.75) and overall experience (.80)

The System Usability Scale (SUS)

It's OK to replace "cumbersome" with "awkward" and make reasonable replacements for "system"

Align items to 0-4 scale:

Pos: $x_i - 1$

Neg: $5 - x_i$

Then sum & multiply by 2.5 (100/40)

	The System Usability Scale Standard Version		Strongly Disagree 1	2	3	4	Strongly Agree 5
1	I think that I would like to use this system frequently.		O	O	O	O	O
2	I found the system unnecessarily complex.		O	O	O	O	O
3	I thought the system was easy to use.		O	O	O	O	O
4	I think that I would need the support of a technical person to be able to use this system.		O	O	O	O	O
5	I found the various functions in this system were well integrated.		O	O	O	O	O
6	I thought there was too much inconsistency in this system.		O	O	O	O	O
7	I would imagine that most people would learn to use this system very quickly.		O	O	O	O	O
8	I found the system very cumbersome to use.		O	O	O	O	O
9	I felt very confident using the system.		O	O	O	O	O
10	I needed to learn a lot of things before I could get going with this system.		O	O	O	O	O

The Sauro-Lewis Curved Grading Scale for the SUS

SUS Score Range	Grade	Grade Point	Percentile Range
84.1 - 100	A+	4.0	96-100
80.8 - 84.0	A	4.0	90-95
78.9 - 80.7	A-	3.7	85-89
77.2 - 78.8	B+	3.3	80-84
74.1 - 77.1	B	3.0	70-79
72.6 - 74.0	B-	2.7	65-69
71.1 - 72.5	C+	2.3	60-64
65.0 -71.0	C	2.0	41-59
62.7 - 64.9	C-	1.7	35-40
51.7 - 62.6	D	1.0	15-34
0.0 - 51.6	F	0.0	0-14

From Sauro & Lewis (2016, Table 8.5)

Based on data from 446 usability studies/surveys

SUS Ratings for Everyday Products

Product	95% CI Lower Limit	Mean (Grade)	95% CI Upper Limit	Sauro-Lewis Grade Range	Std Dev	n
Excel	55.3	**56.5 (D)**	57.7	**D to D**	18.6	866
GPS	68.5	**70.8 (C)**	73.1	**C to B-**	18.3	252
DVR	71.9	**74.0 (B-)**	76.1	**C+ to B**	17.8	276
PowerPoint	73.5	**74.6 (B)**	75.7	**B- to B**	16.6	867
Word	75.3	**76.2 (B)**	77.1	**B to B**	15	968
Wii	75.2	**76.9 (B)**	78.6	**B to B+**	17	391
iPhone	76.4	**78.5 (B+)**	80.6	**B to A-**	18.3	292
Amazon	80.8	**81.8 (A)**	82.8	**A to A**	14.8	801
ATM	81.1	**82.3 (A)**	83.5	**A to A**	16.1	731
Gmail	82.2	**83.5 (A)**	84.8	**A to A+**	15.9	605
Microwaves	86.0	**86.9 (A+)**	87.8	**A+ to A+**	13.9	943
Landline phone	86.6	**87.7 (A+)**	88.8	**A+ to A+**	12.4	529
Browser	87.3	**88.1 (A+)**	88.9	**A+ to A+**	12.2	980
Google search	92.7	**93.4 (A+)**	94.1	**A+ to A+**	10.5	948

Based on Kortum & Bangor (2013, Table 2) – Mostly best in class products

The Technology Acceptance Model (TAM)

12 positive-tone items

Two factors

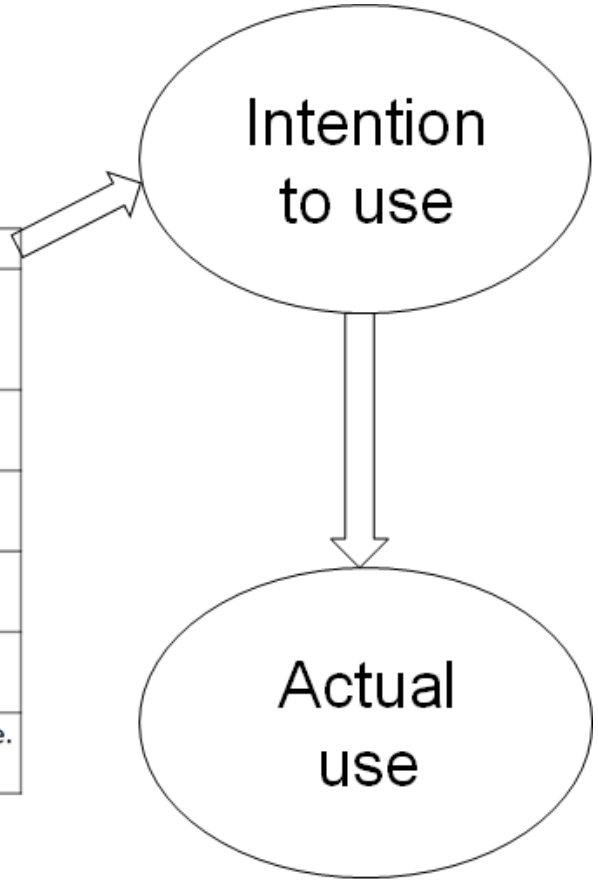
Perceived Usefulness

Perceived Ease of Use

- Developed by Davis (1989)
  - Developed during same period as first standardized usability questionnaires
  - Information Systems (IS) researchers dealing with similar issues
  - Influential in market and IS research (e.g., Sauro, 2019a; Wu et al., 2007)
  - Perceived usefulness/ease-of-use > intention to use > actual use
- Psychometric evaluation
  - Started with 14 items per construct – ended with 6
  - Started with mixed tone – due to structural issues, ended with all positive
  - Reliability: PU (.98); PEU (.94)
  - Factor analysis showed expected item-factor alignment
  - Concurrent validity with predicted likelihood of use (PU: .85; PEU: .59)

The Technology Acceptance Model (TAM)

Perceived Usefulness	Perceived Ease-of-Use
Using [this product] in my job would enable me to accomplish tasks more quickly.	Learning to operate [this product] would be easy for me.
Using [this product] would improve my job performance.	I would find it easy to get [this product] to do what I want it to do.
Using [this product] in my job would increase my productivity.	My interaction with [this product] would be clear and understandable.
Using [this product] would enhance my effectiveness on the job.	I would find [this product] to be flexible to interact with.
Using [this product] would make it easier to do my job.	It would be easy for me to become skillful at using [this product].
I would find [this product] useful in my job.	I would find [this product] easy to use.



Using [this product] in my job would enable me to accomplish tasks more quickly.

likely | extremely | quite | slightly | neither | slightly | quite | extremely | unlikely

Item content and format from Davis (1989)

The Net Promoter Score (NPS)

- Introduced in 2003 by Fred Reichheld
- Net Promoter Score, Net Promoter and NPS are registered trademarks of Bain & Company, Satmetrix Systems and Fred Reichheld
- Popular metric of customer loyalty, based on likelihood to recommend

	Not at all Likely					Neutral					Extremely Likely
	0	1	2	3	4	5	6	7	8	9	10
How likely are you to recommend this website to a friend or colleague?	O	O	O	O	O	O	O	O	O	O	O

- Computing NPS
- Type of top-box-minus-bottom-box metric
- Respondents rate likelihood to recommend (LTR) using 11-point scale
- Ratings of 9-10 are promoters; 0-6 are detractors; 7-8 are passives
- NPS is the percentage of promoters minus the percentage of detractors
- NPS can range from -100 to +100

The Usability Metric for User Experience (UMUX)

No license required

Best source for citation is Finstad (2010)

- Developed by Kraig Finstad at Intel
- Published in 2010
- Designed to act as four-item proxy for SUS
- Items based on ISO definition of usability
- Psychometric evaluation
- Initial pool of 12 items (item analysis n = 42)
- Selected best three for effectiveness, efficiency, satisfaction (highest SUS r)
- Collected SUS and UMUX data for two systems (total n = 558)
- High reliability: .94
- Concurrent validity correlation with SUS: .96
- Sensitive to large system differences
- Replicated by Lewis et al. (2013) – lower values but still impressive

The Usability Metric for User Experience (UMUX)

	The Usability Metric for User Experience Version 1	Strongly Disagree							Strongly Agree
			1	2	3	4	5	6	7
1	This system's capabilities meet my requirements.		O	O	O	O	O	O	O
2	Using this system is a frustrating experience.		O	O	O	O	O	O	O
3	This system is easy to use.		O	O	O	O	O	O	O
4	I have to spend too much time correcting things with this system.		O	O	O	O	O	O	O

- Four 7-point scales (alternating tone)
- Labeled from 1 (strongly disagree) to 7 (strongly agree)
- Like SUS, need to recode to 0-6 scale where larger number is better
- Sum the item scores, multiply by 100, then divide by 24 (4 x 6)
- Final UMUX scores range from 0 to 100

Cutting the UMUX in Half – The UMUX-LITE

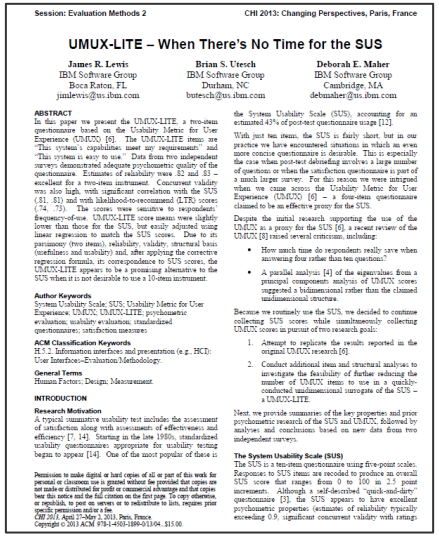
No license required

Best source for citation is Lewis, Utesch, and Maher (2013)

- Derived from UMUX by Lewis et al. (2013)
- Concerns with UMUX structure – apparent bidimensionality with 4 items
- Known usability issues with mixed-tone questionnaires (Sauro & Lewis, 2011)
- Possible to reduce items to get even more concise instrument?
- Current version
- Two 7-point UMUX items (those with positive tone)
- Content consistent with Technology Acceptance Model (useful and easy)
- Aligned in factor analysis of UMUX
- Highest correlations with SUS (both versions)

	The UMUX-LITE Version 1	Strongly Disagree							Strongly Agree
			1	2	3	4	5	6	7
1	This system's capabilities meet my requirements.		o	o	o	o	o	o	o
2	This system is easy to use.		o	o	o	o	o	o	o

UMUX-LITE Psychometric Evaluation


Session: Evaluation Methods 2 CHI 2013: Changing Perspectives, Paris, France

UMUX-LITE – When There's No Time for the SUS

James R. Lewis, IBM Software Group, Boca Raton, FL, jimlewis@us.ibm.com
Brian S. Utesch, IBM Software Group, Durham, NC, butesch@us.ibm.com
Deborah E. Maher, IBM Software Group, Cambridge, MA, debmaher@us.ibm.com

ABSTRACT
In this paper we present the UMUX-LITE, a two-item questionnaire based on the Usability Metric for User Experience (UMUX) [6]. The UMUX-LITE items are "This system's capabilities meet my requirements" and "This system is easy to use." Data from two independent surveys demonstrated adequate psychometric quality of the questionnaire. Estimates of reliability were .82 and .83 – excellent for a two-item instrument. Concurrent validity was also high, with significant correlation with the SUS (.81, .81) and with likelihood-to-recommend (LTR) scores (.74, .73). The scores were sensitive to respondents' frequency-of-use. UMUX-LITE score means were slightly lower than those for the SUS, but easily adjusted using linear regression to match the SUS scores. Due to its parsimony (two items), reliability, validity, structural basis (usefulness and usability) and, after applying the corrective regression formula, its correspondence to SUS scores, the UMUX-LITE appears to be a promising alternative to the SUS when it is not desirable to use a 10-item instrument.

Author Keywords
System Usability Scale; SUS; Usability Metric for User Experience; UMUX; UMUX-LITE; psychometric evaluation; usability evaluation; standardized questionnaires; satisfaction measures

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces—Evaluation/Methodology.

General Terms
Human Factors; Design; Measurement.

INTRODUCTION

Research Motivation
A typical summative usability test includes the assessment of satisfaction along with assessment of effectiveness and efficiency [7, 14]. Starting in the late 1980s, standardized usability questionnaires appropriate for usability testing began to appear [14]. One of the most popular of these is the System Usability Scale (SUS), accounting for an estimated 43% of post-test questionnaire usage [12].

With just ten items, the SUS is fairly short, but in our practice we have encountered situations in which an even more concise questionnaire is desirable. This is especially the case when post-test debriefing involves a large number of questions or when the satisfaction questionnaire is part of a much larger survey. For this reason we were intrigued when we came across the Usability Metric for User Experience (UMUX) [6] – a four-item questionnaire claimed to be an effective proxy for the SUS.

Despite the initial research supporting the use of the UMUX as a proxy for the SUS [6], a recent review of the UMUX [8] raised several criticisms, including:

- How much time do respondents really save when answering four rather than ten questions?
- A parallel analysis [4] of the eigenvalues from a principal components analysis of UMUX scores suggested a bidimensional rather than the claimed unidimensional structure.

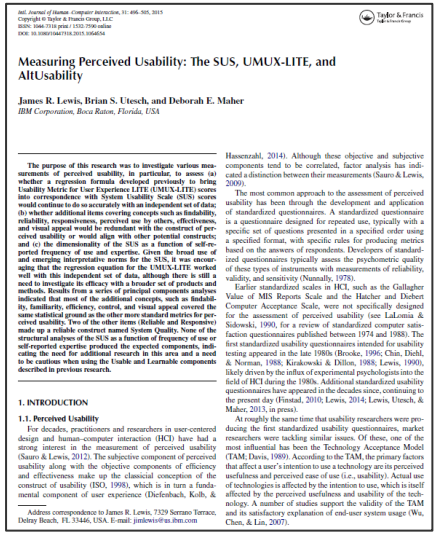
Because we routinely use the SUS, we decided to continue collecting SUS scores while simultaneously collecting UMUX scores in pursuit of two research goals:

1. Attempt to replicate the results reported in the original UMUX research [8].
2. Conduct additional item and structural analyses to investigate the feasibility of further reducing the number of UMUX items to use in a quickly-conducted unidimensional surrogate of the SUS – a UMUX-LITE.

Next, we provide summaries of the key properties and prior psychometric research of the SUS and UMUX, followed by analyses and conclusions based on new data from two independent surveys.

The System Usability Scale (SUS)
The SUS is a ten-item questionnaire using five-point scales. Responses to SUS items are recoded to produce an overall SUS score that ranges from 0 to 100 in 2.5 point increments. Although a self-described "quick-and-dirty" questionnaire [3], the SUS appears to have excellent psychometric properties (estimates of reliability typically exceeding 0.9, significant concurrent validity with ratings

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright © 2013 ACM. 978-1-4503-1899-0/13/04...$15.00.


Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability

James R. Lewis, Brian S. Utesch, and Deborah E. Maher
IBM Corporation, Boca Raton, Florida, USA

The purpose of this research was to investigate various measurements of perceived usability, in particular, to assess (a) whether a regression formula developed previously to bring Usability Metric for User Experience LITE (UMUX-LITE) scores into correspondence with System Usability Scale (SUS) scores would continue to do so accurately with an independent set of data; (b) whether additional items covering concepts such as findability, reliability, responsiveness, perceived use by others, effectiveness, and visual appeal would be redundant with the construct of perceived usability or would align with other potential constructs; and (c) the dimensionality of the SUS as a function of self-reported frequency of use and expertise. Given the broad use of and emerging interpretative norms for the SUS, it was encouraging that the regression equation for the UMUX-LITE worked well with this independent set of data, although there is still a need to investigate its efficacy with a broader set of products and methods. Results from a series of principal components analyses indicated that most of the additional concepts, such as findability, reliability, responsiveness, control, and visual appeal covered the same statistical ground as the other more standard metrics for perceived usability. Two of the other items (Reliable and Responsive) made up a reliable construct named System Quality. None of the structural analyses of the SUS as a function of frequency of use or self-reported expertise produced the expected components, indicating the need for additional research in this area and a need to be cautious when using the Usable and Learnable components described in previous research.

1. INTRODUCTION

1.1. Perceived Usability

For decades, practitioners and researchers in user-centered design and human–computer interaction (HCI) have had a strong interest in the measurement of perceived usability (Sauro & Lewis, 2012). The subjective component of perceived usability along with the objective components of efficiency and effectiveness make up the classical conception of the construct of usability (ISO, 1998), which is in turn a fundamental component of user experience (Diefenbach, Kolb, & Hassenzahl, 2014). Likely driven by the influx of experimental psychologists into the field in the early 1980s, the first standardized usability questionnaires intended for usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Two of the most widely-used standardized usability questionnaires with roots in the 1980s are the System Usability Scale (SUS; Brooke, 1996) and the Computer System Usability Questionnaire (CSUQ; Lewis, 1995). The Usability Metric for User Experience (UMUX; Finstad, 2010) and metrics derived from the UMUX (UMUX-LITE and UMUX-LITE+; Lewis, Utesch, & Maher, 2013, 2015) are more recently published usability questionnaires. Previous research has generally shown that the UMUX and associated metrics provide scores very similar to the SUS, but with fewer items (the SUS has ten items, the UMUX has four, and UMUX-LITE has two), and that, after appropriate transformation, the 16-item CSUQ also produces SUS-like scores (Lewis, 2018a, 2018b).

Address correspondence to James R. Lewis, 7329 Serrano Terrace, Delray Beach, FL 33446, USA. E-mail: jimlewis@us.ibm.com

Hassenzahl, 2014). Although these objective and subjective components tend to be correlated, factor analysis has indicated a distinction between their measurements (Sauro & Lewis, 2009).

The most common approach to the assessment of perceived usability has been through the development and application of standardized questionnaires. A standardized questionnaire is a questionnaire designed for repeated use, typically with a specific set of questions presented in a specified order using a specified format, with specific rules for producing metrics based on the answers of respondents. Developers of standardized questionnaires typically assess the psychometric quality of these types of instruments with measurements of reliability, validity, and sensitivity (Nunnally, 1978).

Earlier standardized scales in HCI, such as the Gallagher Value of MIS Reports Scale and the Hatcher and Diebert Computer Attitude Scale, were not specifically designed for the assessment of perceived usability (see LaLomia & Sidowski, 1990, for a review of standardized computer satisfaction questionnaires published between 1974 and 1988). The first standardized usability questionnaires intended for usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990), likely driven by the influx of experimental psychologists into the field of HCI during the 1980s. Additional standardized usability questionnaires have appeared in the decades since, continuing to the present day (Finstad, 2010; Lewis, 2014; Lewis, Utesch, & Maher, 2013, in press).

At roughly the same time that usability researchers were producing the first standardized usability questionnaires, market researchers were tackling similar issues. Of these, one of the most influential has been the Technology Acceptance Model (TAM; Davis, 1989). According to the TAM, the primary factors that affect a user's intention to use a technology are its perceived usefulness and perceived ease of use (i.e., usability). Actual use of technologies is affected by the intention to use, which is itself affected by the perceived usefulness and usability of the technology. A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007).


Measuring Perceived Usability: The CSUQ, SUS, and UMUX

James R. Lewis
IBM Corporation, Boca Raton, FL, USA

ABSTRACT
The primary purpose of this research was to investigate the relationship between two widely used questionnaires designed to measure perceived usability: the Computer System Usability Questionnaire (CSUQ) and the System Usability Scale (SUS). The correlation between concurrently collected CSUQ and SUS scores was 0.76 (over 50% shared variance). After converting CSUQ scores to a 0–100-point scale (to match the range of the SUS scores), there was a small but statistically significant difference between CSUQ and SUS means. Although this difference (just under 2 scale points out of a possible 100) was statistically significant, it did not appear to be practically significant. Although usability practitioners should be cautious pending additional independent replication, it appears that CSUQ scores, after conversion to a 0–100-point scale, can be interpreted with the Sauro-Lewis curved grading scale. As a secondary research goal, investigation of variations of the Usability Metric for User Experience (UMUX) replicated previous findings that the regression-adjusted version of the UMUX-LITE (UMUX-LITE+) had the closest correspondence with concurrently collected SUS scores. Thus, even though these three standardized questionnaires were independently developed and have different item content and formats, they largely appear to be measuring the same thing, presumably, perceived usability.

1. Introduction

1.1. Perceived usability

An important component of the higher-level construct of usability is perceived usability (Brooke, 2013; Lewis, Utesch, & Maher, 2015; Sauro & Lewis, 2009, 2016). The subjective component of perceived usability, along with the objective components of efficiency and effectiveness, makes up the classical conception of the construct of usability (ISO, 1998), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). Likely driven by the influx of experimental psychologists into the field in the early 1980s, the first standardized usability questionnaires intended for usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Two of the most popular standardized questionnaires used to assess perceived usability are the Computer System Usability Questionnaire (CSUQ; Lewis, 1995) and the System Usability Scale (SUS; Brooke, 1996). They were independently developed in the 1980s at, respectively, IBM and DEC, and published in the mid-1990s. Although both are widely used, over the years the SUS has become the more popular questionnaire. Sauro and Lewis (2009) reported that the SUS accounted for 43% of post-study questionnaire usage in a study of unpublished usability studies while the CSUQ accounted for about 15%. Google Scholar citations (examined 9/12/2017) showed 5015 citations for the paper that introduced the SUS (Brooke, 1996) and 1693 for the paper that introduced the CSUQ (Lewis, 1995). These independent measurements of questionnaire "popularity" show that the CSUQ is popular, but the SUS is almost three times as popular.

The primary goal of this research was to investigate whether the CSUQ and the SUS appear to be measuring the same thing or something different. Despite these questionnaires having been available in the public domain for over 20 years, there has been very little investigation of their psychometric relationship. This is an issue of both theoretical and practical significance.

From a theoretical perspective, if the CSUQ and SUS – possibly the two most popular instruments for the assessment of perceived usability – do not appear to substantially measure the same thing, then what does this mean for the validity of the construct of usability, which has been recently questioned (Tractinsky, 2017)? From a practical perspective, if they do essentially measure the same thing, then practitioners should be able to put data from both questionnaires on a common scale for interpretation as indicating relatively poor, average, or good levels of perceived usability based on grading scale norms developed for the SUS (Sauro & Lewis, 2016).

A secondary goal of this research was to investigate the relationship between the SUS and concurrently collected scores from the Usability Metric for User Experience (UMUX; Finstad, 2010) and metrics derived from the UMUX (UMUX-LITE and UMUX-LITE+; Lewis, Utesch, & Maher, 2013, 2015). Previous research has generally shown that the UMUX and associated metrics provide scores very similar to the SUS, but with fewer items (the SUS has 10 items, the UMUX has 4, and UMUX-LITE has 2). Although not yet as popular as the CSUQ or SUS (144 citations of Finstad, 2010 in Google Scholar on 9/12/2017), the UMUX-related metrics have the potential to become more widely used as researchers and practitioners become more aware of them and, of course, depending on whether research continues to support their use.

CONTACT James R. Lewis jimlewis@us.ibm.com 7329 Serrano Terrace, Delray Beach, FL 33446, USA.


Measuring Perceived Usability: SUS, UMUX, and CSUQ Ratings for Four Everyday Products

James R. Lewis
IBM Hybrid Cloud, IBM Corporation, Boca Raton, FL, USA

ABSTRACT
This research continued previous investigation of the relationships among measures of perceived usability: the System Usability Scale (SUS), three metrics derived from the Usability Metric for User Experience (UMUX), and the Computer System Usability Questionnaire (CSUQ), this time with ratings of four everyday products (Excel, Word, Amazon, and Gmail). SUS ratings of these products were generally consistent with previous reports. Significant differences in SUS means across studies could be due to differences in frequency of use, with implications for using these data as usability benchmarks. Correspondence among the various measures of perceived usability was also consistent with previous research. Considering frequency of use, mean differences ranged from -2.0 to 1.8 (average shift in Sauro-Lewis grade range from -0.6 to 0.8). When SUS scores were above average, the range restriction of the UMUX-LITE+ led to relatively large discrepancies with SUS, suggesting it might not always be better than unadjusted UMUX-LITE.

1. Introduction

1.1. Perceived usability

An important component of the higher-level construct of usability is perceived usability (Brooke, 2013; Lewis, Utesch, & Maher, 2015; Sauro & Lewis, 2009, 2016). The subjective component of perceived usability along with the objective components of efficiency and effectiveness make up the classical conception of the construct of usability (ISO, 1998), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). Likely driven by the influx of experimental psychologists into the field in the early 1980s, the first standardized usability questionnaires intended for usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Two of the most widely-used standardized usability questionnaires with roots in the 1980s are the System Usability Scale (SUS; Brooke, 1996) and the Computer System Usability Questionnaire (CSUQ; Lewis, 1995). The Usability Metric for User Experience (UMUX; Finstad, 2010) and metrics derived from the UMUX (UMUX-LITE and UMUX-LITE+; Lewis, Utesch, & Maher, 2013, 2015) are more recently published usability questionnaires. Previous research has generally shown that the UMUX and associated metrics provide scores very similar to the SUS, but with fewer items (the SUS has ten items, the UMUX has four, and UMUX-LITE has two), and that, after appropriate transformation, the 16-item CSUQ also produces SUS-like scores (Lewis, 2018a, 2018b).

1.2. Perceived usability of everyday products

Kortum and Bangor (2013) published research on retrospective ratings of 14 everyday products and product classes using the SUS (Excel, GPS, DVR, PowerPoint, Word, Wii, iPhone, Amazon, ATM, Gmail, Microwaves, Landline, Browser, and Google Search). They used an online survey to collect data during 2010–2011 from just over 1,000 respondents, with those respondents about evenly split between Rice University undergraduate students and participants recruited using Amazon's Mechanical Turk. There were statistically significant differences in mean SUS ratings between the Rice and Turk groups for eight products, but the means across all 14 products were almost identical (79.2 for Rice; 79.6 for Turk) and the correlation of the scores across products for the two groups was very high ($r(12) = 0.95$, $p < 0.0001$). One of the strongest factors affecting the ratings was the amount of experience the respondent reported having with the product, but the higher the average SUS score, the less experience mattered, indicating that "products that have superior usability are usable by novices and experts alike, whereas hard to use products may get easier over time but never reach superior usability even with heavy experience" (p. 74).

1.3. Research goals

As part of a continuing investigation into the relationships among various measures of perceived usability, a major goal of this study was to replicate the method of Kortum and Bangor (2013) for a subset of their everyday products that were (1) likely to have been used by members of the IBM User Experience panel

- Lewis et al. (2013, 2015, 2018, 2019)
- Multiple surveys (n = 402, 389, 397, 746, 390, 453, 338, 256)
- Acceptable reliability: .83, .82, .86, .79, .76
- Concurrent validity (correlation) with SUS: .81, .85, .83, .74, .86
- Concurrent validity (correlation) with LTR: .73, .74, .72
- **Correspondence of UMUX-LITE with SUS**
- Initial results suggested possibility of improvement through regression
- Latest review of all available concurrently collected data indicates best practice is to use UMUX-LITE without any adjustment
- Correspondence and psychometric properties similar for 5-point version of UMUX-LITE, sometimes used for consistency with SUS format
- When reporting UMUX-LITE, carefully document the version you're using

UMUX-LITE: Latest Research (Lah et al., 2020)

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2020.1727262

Taylor & Francis
Taylor & Francis Group

Check for updates

Perceived Usability and the Modified Technology Acceptance Model

Urška Lah[a,b], James R. Lewis [c], and Boštjan Šumak[b]

[a]EMO - Orodjarna Proizvodna družba, d. o. o, Celje, Slovenia; [b]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia; [c]MeasuringU, Denver, Colorado, USA

ABSTRACT

In response to recent criticism of the usefulness of the construct of usability, we investigated the relationships between measures of perceived usability and the components of a modified version of the Technology Acceptance Model (mTAM) – Perceived Usefulness (PU) and Perceived Ease-of-Use (PEU). In three surveys, respondents used SUS, UMUX-LITE and mTAM to rate their actual (as opposed to expected) experience with three software products. As expected, the correlations between PEU and other measures of perceived usability tended to be significantly stronger than those with PU. Additional findings support the use of the UMUX-LITE as a compact measure of perceived usability that has a strong relationship to the mTAM and strong correspondence with concurrently collected SUS scores. The main theoretical result of this research were regression results providing evidence that the PEU component of the mTAM appears to be another measure of the construct of perceived usability, connecting the TAM to the construct of perceived usability through the mTAM and providing evidence against the claim that the construct of usability is a theoretical dead end.

1. Introduction

1.1. Perceived usability and technology acceptance

In addition to the objective components of efficiency and effectiveness, perceived usability is an important component of the classical conception of usability (Brooke, 2013; ISO, 1998; Lewis, Utesch, & Maher, 2013, 2015; Sauro & Lewis, 2009, 2016), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). The first standardized usability questionnaires intended for application in usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Around the same time that usability researchers were producing the first standardized questionnaires to assess perceived usability, market researchers who studied the adoption of information systems were addressing similar issues. Of these, one of the most influential has been the Technology Acceptance Model (TAM, Davis, 1989). According to TAM, the primary factors that affect a user's intention to use a technology are its perceived usefulness (PU) and perceived ease of use (PEU). This model addressed early criticism of focusing only on usability without consideration of whether a product or system was useful (Pearson & Bailey, 1980).

A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007). In the TAM, PU is the extent to which a person believes a technology will enhance job performance, and PEU is the extent to which a person believes that using the technology will be effortless. The more someone holds these beliefs before use, the greater their intention to use, and the more likely they are to try the technology. Figure 1 illustrates this model, and shows the wording of the items that Davis (1989) used to measure its constructs.

Recently, Tractinsky (2018), in a paper entitled "The Usability Construct: A Dead End?", argued against the usefulness of the construct of usability as a part of theory construction in human-computer interaction, in part due to "the inadequate modeling of the relations between the construct and its measures" (p. 133). In the same paper, he later cited the TAM as a good example of the use of constructs in scientific and practical model building, writing:

> Constructs contribute to a theory if they add to our understanding of the phenomenon under study. For example, the general domain that serves as the background for the emergence of the usability construct, namely, the use (often termed adoption or acceptance) of information technology, is a point of contact with various other theories. A notable such theory is the technology acceptance model (TAM; Davis, Bagozzi, & Warshaw, 1989). In TAM, the constructs "perceived ease-of-use" (a close relative of the intuitive meaning of "usability") and "perceived usefulness" are instrumental in explaining variations in the construct "behavioral intention" (people's intention to use a certain information system or product). Unfortunately, as previously mentioned, it is hard to name any influential theory in which the construct of "usability" plays a similarly useful role. (Tractinsky, 2018, p. 141)

The general reaction to the Tractinsky paper was that it offered valuable arguments regarding the difficulty of measuring usability and user experience, but those arguments were not universally accepted as the final word on the topic, especially with regard to the usefulness of usability as a construct

CONTACT James R. Lewis drjimmfl@gmail.com @drjimmfl; linkedin.com/in/jlewis6; MeasuringU, Denver, Colorado, USA

- Exploration of relationship between measures of perceived usability and TAM
- Three new surveys
  - PowerPoint – English – IBM Panel – n=483
  - Gmail – Slovenian – industrial/academic – n=397
  - Notes – English – IBM Panel – n=546
- Three standardized questionnaires
  - SUS: Standard version
  - UMUX: Standard version
  - mTAM: TAM modified to assess experience rather than intention to use
  - Latin square counterbalancing for order of presenting questionnaires

UMUX-LITE: Latest Research - Psychometrics

Reliability	PowerPoint	Gmail	Notes
SUS	0.91	0.88	0.94
UMUX	0.85	0.79	0.91
LITE	0.73	0.69	0.84
mTAM	0.95	0.95	0.98
PU	0.95	0.93	0.98
PEU	0.95	0.95	0.97

r(SUS)	PowerPoint	Gmail	Notes
LITE	0.82	0.74	0.89
LITE-PU	*0.64*	*0.57*	*0.77*
LITE-PEU	0.80	0.73	0.88
mTAM	0.80	0.70	0.90
PU	*0.61*	*0.52*	*0.83*
PEU	0.84	0.78	0.90

No effects of questionnaire presentation order

- Acceptable levels of reliability
  - UMUX-LITE tends to have lowest reliability, but only has two items
  - Can compensate for this with slightly larger sample sizes
- Items mostly aligned with constructs as expected
  - Parallel analysis: SUS and UMUX one factor; mTAM two factors
  - Misalignment of mTAM06 in Slovenian version
- Convergent/divergent validity
  - All correlations statistically significant, but different magnitudes
  - PU correlations with SUS lower than PEU correlations with SUS

UMUX-LITE: Latest Research - Regressions

Predicting (Study 1: PowerPoint)	R^2adj	Beta 1	Beta 2
LTR with PU and PEU	65%	0.446	0.446
LTR with LITE-PU and LITE-PEU	56%	0.486	0.355
LTR with PU and SUS	67%	0.436	0.477
OverExp with PU and PEU	69%	0.314	0.570
OverExp with LITE-PU and LITE-PEU	61%	0.429	0.448
OverExp with PU and SUS	72%	0.342	0.593

Predicting (Study 2: Gmail)	R^2adj	Beta 1	Beta 2
LTR with PU and PEU	43%	.342	.386
LTR with LITE-PU and LITE-PEU	38%	.326	.382
LTR with PU and SUS	46%	.386	.394
OverExp with PU and PEU	46%	.271	.474
OverExp with LITE-PU and LITE-PEU	44%	.341	.420
OverExp with PU and SUS	49%	.330	.471

Predicting (Study 3: Notes)	R^2adj	Beta 1	Beta 2
LTR with PU and PEU	82%	0.483	0.458
LTR with LITE-PU and LITE-PEU	76%	0.361	0.575
LTR with PU and SUS	83%	0.450	0.503
OverExp with PU and PEU	88%	0.533	0.442
OverExp with LITE-PU and LITE-PEU	82%	0.475	0.499
OverExp with PU and SUS	88%	0.528	0.453

- All regression models significant
- Reasonably consistent across surveys
- Highest R^2 for Notes; lowest for Gmail
- Possibly due to different levels of choice in using
- Substituting SUS for PEU
- Models almost identical – SUS and PEU interchangeable
- PEU another measure of the construct of perceived usability
- Substituting UMUX-LITE items for TAM
- Similar regression models
- Slightly smaller coefficients of determination (R^2)

UMUX-LITE: Latest Research - Correspondence

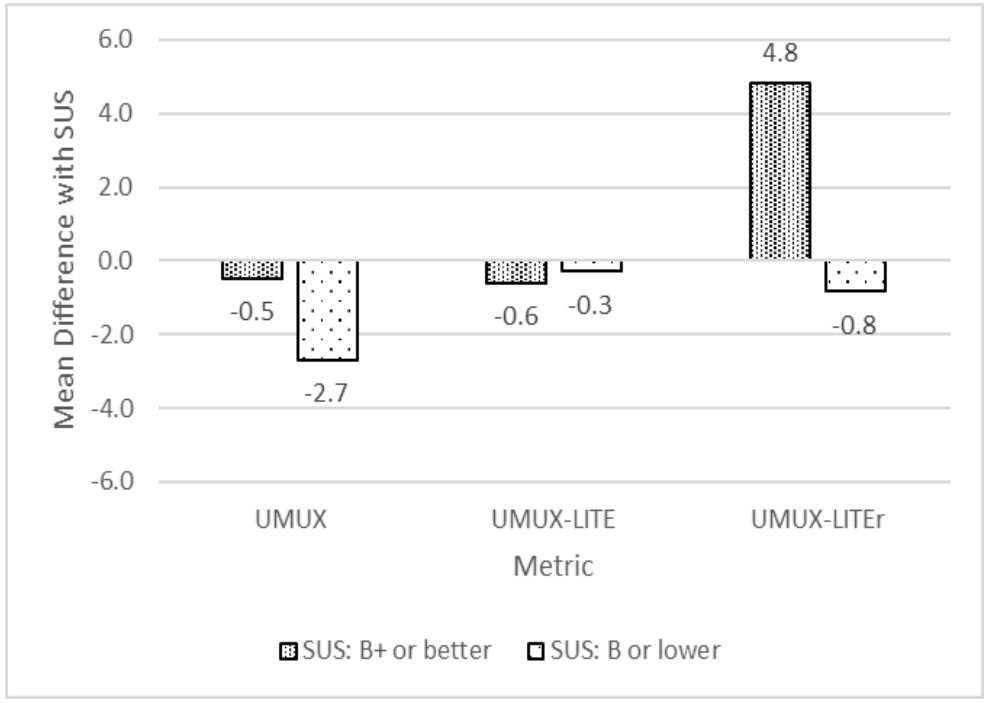
Product (Study)	SUS Mean	UMUX-LITE Mean	Mean Diff	SUS CGS	UMUX-LITE CGS	SUS GPA	UMUX-LITE GPA	GPA Diff
Mind Maps (Berkman & Karahoca, 2016)	79.5	78.5	1.0	A-	B+	3.7	3.3	0.4
PowerPoint (Lah et al., 2020)	70.8	74.3	-3.5	C	B	2.0	3.0	-1.0
Gmail (Lah et al., 2020)	79.3	81.2	-1.9	B+	A	3.7	4.0	-0.3
Notes (Lah et al., 2020)	56.8	59.3	-2.5	D	D	1.0	1.0	0.0
Apple OS (Lewis, 2018b)	76.8	79.9	-3.1	B	A-	3.0	3.7	-0.7
Windows OS (Lewis, 2018b)	66.9	68.5	-1.6	C	C	2.0	2.0	0.0
Excel (Lewis, 2019a)	69.6	74.0	-4.4	C	B-	2.0	2.7	-0.7
Word (Lewis, 2019a)	75.5	78.0	-2.5	B	B+	3.0	3.3	-0.3
Amazon (Lewis, 2019a)	84.8	86.6	-1.8	A+	A+	4.0	4.0	0.0
Gmail (Lewis, 2019a)	78.0	77.7	0.3	B+	B+	3.3	3.3	0.0
Various (Lewis et al., 2013)	53.5	50.3	3.2	D	F	1.0	0.0	1.0
Various (Lewis et al., 2013)	58.8	55.1	3.7	D	D	1.0	1.0	0.0
Various (Lewis et al., 2015)	58.1	52.4	5.7	D	D	1.0	1.0	0.0

Mean difference for SUS - UMUX-LITE: -0.57 (95% CI: -2.45 to 1.31)

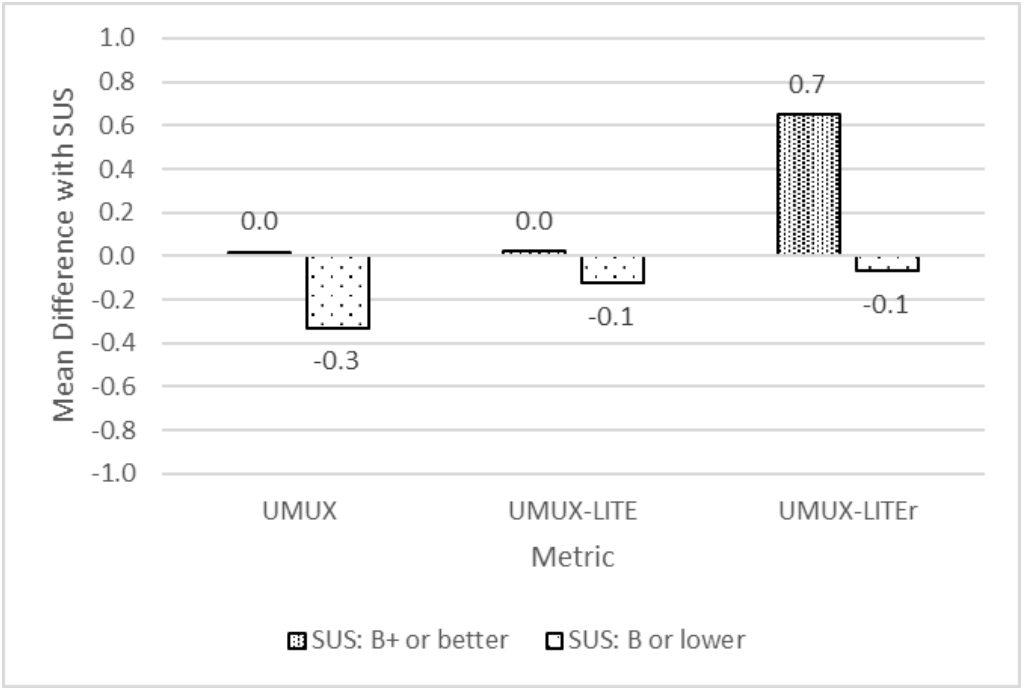
Mean GPA difference: -0.12 (95% CI: -0.43 to 0.19)

CIs narrow; 0 plausible; large differences not plausible

UMUX-LITE: Latest Research - Correspondence



Score Correspondence



GPA Correspondence

Based on 13 independent estimates of correspondence with SUS

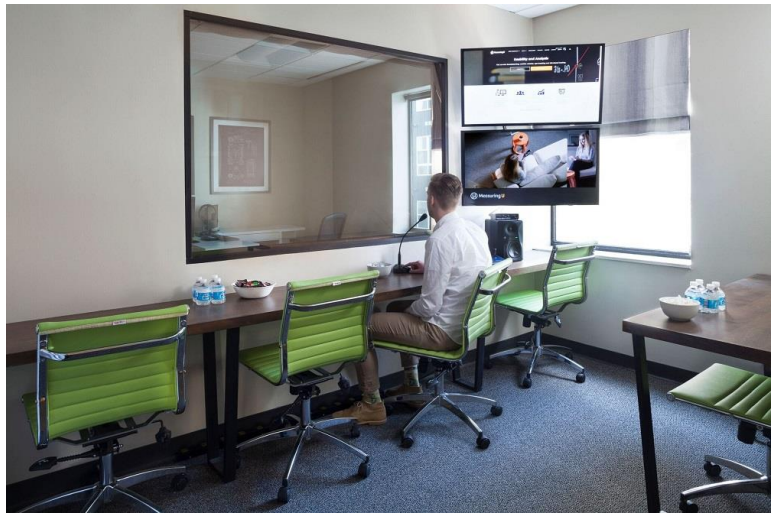
Wide range of CGS grade levels from D to A+

Best correspondence is with unadjusted UMUX-LITE

When to Use the UMUX-LITE

- As ultra-short standardized measure of perceived usability
- As ultra-short proxy for TAM-like measure of UX – one item for PU and one for PEU
- As easily-understood business metric to use in place of or in addition to NPS, especially when users are unlikely to engage in recommendation behavior
- Especially useful in surveys when there is limited “real estate” for global measurement of UX
- Consider using it in usability studies in combination with the SUS, using UMUX-LITE between tasks and SUS at the end
- If currently using the SUS and interested in replacing the SUS with the UMUX-LITE, use them concurrently for some period of time to ensure their correspondence in your context of measurement.

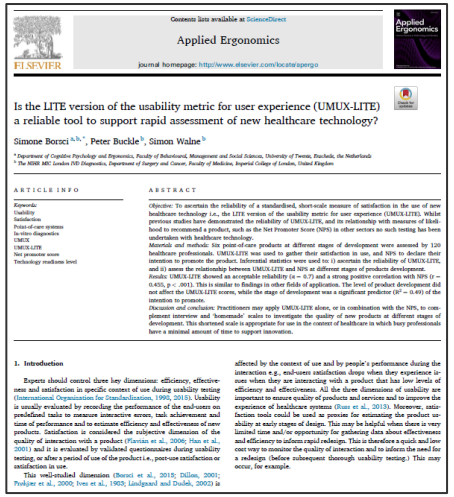
How to Use the UMUX-LITE



- Research Contexts
- Traditional usability testing
- Traditional experimental designs (e.g., between- and within-subjects)
- Retrospective evaluation (e.g., surveys)

- Standard Analyses
- Confidence interval estimation
- Comparing means
- Normative analysis using the curved grading scale

The Future of the UMUX-LITE



Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: http://www.elsevier.com/locate/apergo

Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology?

Simone Borsci [a,b,*], Peter Buckle [b], Simon Walne [b]

[a] Department of Cognitive Psychology and Ergonomics, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, the Netherlands
[b] The NIHR MIC London IVD Diagnostics, Department of Surgery and Cancer, Faculty of Medicine, Imperial College of London, United Kingdom

ARTICLE INFO

Keywords:
Usability
Satisfaction
Point-of-care systems
In-vitro diagnostics
UMUX
UMUX-LITE
Net promoter score
Technology readiness level

ABSTRACT

Objective: To ascertain the reliability of a standardised, short-scale measure of satisfaction in the use of new healthcare technology i.e., the LITE version of the usability metric for user experience (UMUX-LITE). Whilst previous studies have demonstrated the reliability of UMUX-LITE, and its relationship with measures of likelihood to recommend a product, such as the Net Promoter Score (NPS) in other sectors no such testing has been undertaken with healthcare technology.

Materials and methods: Six point-of-care products at different stages of development were assessed by 120 healthcare professionals. UMUX-LITE was used to gather their satisfaction in use, and NPS to declare their intention to promote the product. Inferential statistics were used to: i) ascertain the reliability of UMUX-LITE, and ii) assess the relationship between UMUX-LITE and NPS at different stages of products development.

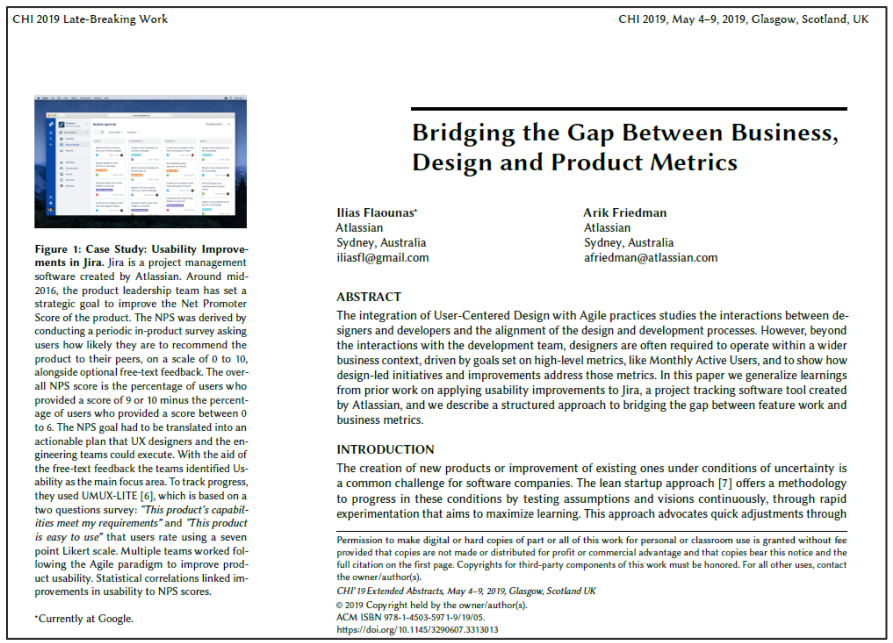
Results: UMUX-LITE showed an acceptable reliability ($\alpha = 0.7$) and a strong positive correlation with NPS ($r = 0.455$, $p < .001$). This is similar to findings in other fields of application. The level of product development did not affect the UMUX-LITE scores, while the stage of development was a significant predictor ($R^2 = 0.49$) of the intention to promote.

Discussion and conclusion: Practitioners may apply UMUX-LITE alone, or in combination with the NPS, to complement interview and 'homemade' scales to investigate the quality of new products at different stages of development. This shortened scale is appropriate for use in the context of healthcare in which busy professionals have a minimal amount of time to support innovation.

1. Introduction

Experts should control three key dimensions: efficiency, effectiveness and satisfaction in specific context of use during usability testing (International Organization for Standardization, 1998, 2018). Usability is usually evaluated by recording the performance of the end-users on predefined tasks to measure interactive scores, task achievement and time of performance and to estimate efficiency and effectiveness of new products. Satisfaction is considered the subjective dimension of the quality of interaction with a product (Hartson et al., 2006; Hao et al., 2009) and it is evaluated by validated questionnaires during usability testing, or after a period of use of the product i.e., post-use satisfaction or satisfaction in use.

This well-studied dimension (Borsci et al., 2015; Dillon, 2001; Frokjer et al., 2000; Ives et al., 1983; Lindgaard and Dudek, 2002) is affected by the context of use and by people's performance during the interaction e.g., end-users satisfaction drops when they experience issues when they are interacting with a product that has low levels of efficiency and effectiveness. All the three dimensions of usability are important to ensure quality of products and services and to improve the experience of healthcare systems (Esan et al., 2013). Moreover, satisfaction tools could be used as proxies for estimating the product usability at early stages of design. This may be helpful when there is very limited time and/or opportunity for gathering data about effectiveness and efficiency to inform rapid redesign. This is therefore a quick and low cost way to monitor the quality of interaction and to inform the need for a redesign (before subsequent thorough usability testing.) This may occur, for example.



CHI 2019 Late-Breaking Work — CHI 2019, May 4–9, 2019, Glasgow, Scotland, UK

Figure 1: Case Study: Usability Improvements in Jira. Jira is a project management software created by Atlassian. Around mid-2016, the product leadership team has set a strategic goal to improve the Net Promoter Score of the product. The NPS was derived by conducting a periodic in-product survey asking users how likely they are to recommend the product to their peers, on a scale of 0 to 10, alongside optional free-text feedback. The overall NPS score is the percentage of users who provided a score of 9 or 10 minus the percentage of users who provided a score between 0 to 6. The NPS goal had to be translated into an actionable plan that UX designers and the engineering teams could execute. With the aid of the free-text feedback the teams identified Usability as the main focus area. To track progress, they used UMUX-LITE [6], which is based on a two questions survey: *"This product's capabilities meet my requirements"* and *"This product is easy to use"* that users rate using a seven point Likert scale. Multiple teams worked following the Agile paradigm to improve product usability. Statistical correlations linked improvements in usability to NPS scores.

*Currently at Google.

Bridging the Gap Between Business, Design and Product Metrics

Ilias Flaounas*
Atlassian
Sydney, Australia
iliasfl@gmail.com

Arik Friedman
Atlassian
Sydney, Australia
afriedman@atlassian.com

ABSTRACT

The integration of User-Centered Design with Agile practices studies the interactions between designers and developers and the alignment of the design and development processes. However, beyond the interactions with the development team, designers are often required to operate within a wider business context, driven by goals set on high-level metrics, like Monthly Active Users, and to show how design-led initiatives and improvements address those metrics. In this paper we generalize learnings from prior work on applying usability improvements to Jira, a project tracking software tool created by Atlassian, and we describe a structured approach to bridging the gap between feature work and business metrics.

INTRODUCTION

The creation of new products or improvement of existing ones under conditions of uncertainty is a common challenge for software companies. The lean startup approach [7] offers a methodology to progress in these conditions by testing assumptions and visions continuously, through rapid experimentation that aims to maximize learning. This approach advocates quick adjustments through

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3313013

- UMUX-LITE has acceptable psychometric properties (reliability, validity, sensitivity) plus it is parsimonious (just 2 items)
- Open-source norms enable interpretation of SUS means, making the SUS the gold standard for assessing correspondence among perceived usability metrics
- Research to date indicates close correspondence between UMUX-LITE and SUS, allowing UMUX-LITE to piggy-back on open-source SUS norms (e.g., grades)
- New research also shows expected relationship between UMUX-LITE items and TAM components
- UMUX-LITE more contextually appropriate than LTR/NPS when users unlikely to engage in recommendation behavior
- UMUX-LITE already adopted for some use by some major corporations, and its use is likely to increase over the coming years
- Currently only available in English, Italian, and Slovene

The Usability Construct – Apparently Not a Dead End

The Usability Construct: A Dead End?

Noam Tractinsky
Ben-Gurion University of the Negev, Israel

"Usability" is a construct conceived by the human–computer interaction (HCI) community to denote a desired quality of interactive systems and products. Despite its prominence and intensive use in HCI research, the usefulness of the usability construct to HCI theories and to our understanding of HCI has been meager. In this article I propose and discuss two reasons for this state of affairs. The first is that usability is an umbrella construct. Umbrella constructs are prevalent in scientific fields that are broad, diverse, and lack a unifying research paradigm. Accordingly, umbrella constructs, such as usability, tend to be vague and loose, characteristics that challenge our ability to accumulate and communicate knowledge and to capture real-world phenomena. The second reason involves the nature of the relations between the usability construct and its measures, a topic rarely discussed in HCI research. There appears to be a mismatch between how the HCI community has (implicitly) conceptualized these relations and how it has empirically examined them. The relations have been conceptualized according to a formative measurement model but have mostly been tested according to a reflective measurement model. The trouble is that representing the usability construct by the reflective model appears inappropriate, and representing it by the formative model involves considerable difficulties. Possible ways of addressing these issues are discussed, each with its advantages and drawbacks. I conclude that for scientific research on this subject to progress, the usability construct ought to be unbundled and replaced by well-defined constructs. The issues discussed in this article are relevant to other HCI umbrella concepts and constructs such as user experience.

Perceived Usability and the Modified Technology Acceptance Model

Urška Lah[a,b], James R. Lewis[c], and Boštjan Šumak[b]

[a]EMO - Orodjarna Proizvodna družba, d. o. o, Celje, Slovenia; [b]Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia; [c]MeasuringU, Denver, Colorado, USA

ABSTRACT
In response to recent criticism of the usefulness of the construct of usability, we investigated the relationships between measures of perceived usability and the components of a modified version of the Technology Acceptance Model (mTAM) – Perceived Usefulness (PU) and Perceived Ease-of-Use (PEU). In three surveys, respondents used SUS, UMUX-LITE and mTAM to rate their actual (as opposed to expected) experience with three software products. As expected, the correlations between PEU and other measures of perceived usability tended to be significantly stronger than those with PU. Additional findings support the use of the UMUX-LITE as a compact measure of perceived usability that has a strong relationship to the mTAM and strong correspondence with concurrently collected SUS scores. The main theoretical result of this research were regression results providing evidence that the PEU component of the mTAM appears to be another measure of the construct of perceived usability, connecting the TAM to the construct of perceived usability through the mTAM and providing evidence against the claim that the construct of usability is a theoretical dead end.

1. Introduction

1.1. Perceived usability and technology acceptance

In addition to the objective components of efficiency and effectiveness, perceived usability is an important component of the classical conception of usability (Brooke, 2013; ISO, 1998; Lewis, Utesch, & Maher, 2013, 2015; Sauro & Lewis, 2009, 2016), which is in turn a fundamental component of user experience (UX; Diefenbach, Kolb, & Hassenzahl, 2014). The first standardized usability questionnaires intended for application in usability testing appeared in the late 1980s (Brooke, 1996; Chin, Diehl, & Norman, 1988; Kirakowski & Dillon, 1988; Lewis, 1990).

Around the same time that usability researchers were producing the first standardized questionnaires to assess perceived usability, market researchers who studied the adoption of information systems were addressing similar issues. Of these, one of the most influential has been the Technology Acceptance Model (TAM, Davis, 1989). According to TAM, the primary factors that affect a user's intention to use a technology are its perceived usefulness (PU) and perceived ease of use (PEU). This model addressed early criticism of focusing only on usability without consideration of whether a product or system was useful (Pearson & Bailey, 1980).

A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007). In the TAM, PU is the extent to which a person believes a technology will enhance job performance, and PEU is the extent to which a person believes that using the technology will be effortless. The more someone holds these beliefs before use, the greater their intention to use, and the more likely they are to try the technology. Figure 1 illustrates this model, and shows the wording of the items that Davis (1989) used to measure its constructs.

Recently, Tractinsky (2018), in a paper entitled "The Usability Construct: A Dead End?", argued against the usefulness of the construct of usability as a part of theory construction in human-computer interaction, in part due to "the inadequate modeling of the relations between the construct and its measures" (p. 133). In the same paper, he later cited the TAM as a good example of the use of constructs in scientific and practical model building, writing:

> Constructs contribute to a theory if they add to our understanding of the phenomenon under study. For example, the general domain that serves as the background for the emergence of the usability construct, namely, the use (often termed adoption or acceptance) of information technology, is a point of contact with various other theories. A notable such theory is the technology acceptance model (TAM; Davis, Bagozzi, & Warshaw, 1989). In TAM, the constructs "perceived ease-of-use" (a close relative of the intuitive meaning of "usability") and "perceived usefulness" are instrumental in explaining variations in the construct "behavioral intention" (people's intention to use a certain information system or product). Unfortunately, as previously mentioned, it is hard to name any influential theory in which the construct of "usability" plays a similarly useful role. (Tractinsky, 2018, p. 141)

The general reaction to the Tractinsky paper was that it offered valuable arguments regarding the difficulty of measuring usability and user experience, but those arguments were not universally accepted as the final word on the topic, especially with regard to the usefulness of usability as a construct

References

Berkman, M. I, & Karahoca, D. (2016). Re-assessing the Usability Metric for User Experience (UMUX) scale. *Journal of Usability Studies*, *11(3)*, 89-109.

Borsci, S., Buckle, P., & Wane, S. (2020). Is the LITE version of the usability metric for user experience (UMUX-LITE) a reliable tool to support rapid assessment of new healthcare technology? *Applied Ergonomics*, *80*, Article 103007.

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London, UK: Taylor & Francis.

Davis, D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319-339.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*, 323–327.

Flaounas, I., & Friedman, A. (2019). Bridging the gap between business, design, and product metrics. *Proceedings of CHI 2019* (LBW0274). Glasgow, Scotland: ACM.

Kortum, P., and Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale. *International Journal of Human-Computer Interaction*, *29*, 67-76.

Lah, U., Lewis, J. R., and Šumak, B. (2020). Perceived usability and the modified Technology Acceptance Model. *International Journal of Human-Computer Interaction*, DOI: 10.1080/10447318.2020.1727262.

Lewis, J. R. (2018a). Is the report of the death of the construct of usability an exaggeration? *Journal of Usability Studies*, *14(1)*, 1-7.

Lewis, J. R. (2018b). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*, *34(12)*, 1148-1156.

Lewis, J. R. (2019a). Comparison of four TAM item formats: Effect of response option labels and order. *Journal of Usability Studies*, *14(4)*, 224-236.

Lewis, J. R. (2019b). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human-Computer Interaction*, *35(15)*, 1404-1419.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE—When there's no time for the SUS. In *Proceedings of CHI 2013* (pp. 2099-2102). Paris, France: Association for Computing Machinery.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, *31(8)*, 496-505.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*, 46-54.

Tractinsky, N. (2018). The usability construct: A dead end? *Human-Computer Interaction*, *33(2)*, 131-177.

MeasuringU

Usability, Customer Experience & Statistics

MeasuringU.com
@MeasuringU

MUIQ

Remote UX Testing Platform
(Desktop & Mobile)

UX Research

SUPR-Q

Measurement
& Statistical Analysis

Eye Tracking &
Lab Based Testing

MeasuringU is a research firm based in Denver, Colorado focusing on quantifying the user experience.

DenverUX BOOTCAMP

MeasuringU
Custom
Training