Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems

Hyewon Suh*, Nina Shahriaree†, Eric B. Hekler‡, Julie A. Kientz*

*Human Centered Design & Engineering, DUB Group
University of Washington
{hyewon25, jkientz}@uw.edu

†Human Computer Interaction + Design
University of Washington
nina8@uw.edu

‡School of Nutrition and Health Promotion
Arizona State University
ehekler@asu.edu

ABSTRACT

Computing systems that place a high level of burden on their users can have a negative affect on initial adoption, retention, and overall user experience. Through an iterative process, we have developed a model for user burden that consists of six constructs: 1) difficulty of use, 2) physical, 3) time and social, 4) mental and emotional, 5) privacy, and 6) financial. If researchers and practitioners can have an understanding of the overall level of burden systems may be having on the user, they can have a better sense of whether and where to target future design efforts that can reduce those burdens. To help assist with understanding and measuring user burden, we have also developed and validated a measure of user burden in computing systems called the User Burden Scale (UBS), which is a 20-item scale with 6 individual sub-scales representing each construct. This paper presents the process we followed to develop and validate this scale for use in evaluating user burden in computing systems. Results indicate that the User Burden Scale has good overall inter-item reliability, convergent validity with similar scales, and concurrent validity when compared to systems abandoned vs. those still in use.

Author Keywords

User burden; user experience; usability; validated measures; measuring usability; evaluation; technology abandonment;

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The growth of computing systems over the last few decades has been tremendous. Wearable sensors, smartphone-based applications, websites, and more are both readily available and affordable to end-users. Designers and developers have created new systems targeting applications in domains as varied as health, finance, social networking, productivity, and entertainment. The Human-Computer Interaction community has been instrumental in developing ways to evaluate these types of systems and assess them along a number of dimensions, including usability, user experience, and usefulness. Despite these efforts, many people still fail to adopt systems that could benefit them greatly or abandon systems after very little use, even if they are seeing the benefit. They may also use a system out of necessity, but it could still have a negative impact on their lives. We suspect one possible cause of these issues is that the burden these systems place on the user is too high.

We define *user burden* as the negative impact that computing systems might place on the user. While burden includes issues with usability and user experience, it can also include other aspects that may be more subjective in nature and more dependent on individual differences. Through a review of the literature, analysis of existing computing systems, and a principal component analysis, we have developed a model of user burden consisting of 6 unique constructs of user burden: 1) difficulty of use, 2) physical, 3) time and social, 4) mental and emotional, 5) privacy, and 6) financial. Each of these types of burdens can make it difficult for people to initially adopt or continue to use a system and may have a negative impact on the overall user experience. In addition, user burden may be present even when the user has continued to adopt and use a system, but it may be decreasing their overall user experience and engagement.

The concept of user burden as we define it is unique to the field of user experience design in that it focuses primarily on all the aspects of a system that may negatively impact a user's ability to use and tolerate it. It is important for designers to be able to understand and assess the amount of user burden in each of these areas in their designs so that they can make efforts to reduce as much of it as possible. Currently, assessing user burden is not something that is part of standard usability practice, although current user experience methods may touch on these aspects in an indirect way (e.g., by assessing error rates in a usability test or task load). We propose to make understanding and evalua-

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05 $15.00
DOI: http://dx.doi.org/10.1145/2858036.2858448

tion of user burden a much more systematic and lightweight than existing methods.

Although user burden may be measured by more objective measures, whether a system places a burden for a particular person needs to be assessed on an individual level. For example, if the cost of a smartphone is fixed at $300 with a $40 per month subscription fee, this may be an easy expense for someone who is wealthy, but very burdensome for someone from a resource-constrained background. Likewise, a wearable technology shirt may be too heavy and bulky for someone with a smaller frame, but quite comfortable and unnoticeable for someone with a larger build. Because of these individual differences, gathering feedback from a large number of diverse users is important. Many existing methods of assessing user burden may not scale well, and thus we have sought to develop a validated measure of user burden through the form of a questionnaire, called the User Burden Scale (UBS).

In this paper, we describe the development and validation of the UBS. Results indicate that the test has good overall inter-item reliability, convergent validity with similar scales, and concurrent validity when compared to systems abandoned vs. those still in use. We believe that researchers and practitioners can use this scale to assess the overall burden that a computing system places on current users of the system, and we discuss future work that will help expand this approach even further.

RELATED WORK

Here we describe related work in the areas of evaluating computing systems and concepts relating to user burden.

Evaluation of Computing Systems

In the field of HCI, there have been many projects aiming to evaluate computing systems from the user's perspective. Usability is one of the most well-known and well-defined concepts in the human-computer interaction (HCI) research. It was originally defined as the degree of efficiency and effectiveness of the system [7, 24, 44, 45] and has been emphasized to be an important factor in making a successful system [11, 21, 42, 49]. In many early usability studies, the evaluators collected specific measures, such as ease of learning, efficiency of use, memorability, speed and accuracy in performing a task, all of which are believed to reflect aspects of the usability of the system. Based on the collected measures, the evaluators make a conclusion that a certain system has good usability and is ready to be adopted by users or that one version of a system is better than others in terms of one measure or another. However, improving the system in terms of the objective measures does not always mean that the users are satisfied thoroughly with the system.

To overcome the limitation of objective measures, researchers in the field tried to extend the usability concept to emphasize the subjective aspects, including emotional and behavioral factors. Because user preference and satisfaction are not as easy to measure directly, researchers have developed a wide variety of evaluation instruments and user-completed scales. Some of these instruments include the Questionnaire for User Interaction Satisfaction (QUIS) [14], the Computer User Satisfaction Inventory [32], the NASA Task-Load Index (NASA-TLX) [27], the Software Usability Measurement Inventory (SUMI) [33], the System Usability Scale (SUS) [10], the Purdue Usability Testing Questionnaire (PUTQ) [36], and the IsoMetrics Usability Inventory [22]. To our knowledge, no other scales have attempted to characterize all aspects of user burden as we define it.

In addition, human behavior and decision-making theories such as the Technology Acceptance Model (TAM) and the Value-based Adoption Model (VAM) help us understand the behavioral aspects of technology adoption. TAM, a popular framework developed by Davis [17], suggests that people's adoption of a new technology is dependent on their intention to use it, which is in turn dependent on their perception of the 1) technology's usefulness and 2) ease of use [16, 17]. To account for affective value that is missing from TAM, VAM brings to light the important role of affective value in consumer adoption of new technology. According to VAM, consumers choose whether to adopt a technology or service by weighing the products' perceived cognitive and affective benefit such as usefulness, enjoyment against perceived sacrifices such as fees or technicality, including mental or physical effort. The VAM's perceived sacrifice is closely related to user burden, but it does not include concerns such as social, emotional, privacy, and financial burdens. It also relates specifically to adoption, whereas the User Burden Scale is intended to account for more than just adoption and look at how a computing system negatively impacts the user even after they have already adopted it.

Measurements for Different Types of Usability Issues

The study of user burden owes much to research into usability. In this section, we describe various instruments used to evaluate usability, which have been developed and had their reliability validated over time [37]. In light of other growing concerns over user experience beyond usability, we also included additional literature beyond usability instruments to provide a more holistic picture of user burden. Our work builds upon these previous works to develop a quick, low-cost scale for assessing many different aspects of user burden.

Generating scales involves defining the construct of interest and generating a candidate list of items from the domain of all possible items representing the construct. Cronbach [15] states that instruments should draw representative items from a universal pool in order to ensure content validity. With this respect, we reviewed literature related to evaluation of computing technologies and reviewing existing instruments for scales that could be modified for UBS.

Many of these instruments use the concept of cognitive load as the primary lens to evaluate usability [13, 28]. In questionnaires such as the SUS [10] or NASA-TLX [27], sub-

jects are asked to self-report on their experiences with the product in terms of how mental, physical, emotional, and temporal factors contribute to their experience in using a particular system. For instance, emotion is measured in terms of when a user experiences frustration with a task they are asked to perform [43].

Growing awareness has led to the development of many tools to measure and evaluate usability in light of accessibility issues, as detailed by [1]. Guidelines like the ISO standards [25] and surveys like the Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST) [18] provide designers and researchers methods for assessing user satisfaction from an accessibility perspective.

For the financial aspects of computing systems, quantitative models to estimate price sensitivity exist in the field of economics with relevance to HCI designers and software developers. Incorporating financial concerns, especially considering how it burdens a particular user, provides a more comprehensive assessment of user burden [26, 35]. In addition, the critical implications of privacy in relation to usability have given rise to various frameworks that can help researchers and designers make these issues more concrete, as demonstrated by [3, 29, 50]. There exists a need to consider privacy as a key aspect of user burden, which we incorporated into our scale.

The social lives of users are increasingly impacted by interactions with computing systems, and thus understanding social burdens is crucial. From the Internet to smartphones and social networking, these ongoing changes create new avenues for psychological researchers seeking to understand the influence these systems can have on the user's social relationships [8]. Including the social dimension in our survey of user burden gives designers and researchers in HCI valuable insights of growing importance.

In cases where user burden has been difficult to measure through subjective self-reporting, researchers and practitioners have used various techniques for measuring physiological and behavioral data in real-time, such as galvanic skin response, eye-tracking, etc. For the purposes of our scale, we specifically wanted to look at different burdens places on the user that they actually feel are burdens. Thus, we focus on what can be learned after the user has experienced the product and measures that can be made quickly and easily from the user's perspective, since it may differ from person to person.

Through a review of the literature, we identified several characteristics or dimensions of usability related issues, which led to the initial user burden scale (Table 1). Although previous studies and theoretical frameworks provide dimensions regarding the subjective preference about the different types of impacts that computing systems may place on the user, there has been not yet been a systematic method for measuring user burdens. Accepting one system or high user satisfaction is not directly translated to having no burden associated with using it. Thus, this study attempts to define a new way of evaluating systems from a user burden point of view and develop a validated instrument that measures it.

Sub-scale	Related Literature
Difficulty of use	[1], [10], [14], [18], [25], [33], [36], [40]
Physical	[2], [27], [31]
Social & Time	[6], [8], [14], [20], [27], [33], [41], [46]
Mental & Emotional	[10], [13], [14], [22], [27], [28], [33], [34], [36], [43]
Privacy	[3], [5], [29], [47], [50]
Financial	[26], [35]

Table 1. UBS-related literature based on the six constructs of user burden

INITIAL DEFINITION OF USER BURDEN

Based on our review of the literature, and a number of rounds of discussion among experts in user-centered design of computing systems, we determined an initial eight types of user burden, which included access, emotional, financial, mental, physical, privacy, social, and time-based burdens. We defined each of those initial categories as follows:

- **Access Burden**: The system does not fit with the abilities or cultural background of the user.
- **Emotional Burden**: The system makes the user feel bad or unnecessarily worry.
- **Financial Burden**: The system costs a significant amount of money for the user to initially purchase or to maintain use.
- **Mental Burden**: The system requires significant attention, concentration, or is distracting.
- **Physical Burden**: The system makes the user physically uncomfortable.
- **Privacy Burden**: The system risks revealing information about a user that he or she would prefer not to share.
- **Social Burden**: The system may disrupt the user's ability to create and sustain social relationships.
- **Time Burden**: The system requires frequent use or a significant amount of time to use.

Refining User Burden Via Interviews

To refine our understanding of initial definitions the user burden categories we developed and to refine them further, we conducted an hour-long, semi-structured, in-person interview study with 12 participants (6 male, 6 female). We recruited participants who were at least 18 years old and use computing technologies on a daily basis via university and group email lists of faculty and students, and by word of mouth. Participants provided a list of five technologies they frequently used, and the research team decided which two technologies about which to interview and survey to ensure a broad spectrum of applications. From 12 participants, we collected data about 24 different computing systems. Because the purpose of the interview was to explore participants' experience with computing technologies and the user

Initial Category	Sample Interview Prompt	Interview Participant Statement
Access	**[Steam]** Have you ever felt difficulty in seeing, hearing, or manipulating the system?	"Sometimes the graphics requirements are too high for my laptop or sometimes it will just mess up. There will be some kind of lag or something, and I can't see properly, and then I'll die."
Emotional	**[MS Word]** Has using the system ever made you feel bad, other than the frustration that you mentioned already?	"Yeah, well I mean, general software crashes. I've definitely lost information. That makes me feel bad. Sometimes the auto-save doesn't catch it all."
Financial	**[MapMyRun]** Was the amount that you spent on the system the amount that you expected to pay?	"No. I was very disappointed actually. Of all the apps that I paid for, and I love MapMyRun, but the features you pay for really aren't worth $6.00 a month."
Mental	**[Garmin]** What did you have to do to learn how to use the system?	"The Garmin was a little more challenging and there was a lot of trial and error and looking things up a little bit. I had a manual because some of the options and features weren't quite intuitive. But the Google Maps navigation is pretty straight forward."
Physical	**[FaceTime]** Have you ever felt physically uncomfortable while using the system?	"Yeah. It actually heats up very fast, so it becomes very hot. Also, as with many Apple products, it is rounded, and it's very thin, so it's very hard to set down."
Privacy	**[Netflix]** Were there any steps that you have taken to ensure your privacy?	"Not really with Netflix, since it's pretty easy to use. I would actually rate movies actively so that it would recommend something that I want to see. … I don't worry about Netflix using my personal information."
Social	**[MapMyRun]** How does using this system impact your relationship with others?	"I think... My dad, for example, really quite likes it because when it goes to my Facebook, he always likes it because to him it's probably like because I'm from England so he's in England. He sees when I go to a new place and go for a run. I think he quite likes that. He can see I'm in Seattle or Vancouver or whatever."
Time	**[Reddit]** Does the time that you spend on Reddit match your desire to use to Reddit?	"No, I'd love to use it a lot less… but I'd usually do it during times of boredom, which means I'm bored a lot."

Table 2. Sample interview prompts and sample coded phrases for each initial category of user burden.

burden associated with it, interview questions were semi-structured and open-ended to facilitate interviewees in bringing up new ideas during the interview. Interview methods were patterned after Weiss' [48] techniques. As a token of appreciation, participants received $25 in gift cards for their participation. All interviews were audio recorded and transcribed.

We analyzed interview transcripts using thematic analysis [9], creating thematic connections of interview data. Several iterations of this process produced refined and distinct themes. In addition to this, all interview statements were examined and for the statements implying or identifying user burdens, two authors assigned them to one of corresponding eight user burden scales (Table 2). The results of these interviews indicated that the initial categories we defined were consistent with expressions of user burden from end users, and that we could move forward to a more formal validation of the categories into constructs and the definition of a scale.

CONSTRUCT VALIDITY & QUESTIONNAIRE DESIGN

We aimed to design a questionnaire that could be used by designers and developers in the field to evaluate existing technologies that have been used in real world situations. Following the requirements and guidelines of scale development procedures used by Yarosh et al. [51], we came up with several requirements for UBS:

- Measure different categories of perceived user burden in using technologies.
- Refer to a specific system, but be generic to be applicable to a wide range of technologies.
- Be quick to administer.
- Demonstrate reliability and validity on multiple metrics.
- Be sensitive enough to detect differences between technologies.

Identification & Development of Scales

After defining the initial eight categories of burden, the research team brainstormed and developed preliminary scales for each category of user burden. This process yielded 15 items in each of eight categories. Review and discussion by three experts led to eliminating and re-writing a significant portion of the questions, resulting a draft of the survey with 64 items (8 items on each scale). Through 5 rounds of pilot studies with over 922 participants Amazon Mechanical Turk (MTurk), we continued eliminating items to make it brief and concise and to remove questions that were confusing, resulting 8 scales with 26 items. At this point, the items were measured using a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5), with a mix of positively and negatively framed questions. Throughout pilot studies, tasks given to participants were always completing the most recent user burden scale questionnaire at the moment. In the first round (15 items in each of eight categories), we asked participants to complete

the survey on a single, specific system: Facebook. However, because we wanted our survey to be generalizable to any interactive systems, each of the subsequent rounds of testing asked participants to choose an interactive system that they use frequently or one that they have previously used that they have now abandoned. Based on these preliminary results and discussion amongst the research team, we decided to convert all of the questions be negatively framed (since user burden was considered a negative experience) and use two different 5-point scales to add additional nuance beyond a simple Likert scale. The two final response types are as follows:

Response Type 1 (Frequency/Occurrence): 0 = Never; 1 = A little bit of the time; 2 = Sometimes; 3 = Very often; and 4 = All of the time.

Response Type 2 (Degree/Magnitude): 0 = Not at all; 1 = A little bit; 2 = Somewhat; 3 = Very much; 4 = Extremely

Principal Component Analysis

To explore whether our initial 8 categories of user burden held up as constructs and to reduce the number of questions with statistical analysis, we deployed the 26-item survey (all negatively phrased with two response types) via Amazon's Mechanical Turk. 300 participants completed this version of the UBS on an interactive system that they frequently use or one that they have previously used but they have now abandoned. A total of 274 responses remained after filtering by location and survey completion time less than 60 seconds. For remaining 274 responses, the principal component analysis was used to extract the components and this was followed by a varimax (orthogonal) rotation.

	Initial user burden group*	Rotated Factor Pattern					
		Difficulty of use	Physical	Time & Social	Mental & Emotional	Privacy	Financial
I need assistance from another person to use [X].	A	.78					
[X] demands too much mental effort.	M	.77					
It takes too long for me to do what I want to do with [X].	T	.69					
[X] is hard to learn.	M	.69					
I get frustrated when using [X].	M	.68					
Information, such as visual cues or sounds, from [X] is hard to understand.	A	.60					
The value of [X] is not worth the cost to me.	F	.47					
Using [X] too much creates physical discomfort.	Ph		.74				
[X] has made me feel physical pain.	Ph		.67				
[X] is not appropriate for my cultural background.	A		.62				
I don't want others to know that I use [X].	S		.61				
Use of [X] is too physically demanding.	Ph		.53				
I spend too much time using [X].	T			.85			
I use [X] more often than I should.	T			.85			
[X] distracts me from social situations.	S			.73			
Using [X] has a negative effect on my social life.	S			.61			
[X] requires me to remember too much information.	M				.66		
[X] presents too much information at once.	M				.65		
Using [X] makes me feel like a bad person.	E				.64		
I feel guilty when I use [X].	E				.64		
[X] forces me to make changes to how I normally use digital technologies.	A				.56		
I am worried about what information gets shared by [X].	Pr					.88	
[X]'s policies about privacy are not trustworthy.	Pr					.83	
[X] requires me to do a lot to maintain my privacy within it.	Pr					.78	
[X] is too expensive.	F						.88
The upfront cost to using [X] is too high.	F						.81

Table 3. Factor loadings and communalities based on a principal components analysis (PCA) with varimax rotation for 26 items (N = 274). Items included in the final questions set are highlighted
(**A**: Access, **E**: Emotional, **F**: Financial, **M**: Mental, **Ph**: Physical, **Pr**: Privacy, **S**: Social, **T**: Time)*

After principal component analysis of factor extraction, only the first six components displayed eigenvalues greater than 1, not eight. The results of a scree test also suggested that the first six components were meaningful. Therefore, the first six components were retained for rotation. Combined, components 1 to 6 accounted for 67% of the total variance.

Questionnaire items and corresponding factor loadings are presented in Table 3. With this result, we carefully selected items to be included in the final question set. We chose the 4 items of highest loading from each factor. The *Privacy* and *Financial* constructs had less than 4 items, and we chose all of them. One exception was made for the *Physical* construct, where we chose to include the two items with highest factor loading and then included the 5th item which had slightly lower factor loading but seemed more relevant to the overall construct. In addition to eliminating items, the PCA also resulted in changes in the user burden constructs. Through the PCA, *Mental and Emotional* burden categories were combined into one construct, as were *Time and Social* burdens. We also developed a new construct, Difficulty of Use burden, and eliminated the *Accessible* burden construct as it was covered in the other constructs. Overall, these analyses indicated that six distinct factors were underlying in the model of user burden and a total of six items were eliminated because they did not contribute to a simple factor structure.

FINAL USER BURDEN CONSTRUCTS & QUESTION SET

Based on our initial exploration and the construct validity using principal component analysis, we refined our initial proposed categories into a model consisting of six constructs, as well as our definitions for each construct.

User Burden Construct Definitions

Below we provide a definition of each of the final constructs and two or three examples of systems that we considered to place a high burden on the user in this area.

Difficulty of Use Burden

The system does not fit with the abilities of the user and is difficult to use. *Example systems:* i) A photo editing software package with a steep learning curve; ii) A website that is not compatible with a blind user's preferred screen reader.

Physical Burden

The system makes the user physically uncomfortable. *Example systems:* i) A body-worn sleep sensor that gives the user a rash if worn too long; ii) A text-entry system that causes repetitive stress injuries in the wrist due to over use.

Time & Social Burden

The system may require a significant amount of time to use or disrupt the user's ability to create and sustain social relationships. *Example systems:* i) A mobile food diary that requires several minutes to enter each item of food consumed throughout the day; ii) A mobile application that makes noise and annoys others who are working nearby the user; iii) A system that automatically sends emails to everyone on a user's contact list.

Mental & Emotional Burden

The system requires significant attention, concentration, or is distracting, or makes the user feel bad or unnecessarily worry. *Example systems:* i) An exergame that shames an overweight person who plays it if they are too heavy; ii) A phone-based news application that constantly sends the user disruptive reminders; iii) a system that overwhelms the user with a confusing visual display.

Privacy Burden

The system risks revealing information about a user that he or she would prefer not to share. *Example systems:* i) A weight scale that by default automatically posts a user's age and weight to their social media accounts; ii) A social networking system that reveals personal information to others without the user's consent.

Financial Burden

The system costs a significant amount of money for the user to initially purchase or to maintain use. *Example systems:* i) A bicycle GPS system that has a high initial cost and is expensive to replace if damaged or stolen; ii) A video streaming service that requires a costly monthly fee.

Final Scale and Question Set

Throughout the questionnaire design, UBS aimed to cover a wide range of user burden associated with using technologies. After multiple iterations of pilot studies and analysis, the final UBS consisted of 6 subscales representing the 6 constructs defined above with 20 total items. To be able to use UBS for both systems that people currently use and systems they used once but have abandoned, UBS has both a past tense as well as present tense version. The full text of the scales and questions are presented in Table 4.

VALIDATION PROCESS

To provide evidence regarding reliability and validity of the refined scales, we deployed the survey on Mechanical Turk and conducted three different analyses: 1) inter-item reliability; 2) convergent validity with existing instruments; and 3) the sensitivity of the instrument.

Methods

To test the User Burden Scale with a large sample, we deployed an online version of UBS again on Mechanical Turk. We asked participants to complete the new 20-item version of the UBS, the NASA Task-Load Index (NASA TLX), and the System Usability Scale (SUS) two times each: one for a computing system they currently use and one for a computing system they once used but have now abandoned (hence also the test of present and past tense). The order of answering on using vs. abandoned technologies was counter-balanced while the order of UBS, TLX and SUS was always in the order written. To be able to take our survey, we required that participants be at least 18 years old, reside in the United States, and be frequent users of any

Scale	#	Final User Burden Scale Question Set and Subscales		Response type
		Currently using (present tense)	Abandoned (past tense)	
Difficulty of Use	1	I need assistance from another person to use [X].	I needed assistance from another person to use [X].	1
	2	[X] demands too much mental effort.	[X] demanded too much mental effort.	1
	3	It takes too long for me to do what I want to do with [X].	It took too long for me to do what I wanted to do with [X].	1
	4	[X] is hard to learn.	[X] was hard to learn.	2
Physical	5	Using [X] too much creates physical discomfort.	Using [X] too much created physical discomfort.	2
	6	[X] has made me feel physical pain.	[X] had made me feel physical pain.	1
	7	Use of [X] is too physically demanding.	Use of [X] was too physically demanding.	1
Time and Social	8	I spend too much time using [X].	I spent too much time using [X].	2
	9	I use [X] more often than I should.	I used [X] more often than I should have.	1
	10	[X] distracts me from social situations.	[X] distracted me from social situations.	1
	11	Using [X] has a negative effect on my social life.	Using [X] had a negative effect on my social life.	1
Mental and Emotional	12	[X] requires me to remember too much information.	[X] required me to remember too much information.	1
	13	[X] presents too much information at once.	[X] presented too much information at once.	1
	14	Using [X] makes me feel like a bad person.	Using [X] made me feel like a bad person.	1
	15	I feel guilty when I use [X].	I felt guilty when I used [X].	1
Privacy	16	I am worried about what information gets shared by [X].	I was worried about what information got shared by [X].	2
	17	[X]'s policies about privacy are not trustworthy.	[X]'s policies about privacy were not trustworthy.	2
	18	[X] requires me to do a lot to maintain my privacy within it.	[X] required me to do a lot to maintain my privacy within it.	1
Financial	19	[X] is too expensive.	[X] was too expensive.	2
	20	The upfront cost to using [X] is too high.	The upfront cost to using [X] was too high.	2

Reporting Value	Response Type 1	Response Type 2
0	Never	Not at all
1	A little bit of the time	A little bit
2	Sometimes	Somewhat
3	Very often	Very much
4	All of the time	Extremely

Table 4. User Burden Scale items and categories. [X] is the name of the system being investigated. The order of item was randomized and two response types were used as appropriate. Of the 20 items, 13 used Response Type 1 and 7 used Response Type 2.

Using System	# of responses	Abandoned System	# of responses
Facebook	78	**Skype**	65
Netflix	65	**Facebook**	55
YouTube	42	**Wii**	31
Gmail	38	**Dropbox**	30
PayPal	18	**Netflix**	23
iPad	17	**PayPal**	22
Kindle	16	**Kindle**	15
Skype	12	**Fitbit**	13
Fitbit	10	**Gmail**	11
iPhone	8	**MS Word**	9
Other	71 (45 distinct)	**Other**	101 (50 distinct)
TOTAL	375 responses with 55 distinct systems	**TOTAL**	375 responses with 60 distinct systems

Table 5. List of systems participants rated using UBS

type of computing system. Qualified participants took the survey on a computing system of their choice. Some of popular systems participants named are in Table 5. There were a number of systems currently used by some participants but had also been abandoned by others. We expected 7-10 minutes to complete two sets of three surveys (UBS, NASA TLX and SUS) and compensated participants $0.80 through Mechanical Turk for their task. A total of 396 participants completed all the surveys. We filtered out some responses (i.e., 9 participants outside the United States and 12 who took too short of time to complete (less than 3 minutes for the entire survey)). The remaining 375 responses were analyzed to examine the reliability and validity of the questionnaire.

Subscale Burden Type	# of items	Cronbach's alpha
Difficulty of Use	4	0.817
Physical	3	0.814
Time and Social	4	0.862
Mental and Emotional	4	0.728
Privacy	3	0.890
Financial	2	0.891
OVERALL	20	0.883

Table 6. Subscale Reliability Using Cronbach's Alpha Coefficient for UBS (N=750).

Results

We report the results of the User Burden Scale validation process in terms of inter-item reliability, convergent validity with existing instruments, and concurrent validity sensitivity in detecting differences between technologies.

Inter-item Reliability

With 750 survey responses from 375 participants (two technologies per participant), we calculated the inter-item reliability metrics of the UBS (Table 6). All of subscales had good internal consistency (> 0.8), with the exception of the *Mental and Emotional* burden subscale, which was still in the acceptable (> 0.7) range [23]. This is promising because it is generally known that Cronbach's alpha is dependent on the number of items: fewer items will likely lead to a low alpha, whereas many items will increase reliability. So, it is common for scales with only a few items per construct (3 to 6) to yield a lower alpha [19, 38]. Our questionnaire achieved good alpha values across even a small number of items within each subscale. The entire UBS also attained a good Cronbach's alpha of 0.883, indicating the survey has good internal consistency as a whole.

Convergent Validity with Existing Validated Instruments

In addition to the UBS, we asked participants to complete the NASA Task Load Index (NASA-TLX) scale and the System Usability Scale (SUS) to see the relationship between the perceived workload of the system and user burden and that of the system's usability and user burden. We hypothesized that those participants who reported having greater user burden with certain technologies would report higher NASA-TLX scores and lower SUS scores. To test these hypotheses, we ran a Spearman's rank-order correlation. In accordance with our assumption, there were significant positive correlations between the UBS and the NASA-TLX ($r(748) = 0.506$, $p < 0.001$). Correlation between the UBS and SUS was significantly negative ($r(748) = -0.366$, $p < 0.001$). This result indicates that although UBS evaluates technologies from different dimensions than the NASA-TLX and SUS, it has good convergence with it.

Concurrent Validity for Technologies Used or Abandoned

We also wanted to determine if the UBS is sensitive enough to detect differences between technologies with lower user burden (those likely to be still in use) than those with higher user burden (those likely to be abandoned). To compare survey responses on two technologies (using vs. abandoned) by 375 participants, we ran an independent samples *t*-test, where the test variable was sum of each scale and grouping was *using* vs. *abandoned*. For all burden scales and the overall questionnaire, the UBS score difference between the two technologies was statistically significant ($p < 0.001$) (Table 7). This suggests that the UBS is sensitive to detect differences between used and abandoned technologies. One thing to note is that for all subscales except *Time and Social*, the user burden was higher for abandoned system than for systems still in use but for *Time and Social*, it was the opposite. It may be that *Time and Social* burdens are not as good of predictors of abandonment than the other constructs. One reason for this may be that technologies often become abandoned when they are no longer used, which would mean they were not imposing a time burden or interfering socially.

Burden Subscale	Using	Abandoned	Conditions
Difficulty of Use	M = 1.22 (SD = 1.832)	M = 3.83 (SD = 3.759)	t(748) = -12.102 (p < 0.001)
Physical	M = 0.52 (SD = 1.375)	M = 0.99 (SD = 2.038)	t(748) = -3.696 (p < 0.001)
Time and Social	M = 4.57 (SD = 3.791)	M = 3.14 (SD = 4.009)	t(748) = 4.997 (p < 0.001)
Mental and Emotional	M = 1.39 (SD = 1.940)	M = 2.37 (SD = 3.112)	t(748) = -5.183 (p < 0.001)
Privacy	M = 2.59 (SD = 2.780)	M = 3.66 (SD = 3.926)	t(748) = -4.326 (p < 0.001)
Financial	M = 0.93 (SD = 1.684)	M = 1.38 (SD = 2.132)	t(748) = -3.231 (p = 0.001)
OVERALL	M = 11.21 (SD = 8.979)	M = 15.38 (SD = 13.132)	t(748) = -5.074 (p < 0.001)

Table 7. Comparing different computing systems based on UBS sum using Independent Samples *t*-Test (N=375 each on using and abandoned technologies).

DISCUSSION

In this section, we provide guidelines for administrating the User Burden Scale, including how to score and interpret the results and general usage guidelines. We also discuss its limitation and future directions.

UBS Scoring and Analysis and Guidelines for Use

The User Burden Scale users two 5-point scales (ranging from 0 to 4). This allows the scale to have a higher score resulting in a higher level of user burden. Given that there are 20 total questions, the maximum score is 80 and the minimum score is 0. So that the score is comparable across systems and across users, we recommend that participants be required to answer every question. If a participant feels that they cannot answer a particular question or that it is not applicable, they should be instructed to choose the 0 value for that item since it is likely that item is not a burden. Because each subscale had a good alpha value, survey administrators may also choose to only administer the subscales relevant to the system of interest. If the administrator

Burden Subscale	# of items	User Burden Score Range				
		A top 15%	B next 30%	C next 40%	D next 10%	F bottom 5%
Difficulty of Use	4	0	0.25	0.5 - 1.5	1.75 - 2.25	2.5 - 4
Physical	3	0	0	0.33 - 0.67	1-1.33	1.67 - 4
Time and Social	4	0	0.25 - 0.5	0.75 - 2.25	2.5 - 2.75	3 - 4
Mental & Emotional	4	0	0.25	0.5 - 1	1.25 - 1.75	2 - 4
Privacy	3	0	0.33	0.67 - 2.33	2.67 - 3.33	3.67 - 4
Financial	2	0	0	0.5 - 1.5	2 - 3	3.5 - 4
OVERALL	20	0 - 0.15	0.2 - 0.45	0.5 - 1.2	1.25 - 1.7	1.75 - 4

Table 8. Score guidelines for UBS (overall and each subscale)

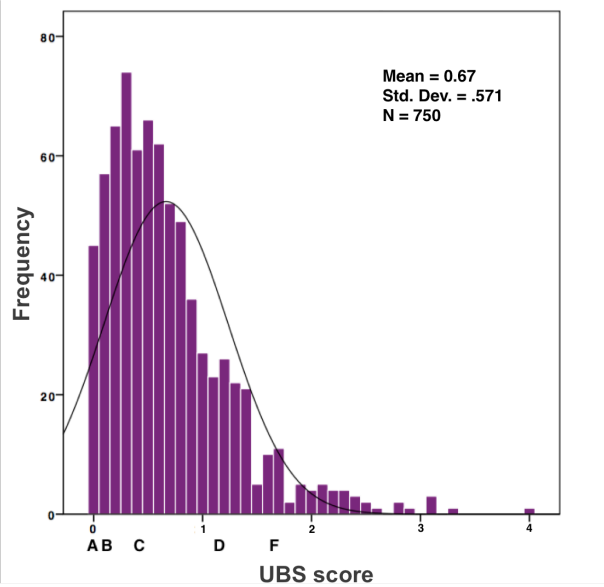

Figure 1. UBS score distribution (N=750)

chooses, he or she can calculate the score for each subscale by calculating the mean for items within each burden category to determine which constructs of user burden seem to be contributing the most burden.

Figure 1 shows UBS score distribution from initial 750 data points. Based on this, we present a score guidelines with letter grades in Table 8 (A: top 15%, B: next 30%, C: next 40%, D: next 10%, and F: bottom 5%.) We provide cutoff scores for each subscale as well as overall UBS.

Our validity tests described above were administered using Survey Gizmo (http://surveygizmo.com), where the first question asked participants the name of the system they were evaluating. We then piped the system name into all of the questions indicated by [X] in Table 4 and had the survey tool randomize the order of the questions for all participants. We believe that the scale could also be administered on paper if the system name was completed ahead of time (or [X] being replaced more generally by "the system") and each sheet printed had the questions in random order, but we have not yet validated the scale for use on paper. The 20-item version of the scale took approximately 2 to 3 minutes to complete. We also encourage administrators to consider whether participants are currently using the system of interest and use the present tense version, and if they are evaluating a technology that has been abandoned to use the past tense version (see Table 4). We should also note that the scale assumes that the user has been using the system for some unspecified amount of time, and thus is intended more for evaluation of fully developed systems or later stage prototypes that are fully deployable, rather than early mockups or low-fidelity, non-functional prototypes.

Another consideration for use is interpreting *why* the system is causing the user to be burdened. In our interviews, there were several times that people had a hard time distinguishing between the burden of the system based on its design or its content. For example, if someone was evaluating Gmail, and they were asked to respond to questions about times when it has made them feel badly, they mentioned the content of specific emails or the quantity of emails as the source of bad feelings, in addition to issues with the interface itself. A few indicated that they were not sure that they would fault Gmail for making them feel bad because of this. The UBS does not currently distinguish between interface design and content, and so the user burden scores should be interpreted as a combination of the two. If further discernment is required, we recommend follow up studies with survey respondents to probe at what about the system is causing the user to feel burdened. Administrators could consider adding open-ended questions after administering the survey to allow participants to elaborate on the specific causes of user burden.

We believe that the UBS can be useful in helping designers to determine different types of trade-offs in their design. While it is good to reduce the overall level of user burden across all categories for a given system, it may not be possible to reduce all of them. In addition, it may be that by reducing one burden, another one increases. For example, to help reduce the privacy burden, a designer may add in a rich set of privacy control features. However, this may increase the difficulty of use burden by requiring the user to spend more time try to understand and maintain their privacy settings. Because the User Burden Scale has valid subscales, designers can look at individual scores for each subscale to determine which aspects of the system are contributing most to user burden. Because there are an unequal

number of questions within each subscale, however, average scores should be used rather than total scores when comparing across burden type.

Limitations

Although we obtained good validity results for the UBS, we acknowledge a few limitations at this time. UBS is still in its early stage of validation and has not yet been tested on a large scale with large Ns for researchers or practitioners aiming to improve their technology design. Validating the questionnaire in practice with designers or researchers could provide more insights on how to best use the UBS. We encourage researchers and practitioners to use our scale and provide us feedback on the usefulness of the results so that we can continue to develop and refine the scale for broader ranges of use. In addition, the scoring guidelines we provide in this paper are based on 750 responses on ~150 technologies. We expect scoring guidelines may change as people use it more and the community builds larger database.

There are other aspects of validity that still need to be tested. For one, we did not conduct a test-retest validation due to the difficulty of following up with online, anonymous participants. We also have not yet run predictive validity to determine if the UBS can predict if someone might abandon a technology later. There are also some general issues and limitations with numerical scales and subjective measures [30, 39]. Although user burden is intended to be inherently subjective, we do encourage administrators to use the UBS in combination with other more objective tests to gain an overall complete picture of user experience. Finally, although the user population on Mechanical Turk, which we used for our tests of validity, is relatively diverse for an Internet sample [11], it would still be prudent to test whether the UBS holds across different populations and different cultures.

Future Directions

Future work will explore the use of the User Burden Scale with larger populations and investigations on different types of systems and interview researchers and practitioners on the overall usefulness of the scores for improving the understanding of the impact of the design of their systems. This includes also exploring whether we can create a version of the UBS that can be used to predict user burden for earlier stage prototypes. We also plan to conduct additional testing with participants across varying demographics, especially relating to age, gender, education level, cultural background, and technology expertise, to ensure that the scale is widely applicable across all populations.

CONCLUSION

In this paper, we described the design and validation of a new scale for assessing user burden in computing systems, called the User Burden Scale. User burden is a model for characterizing the ways that computing systems might have a negative impact on the user across six different constructs: 1) difficulty of use, 2) physical, 3) time and social, 5) mental and emotional, 5) privacy, and 6) financial. We believe that user burden is a unique but important view of overall user experience that has not yet been supported through specific, lightweight measures. The User Burden Scale is intended for use with current or past users of systems to help researchers and practitioners to understand different aspects of user burden their users are experiencing. We hope that this scale can be useful for researchers and practitioners alike in understanding the ways that computing systems can have an impact on the user's lives beyond just issues of usability and enjoyment. Although our scale has been validated, there are some limitations to the scale's use that should be considered during use. Future work will seek to understand more about how the UBS can be used in broader contexts and how useful it is in helping to improve the design of computing systems.

ACKNOWLEDGMENTS

This research was reviewed and approved for exemption by University of Washington Human Subjects Division. Research was funded by National Science Foundation grants #0952623 and #1344613. We also thank Svetlana, Yarosh, Divya Addepalli, Chih-Wei Chen, Cynthia Bennett, and Nicole Tidwell for their assistance in this research. Finally, we would like to thank the reviewers and associate chairs for their very helpful and insightful reviews on this paper.

REFERENCES

1. Sajay Arthanat, Stephen M. Bauer, James A. Lenker, Susan M. Nochajski, and Yow Wu B. Wu. 2007. Conceptualization and measurement of assistive technology usability. *Disability & Rehabilitation: Assistive Technology* 2, 4: 235-248.
2. Ilhan Aslan, Martin Murer, Verena Fuchsberger, Andrew Fugard, and Manfred Tscheligi. 2013. Workload on your fingertips: the influence of workload on touch-based drag and drop. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces* (ITS '13), 417-420.
3. Sasikanth Avancha, Amit Baxi, and David Kotz. 2012. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys* (CSUR) 45, 1: 3.
4. James E. Bailey and Sammy W. Pearson. 1983. Development of a tool for measuring and analyzing computer user satisfaction. Management science 29, 5: 530-545.
5. Louise Barkhuus and Anind K. Dey. 2003. Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. In *INTERACT*, 3: 702-712.
6. Thomas Beauvisage. 2009. Computer usage in daily life. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '09), 575-584.
7. John L. Bennett 1984. Managing to meet usability requirements: establishing and meeting software development goals. *Visual display terminals*, 161-84.
8. Mark Bilandzic. 2010. The embodied hybrid space: designing ubiquitous computing towards an amplification

of situated real world experiences. *In Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (OzCHI '10), 422-427.

9. Richard E. Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* Sage.
10. John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194: 4-7.
11. Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science* 6, 1: 3-5.
12. Rudy Den Buurman. 1997. User-centred design of smart products. *Ergonomics* 40, 10: 1159-1169.
13. Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M. Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems* (TiiS '12) 2, 4: 22.
14. John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '88), 213-218.
15. Lee J. Cronbach and R. L. Thorndike. 1971. Educational measurement. *Test validation*, 443-507.
16. Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
17. Fred D. Davis. 1993. User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *International journal of man-machine studies* 38, 3: 475-487.
18. Louise Demers, Rhoda Weiss-Lambrou and Bernadette Ska. 2002. The Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST 2.0): an overview and recent progress. Technology and Disability 14, 3: 101-105.
19. David De Vaus. 2002. *Analyzing social science data: 50 key problems in data analysis.* Sage.
20. Judith Donath. 2014. How social media design shapes society. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems* (CHI '14), 1057-1058.
21. Joseph S. Dumas and Janice Redish. 1999. A practical guide to usability testing. *Intellect Books*.
22. Günther Gediga, Kai-Christoph Hamborg, and Ivo Düntsch. 1999. The IsoMetrics usability inventory: an operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology* 18, 3: 151-164.
23. Joseph A. Gliem and Rosemary R. Gliem. 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*.
24. John D. Gould and Clayton Lewis. 1985. Designing for usability: key principles and what designers think. *Communications of the ACM* 28, 3: 300-311.
25. Jeroen B. Guinée. 2002. Handbook on life cycle assessment operational guide to the ISO standards." The international journal of life cycle assessment 7, 5: 311-313.
26. Robert Harmon, David Raffo, and Stuart Faulk. 2003. Incorporating price sensitivity measurement into the software engineering process. In *Management of Engineering and Technology* (PICMET'03), *Technology Management for Reshaping the World. Portland International Conference on*, 316-323.
27. Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology,* 52: 139-183..
28. Leanne M. Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert JK Jacob, Angelo Sassaroli, and Sergio Fantini. 2009. Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 2185-2194.
29. Jason I. Hong, Jennifer D. Ng, Scott Lederer, and James A. Landay. 2004. Privacy risk models for designing privacy-sensitive ubiquitous computing systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (DIS '04), 91-100.
30. Salar Jahedi and Fabio Méndez. 2014. On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization* 98: 97-114.
31. Beom Suk Jin and Yong Gu Ji. 2010. Usability risk level evaluation for physical user interface of mobile phone. *Computers in Industry* 61, 4: 350-363.
32. Jurek Kirakowski and A. Dillon. 1988. The computer user satisfaction inventory (CUSI): Manual and scoring key. *Cork, Ireland: Human Factors Research Group, University College of Cork*.
33. Jurek Kirakowski and Mary Corbett. 1993. SUMI: The software usability measurement inventory. *British journal of educational technology* 24, 3: 210-212.
34. Jonathan Klein, Youngme Moon, and Rosalind W. Picard. 2002. This computer responds to user frustration:: Theory, design, and results. *Interacting with computers* 14, no. 2 (2002): 119-140.

35. Shelah Leader, Phillip Jacobson, James Marcin, Ralph Vardis, Mark Sorrentino, and Dennis Murray. 2002. A method for identifying the financial burden of hospitalized infants on families. *Value in Health* 5, 1: 55-59.
36. Han X. Lin, Yee-Yin Choong, and Gavriel Salvendy. 1997. A proposed index of usability: a method for comparing the relative usability of different software systems. *Behaviour & information technology* 16, 4-5: 267-277.
37. Jakob Nielsen. 1991. Usability metrics and methodologies. *ACM SIGCHI Bulletin* 23, 2: 37-39.
38. Jum C. Nunnally, Ira H. Bernstein, and Jos MF ten Berge. 1967. *Psychometric theory.* Vol. 226. New York: McGraw-Hill.
39. Jane Ogden and Jessica Lo. 2012. How meaningful are data from Likert scales? An evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of health psychology* 17, 3: 350-361.
40. Helen Petrie and Nigel Bevan. 2009. The evaluation of accessibility, usability and user experience. *The universal access handbook*: 10-20.
41. Ahmad Rahmati, Chad Tossell, Clayton Shepard, Philip Kortum, and Lin Zhong. 2012. Exploring iPhone usage: the influence of socioeconomic differences on smartphone adoption, usage and usability. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services* (Mobile HCI '12), 11-20.
42. Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: how to plan, design and conduct effective tests.* John Wiley & Sons.
43. Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W. Picard. 2002. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with computers* 14, 2: 93-118.
44. Brian Shackel. 1984. The concept of usability. *Visual display terminals: Usability issues and health concerns*: 45-87.
45. Ben Shneiderman. 1992. *Designing the user interface: strategies for effective human-computer interaction.* Vol. 3. Reading, MA: Addison-Wesley.
46. Deb Sledgianowski and Songpol Kulviwat. 2008. Social network sites: antecedents of user adoption and usage. *Americas Conference on Information Systems (AMCIS) Proceedings:* 83.
47. Godwin J. Udo. 2001. Privacy and security concerns as major barriers for e-commerce: a survey study. *Information Management & Computer Security* 9, 4: 165-174.
48. Robert S. Weiss. 1995. *Learning from strangers: The art and method of qualitative interview studies.* Simon and Schuster.
49. Anna M. Wichansky. 2000. Usability testing in 2000 and beyond. *Ergonomics* 43, 7: 998-1006.
50. David Wright. 2011. Should privacy impact assessments be mandatory?. *Communications of the ACM* 54, 8: 121-131.
51. Svetlana Yarosh, Panos Markopoulos, and Gregory D. Abowd. 2014. Towards a questionnaire for measuring affective benefits and costs of communication technologies. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14), 84-96.