Applying Psychometrics to Measure User Comfort when Constructing a Strong Password

S M Taiabul Haque[#] Shannon Scielzo[*] Matthew Wright[#]
eresh03@gmail.com, scielzo@uta.edu, mwright@cse.uta.edu
[#]Department of Computer Science and Engineering
[*]Department of Psychology
University of Texas at Arlington, USA

ABSTRACT

As mobile devices become increasingly common for accessing services online, the security of these services in turn depends more on password entry on these devices. Unfortunately, users are not comfortable with existing textual password entry mechanisms on mobile phone handsets. In this study, we investigate this issue of user comfort from the viewpoint of psychometrics. By applying standard techniques of psychometrics, we develop a questionnaire (known as a scale in psychometrics) that measures the comfort of constructing a strong password when using a particular interface. We establish the essential psychometric properties (reliability and validity) of this scale and demonstrate how the scale can be used to profile password construction interfaces of popular smartphone handsets. We also theoretically conceptualize user comfort across different dimensions and use confirmatory factor analysis to verify our theory. Finally, we highlight several issues related to scale development and discuss how psychometric approaches may be useful in general for measuring various subjective concepts that are related to usable security.

Categories and Subject Descriptors

Security and Privacy [**Human and Societal Aspects of Security and Privacy**]: Usability in Security and Privacy; Human-centered Computing [**Human Computer Interaction (HCI)**]: Empirical Studies in HCI

General Terms

Security; Measurement; Human Factors

Keywords

Psychometrics; Questionnaire; User Comfort; Mobile Password Entry

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

1. INTRODUCTION

Password entry on input-constrained devices such as mobile phone handsets is fraught with usability problems. Jakobsson et al. report that the time-consuming and error-prone operation of password entry annoys users more than lack of coverage, small screen size, or poor voice quality [19].

This poor user experience clearly undermines the usability of sensitive security systems that are developed for mobile platforms (mobile banking, for example). According to Whitten and Tygar in their seminal paper, a security system is deemed to be usable if "people are sufficiently comfortable with the interface to continue using it" [43]. In a mobile banking system, a user is required to type her entire password by using the mobile phone keypad (i.e. no "remember me" option) each time she intends to log in to her bank account. Thus, user frustrations over password entry on mobile handsets could undermine the usability of mobile banking as a whole, as well as other security systems on mobile devices.

Since "frustration" and "comfort" are subjective psychological concepts, it is not a straightforward task to measure the level of comfort a user feels when using the interface of a security system. According to psychology researchers [32, 28, 41], merely asking "How comfortable are you with the interface of this security system?" is not enough in this case for three reasons. First, a single question lacks scope to represent a complex psychological concept such as comfort. Just as a single question can not measure intelligence, a single question is not sufficient for measuring one's level of comfort. Second, a single question can only categorize people into a small number of groups, thus limiting the ability to finely discriminate levels of comfort. Third, any individual question has a considerable amount of measurement error associated with it. When multiple questions are asked and the response scores are summed to get a total score, this error tends to average out.

For these reasons, to measure complex psychological concepts such as "frustration" or "comfort", psychology researchers develop a set of questions that meets some widely agreed upon specific statistical criteria. In fact, a separate branch of psychology has evolved in this regard, which is known as *psychometrics*. Psychometrics concentrates on developing and validating questionnaires or tests that are used for assessing knowledge, attitudes, abilities, or personality traits.

In this work, we adopt the methods of psychometrics to develop a questionnaire, also called a *scale*, for measuring the comfort of constructing a strong password when using a particular interface. We first use expert opinions to guide the

creation and selection of questions and then assess our questionnaire for reliability and validity, the two essential psychometric properties of a scale. To this end, we conducted two user studies where we administered the questionnaire to undergraduate students from different majors and analyzed their responses. We find that our questionnaire meets all of the requirements for reliability and validity for a psychometric scale: it is consistent, complete, accurately focused, and capable of predicting certain real-world outcomes.

Through a separate user study, we evaluated the password construction interfaces of popular smartphone handsets by using our scale, where the interface of iPhone was rated the most comfortable by the participants. The results of these studies demonstrate that our scale can be used effectively to measure user comfort during a password entry operation on a mobile handset.

Based on certain observations, we further shorten our scale while maintaining the diversity of interface quality evaluation. We hypothesize a specific theory about user comfort in constructing a strong password by conceptualizing comfort across several factors and build a four-factor model. This model is helpful to explain why a particular interface is more comfortable to use than another one. We employ confirmatory factor analysis, a widely used statistical method in psychometrics, and find that our collected user responses fit the model we developed.

To the best of our knowledge, despite being a well-developed field, psychometrics has not been applied in usable security to develop and evaluate questionnaires. We believe that our work paves the way for applying the techniques of psychometrics in measuring various subjective concepts that are associated with usable security. In particular, our psychometric approach of measuring comfort can be generalized to measure whether people are sufficiently comfortable with other security-related interfaces, such as anti-virus systems, personal firewalls, privacy tools for the Web, and encryption software. This, in turn, would be helpful to understand in what ways that interface is usable or not, according to the working definition of usable security as provided by Whitten and Tygar [43].

The paper is organized as follows. In Section 2, we briefly discuss psychometrics and highlight related works. We describe all the steps of our scale development effort in Section 3. Section 4 illustrates how using our scale can profile password construction interfaces of different handsets. We present our factor analysis results in Section 5. We discuss several issues related to scale development in Section 6 and shed light on future research directions in Section 7.

2. BACKGROUND AND RELATED WORK

In this section, we first briefly cover related work on password entry on mobile devices. We then provide background on psychometrics and discuss prior efforts in measurement in the fields of HCI and usable security.

2.1 Mobile Device Password Entry

Several researchers have sought a better method for entering passwords on mobile devices. By taking advantage of the auto-correction and auto-completion features of mobile handsets, Jakobsson and Akavipat implement a novel password-entry method called *fastword* [18]. Haque et al. propose a modified keyboard layout for inserting digits and special characters [14]. Despite a few limitations, both methods effectively help users to construct stronger passwords on mobile handsets [18, 14]. Other researchers have evaluated the usability of graphical password schemes on smartphones [3, 38].

The basic motivation for all these works on password entry on mobile devices stems from the realization that users are not comfortable with the existing textual password construction interfaces. In our current work, we attempt to systematically measure this user comfort in various dimensions.

2.2 Psychometrics

Psychometrics is the study of measuring complex psychological concepts, or *constructs*, such as a person's motivation, anger, personality, intelligence, attachment, or fear [30]. Since a construct is not a concrete material in the visible world, measuring a construct is not a straightforward task. For example, we know how anger looks, but we cannot describe in meters or grams how much anger a person feels. Psychometrics provides guidance to systematically develop and test a scale to measure this kind of psychological construct. In psychometrics, the basic component of a scale is referred to as an *item*. Items can be questions, true-false statements, or rating scales.

Although the field of psychometrics has been developed for measuring psychological constructs, we observe that the techniques of psychometrics may be suitable for other abstract constructs that concern human feelings and performance. The core function of psychometrics is to assign numbers to observations in a way that best allows people to summarize the observations. In other words, it tries to measure the psychological construct in a meaningful and interpretable way. Since usability is also an abstract construct [29], we believe that the techniques of psychometrics would be helpful in measuring the usability features of a security system in a meaningful and interpretable way.

2.2.1 Reliability and Validity

Let C be an arbitrary construct, such as happiness. At any given point in time, a person has a true level of happiness, namely X_T. A psychometric scale developed for measuring happiness, if administered on that person, will produce an observed level of happiness, namely X_O. The core job of a psychometrician is to develop a scale that produces a score X_O that approximates X_T as closely as possible.

The relationship between X_T and X_O can be formulated in this way [5]:

$$X_T = X_O + X_S + X_R, \tag{1}$$

where X_S comes from systematic sources of error and X_R comes from random sources of error. X_S refers to errors resulting from underlying stable characteristics of the construct, while X_R refers to errors that result from transient personal factors.

A scale may be characterized by two properties: reliability and validity. Reliability is the degree to which a scale produces stable and consistent results, and high reliability is indicated by low values of X_R (low random error). For example, if a person measures the weight of a penny several times by the same scale and always receives the same result, the scale is reliable. Validity is the degree to which a scale measures what it is purported to measure, and high validity

is indicated by low values of both X_R and X_S (both low random and systematic errors). Note that a reliable scale may not be valid, such as if the scale consistently indicates that the penny weighs 100 kg., it suffers from high systematic error.

Although there are many validity classifications, one of the most prevalent frameworks recommends assessing validity from three perspectives: *content validity*, *construct validity*, and *criterion-related validity* [31, 13, 1]. Content validity refers to the extent that a scale represents a given construct, i.e. the extent to which the content domain of the construct is represented in its entirety, and also the extent that items in the scale only represent the construct of interest. Construct validity refers to the extent that a scale assesses the underlying construct it is supposed to assess, i.e. whether the scale is accurately focused. Criterion-related validity, on the other hand, is the degree to which a scale score predicts meaningful outcomes in a real-life situation.

A sound psychometric scale should be reliable and valid in all three ways to have any meaningful application. As pointed out by Nunnally, however, validity is not an all-or-nothing property, rather it is a matter of degree [31].

2.2.2 Framework

The appropriate steps for developing a scale ultimately depend on the construct. Psychometricians have to know a number of tools and methodologies and have a thorough understanding of the construct to be measured so as to find the best mechanisms for developing and assessing the efficacy of the intended scale. There is substantial debate in the field in regards to the specific steps to employ to ensure the highest levels of validity. For example, marketing researchers often focus solely on differences due to stimuli changes, whereas psychology researchers are oftentimes interested in individual differences [5, 32]. However, there appears to be a consensus that comprehensive efforts that employ techniques from numerous perspectives are the most effective. We thus sought recommendations from various sources and applied heuristics from both marketing and psychology perspectives.

Our work is primarily based on the approach outlined by Nunnally in his various books [30, 31, 32][1], but we also considered the recommendations provided by other notable psychometricians and statisticians, including Churchill [5], Parasuraman [33], and Kaiser [21].

2.3 Psychometrics in HCI

In the existing literature on usable security, we have not found any instance of applying psychometrics to address a particular usability issue. HCI researchers, however, have adopted psychometric approaches to measure user satisfaction. Usability questionnaires such as SUMI (Software Usability Measurement Inventory) [23], QUIS (Questionnaire for User Interaction Satisfaction) [4], and MPUQ (Mobile Phone Usability Questionnaire) [36] were developed by following psychometric approaches. Sauro and Lewis employed factor analysis, a statistical method widely used in psychometrics, to identify the underlying factors or dimensions of usability [37].

In another work, Lewis evaluated the psychometric properties of four existing IBM questionnaires that were developed for measuring user satisfaction with computer system usability [27]. He provided the questionnaires to different users after they had completed certain computer tasks and asked them to express their opinion about the computer system they had just interacted with. By analyzing the response scores and measuring the reliability and validity, he concluded that all the questionnaires have acceptable psychometric properties, thus allowing usability practitioners to use them with confidence for measuring user satisfaction with different computer systems.

2.3.1 Psychological Approach in Usable Security

Our effort to apply techniques from psychometrics to address a usability problem is inspired from the observation that psychological approaches have been helpful to solve other usable security problems. A notable example of this is the work of Jaferian et al., who applied *activity theory*, a revolutionary theory originating in Soviet psychology [26], to develop a set of heuristics for evaluating the usability of IT security management tools [17]. Their results demonstrated that the heuristics performed well in identifying usability problems.

3. SCALE DEVELOPMENT STEPS

We now describe all the steps of our scale development procedure. We use the terms "layout" and "interface" interchangeably in the remainder of this paper. For performing some of the statistical calculations, we used R packages such as *psych* and *nFactors*.

3.1 Domain Specification and Initial Item Pool Generation

The first step in developing a scale for measuring a construct is to specify the domain of the construct. A researcher must understand the construct thoroughly and determine its scope: what to be included and what to be excluded [5]. For example, in the context of our current work, "comfort of constructing a password" and "comfort of constructing a strong password" are two different constructs. The former mainly refers to general typing experience and may undervalue issues like "How easy is it to insert a special character in this layout?", which is an important consideration for a strong password.

Churchill recommends performing a literature search and an experience survey for specifying the domain of the construct and generating the initial item pool [5]. An experience survey involves consulting a group of people who are considered to be knowledgeable in the domain. We conducted such a survey by forming a panel of two password researchers and two mobile UI specialists. We also consulted with expert researchers from marketing and psychology to obtain more substantive insights about the scale development procedures. A one-on-one session was held with each of the panel members.

The marketing expert recommended to review the existing scales that have been developed to measure user engagement or customer satisfaction for various activities performed with a computer or mobile phone (online shopping, for example). The psychology expert suggested that we consider emotional

[1] Nunnally's seminal book "Psychometric theory", published in 1967, had been widely used as the primary textbook in basic psychometric courses [30]. Eleven years later, he published a second edition by incorporating the new ideas that had been introduced over the decade [31]. He also co-authored a book with Bernstein, a notable clinical psychologist [32].

or cognitive hindrances such as frustration or confusion that might affect the password construction activity. The mobile UI specialists recommended that we consider subtle typing issues, such as key sensitivity and inter-key distance, which are associated with the user experience when typing on a particular layout. The password researchers focused more on entropy and were interested to observe how different keypad layouts would affect the frequency of using capital letters, digits, and special characters when constructing a new password by using those layouts.

After consulting with all the panel members and reviewing the relevant literature, we generated an initial pool of 32 items.

The first set of items was developed to assess the ease of using a specific layout to construct a strong password. The strength of a password is associated with its length and the frequency of uppercase letters, digits, and special characters (see Section 6.2 for more discussion about password strength). Items in this category directly focused on assessing how easily a user could type an uppercase letter and insert digits and special characters by using a specific layout. We also conjecture that a user would be motivated to type a longer password if her general typing experience is good when using a specific layout. Thus, we tried to capture the general typing experience of a user through this set of items. Accordingly, items in this category focused on issues like ease of editing, key sensitivity, and inter-key distance.

1. It was easy to type an uppercase letter in this layout.
2. It was easy to insert a numeric digit in this layout.
3. It was easy to insert a special character in this layout.
4. It was easy, overall, to type passwords using this layout.
5. I could easily type the exact letter that I wanted to type in this layout.
6. The distance between the keys was not very close in this layout.
7. The keypad of this layout was too much sensitive to my touch.
8. The keypad of this layout was too little sensitive to my touch.
9. It was easy to make edits when typing in this layout.
10. The keys were marked with familiar symbols in this layout.
11. I could clearly see the keys in this layout.
12. It was easy to type using both hands in this layout.

As pointed out by the psychology expert, emotional and cognitive hindrances might adversely affect the password construction activity of a user. The second set of items reflected this direction and were written as reverse-coded items [28]. Consequently, the wording of the items reflected negative connotations such as "annoyance", "error", "confusion", and "restriction".

13. I felt annoyed when typing an uppercase letter in this layout.
14. I felt annoyed when inserting a numeric digit in this layout.
15. I felt annoyed when inserting a special character in this layout.
16. I felt frustrated, overall, when typing passwords using this layout.
17. I made more errors in this layout when typing.
18. It was confusing trying to find some keys in this layout.
19. I found this layout confusing to use when I was typing an uppercase letter.
20. I found this layout confusing to use when I was inserting a numeric digit.
21. I found this layout confusing to use when I was inserting a special character.
22. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.
23. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.
24. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.
25. The current method of typing in this layout restricted me from typing a longer password.

The final set of items targeted user satisfaction. Items in this category addressed whether the users felt that there should be an easier way to insert digits or special characters, whether they were able to type quickly by using the layout, and so on.

26. I want an easier method of typing an uppercase letter in this layout.
27. I want an easier method of inserting a numeric digit in this layout.
28. I want an easier method of inserting a special character in this layout.
29. I was able to quickly type an uppercase letter in this layout.
30. I was able to quickly insert a numeric digit in this layout.
31. I was able to quickly insert a special character in this layout.
32. I was able to quickly type passwords using this layout.

3.2 Content Validity Assessment

After generating the initial item pool, the items were subjected to an assessment of content validity. As mentioned before, content validity refers to the extent to which a scale represents the content domain of a construct [12]. Content validity should be assessed immediately after developing the items, as this provides an opportunity to refine the items before making large investments in administering the items to a sample population [39, 35].

We assessed the content validity of each item by following Lawshe's guidelines [25]. Lawshe proposes forming a panel of subject matter experts and asking each of them to rate each item in terms of whether the knowledge or skills measured by that item is "essential", "useful, but not essential", or "not necessary" to the performance of measuring the construct. He developed a formula for measuring the content validity of each item [25]:

$$CVR = \frac{(n_e - \frac{N}{2})}{\frac{N}{2}}, \tag{2}$$

where CVR stands for *content validity ratio*, n_e is the number of panelists indicating that the item is "essential", and N is the total number of panelists. Lawshe also provides a table of critical values of CVR for a given size of subject matter expert panel [25]. According to his recommendation,

an item can be retained if its CVR value exceeds the critical value. Accordingly, we formed a panel of eight subject matter experts and asked them to evaluate our initial set of items. We recruited mobile application developers with at least two years of experience in working with mobile UI as our subject matter experts. They were explained beforehand about the purpose of the scale and the association between a strong password and a particular layout.

Out of 32 items, 19 items were retained (see Table 1 for the list of retained items), as their CVR was higher than the 0.75 threshold recommended in Lawshe's table for a panel of eight subject matter experts. Two psychometricians reviewed the wordings of the retained items to avoid ambiguity.

3.3 Initial Scale Administration – Study 1

Using the retained items, we then conducted a laboratory study for the purpose of testing the psychometric properties of the selected items. Specifically, the study was designed to not only collect responses from participants to examine their patterns, but to also examine whether participants' responses would change systematically in response to changes in the stimuli (the interface) being rated.

The study was administered through the *research pool* of the Department of Psychology of the University of Texas at Arlington (UTA). The pool is used to assign partial course credits to students taking an introductory course in psychology and extra credit for some advanced elective courses. Any study conducted through the pool can draw a diverse set of participants, because most of these courses are offered to majors from all departments of the university.

Researchers who collaborate with the Department of Psychology can post a brief description about their studies to the pool. Students in the research pool can view all the studies and sign up for those that interest them.

Participants.

A total of 49 undergraduate students (28 female and 21 male) signed up and participated in our study for course credit. Written informed consent was obtained from each participant.

Material.

Three layouts were used as the conditions for the study: (a) mobile phone with touchscreen keypad layout, (b) mobile phone with physical keyboard layout, and (c) computer keyboard layout. We used a within-group experimental model where each participant used all the three layouts to construct passwords.

For this study, we used a Motorola MILESTONE A853 mobile handset running Android 2.1. This handset features both a QWERTY-type touchscreen keypad and a slide-out physical keyboard. Each participant was asked to construct passwords by using both of these layouts and also a standard desktop computer keyboard.

Procedure.

First, we asked each participant to construct new passwords by using the one layout for two banking websites: Chase.com and Wellsfargo.com. We wanted the participants to construct long passwords that would contain uppercase letters, digits, and special characters. To protect their security, we explicitly told the participants not to provide any of their existing passwords or any of the passwords they had previously used. For Chase.com, the participants were presented with the following scenario:

> "Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. You have reached the final step of creating your new account and you need to create a strong password (a password that is long and contains uppercase and lowercase letters, digits, and special characters). Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password."

When they clicked OK, the password construction page appeared. Once they constructed the password for Chase.com, a similar scenario was presented for Wellsfargo.com.

Next, the participants were asked to type five fixed passwords. These fixed passwords were from seven to thirteen characters long and contained multiple uppercase letters, digits, and special characters (TRoub@dor!123, for example).

After a participant finished typing the fixed passwords, she was asked to evaluate the layout by using the 19-item scale. The items were randomly ordered to avoid any ordering effects. A 5-point Likert scale (anchored by 1 = "strongly disagree", 5 = "strongly agree") was used to capture the participants' responses. We note the difference between a Likert-type item and a Likert scale. A Likert-type item is a single question or statement and it falls into the category of ordinal level data. A Likert scale, on the other hand, is composed of multiple Likert-type items. The responses for the individual items are combined and then averaged to obtain a final scale score. Likert scale data are analyzed at the interval measurement scale and descriptive statistics like mean/standard deviation and statistical methods like ANOVA could be used in this regard [15].

The same process was then repeated for the second and third layouts. The order of the layouts was randomized for each participant.

Overall, each participant typed seven passwords in each layout. Out of these seven passwords, two were selected by the participant and five were given by us. The only reason for asking them to construct two of their own passwords was to ensure that they would be able to properly respond to the four items related to "restriction" (items 9-12 in Table 1). When administering the scale to the participants, we also modified these items slightly to emphasize new password construction. For example, item 9 was written in this way "When I was constructing a new passsword for the two banking websites, the current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords".

We note that we did not use deception in this study; the participants were directly asked to construct and type passwords. We also did not store any of their passwords. Given the nature of the scale and the relative lack of consequences (e.g. no embarrassment, no reason for responding dishonestly), there was no reason for hiding the true intent of the study at this stage of scale development. Similarly, participants were free to provide suggestions or concerns regarding

Table 1: Reliability Analysis. Cronbach's α value is 0.96.

Item	Corrected Item-total Correlation
1. It was easy to type an uppercase letter in this layout.	0.76
2. It was easy to insert a numeric digit in this layout.	0.80
3. It was easy to insert a special character in this layout.	0.83
4. It was easy, overall, to type passwords using this layout.	0.90
5. I felt annoyed when typing an uppersace letter in this layout.	0.73
6. I felt annoyed when inserting a digit in this layout.	0.77
7. I felt annoyed when inserting a special character in this layout.	0.75
8. I felt frustrated, overall, when typing passwords using this layout.	0.83
9. The current method of typing an uppercase letter in this layout restricted me from using more uppercase letters in my passwords.	0.77
10. The current method of inserting a numeric digit in this layout restricted me from using more digits in my passwords.	0.78
11. The current method of inserting a special character in this layout restricted me from using more special characters in my passwords.	0.75
12. The current method of typing in this layout restricted me from typing a longer password.	0.78
13. I could easily type the exact letter that I wanted to type in this layout.	0.75
14. It was easy to make edits when typing in this layout.	0.75
15. It was easy to type using both hands in this layout.	0.62
16. I was able to quickly type an uppercase letter in this layout.	0.77
17. I was able to quickly insert a numeric digit in this layout.	0.82
18. I was able to quickly insert a special character in this layout.	0.81
19. I was able to quickly type passwords using this layout.	0.89

the items and the layouts. Upon completion of the required tasks for each condition, participants were asked to evaluate their experience by using the item list.

As each participant evaluated three layouts, we collected a data set with a total of 147 evaluations. The scores of the reverse-coded items were inversed before adding them to the data set. There were no missing data points. We used this data set to assess the reliability and the validity of our 19-item scale.

3.4 Reliability Analysis

We first assessed the reliability of our scale. Nunnally points out that reliability is a necessary precondition for validity [31]. There are several types of reliability estimates: inter-rater reliability, test-retest reliability, parallel-forms reliability, and internal consistency. In his landmark paper, Churchill strongly emphasizes internal consistency over the other types of reliability [5]. For a Likert scale like ours, internal consistency is the reliability estimate that is most frequently reported [11].

Internal consistency of a scale is calculated based on the covariations between different items of that scale. It measures whether multiple items that are generated to measure the same general construct produce similar scores. For example, if a participant expresses agreement with the item "It was easy to type an uppercase letter in this layout" and disagreement with the item "I felt annoyed when typing an uppercase letter in this layout", it would indicate good internal consistency. Internal consistency can be measured statistically by calculating the Cronbach's alpha [7].

Since Cronbach's alpha usually increases as the covariations among items increase, a low Cronbach's alpha value suggests that the items are possibly not measuring the same construct. Along with the Cronbach's alpha value of the entire scale, the corrected item-total correlation values of the individual items also need to be calculated. The corrected item-total correlation value is an estimate of whether a given item is consistent with the averaged behavior of the other items. A low corrected item-total correlation value of an item would indicate that the item should be removed, as that particular item is ultimately not discriminating participants well in regards to what the remainder of the items are measuring. Nunnally recommends removing the items with corrected item-total correlation values lower than 0.30 [32]. Once these items are removed, the Cronbach's alpha should be recalculated to see whether a satisfactory value is achieved. However, if the value of Cronbach's alpha is too low, a researcher should loop back to the previous step of domain specification and item generation to find out what might have gone wrong [5].

The reliability results of our data set are shown in Table 1. Cronbach's alpha for the scale is 0.96, which is excellent according to the recommendation of George and Mallery [9]. This value is even arguably high, and suggests that some items could be removed and still maintain the general essence of what is being measured. We discuss this further in Section 5. Furthermore, all the corrected item-total correlation values were much larger than the cutoff value of 0.30, with the lowest correlation at 0.62. We therefore retained all the items at this point.

3.5 Construct Validity Assessment

We assessed the construct validity of our scale through a technique called the *known-groups* method [16], which involves administering the scale to conditions/groups expected to differ due to known characteristics [34]. For example, a scale to measure the construct of "fun" should show a large difference between subjects playing a video game and subjects made to wait with nothing to do. If the conditions/groups have a significant difference between their mean scores on the scale, this provides evidence for the scale's construct validity, since this indicates that it is able to discriminate among conditions/groups that are known to be different. In other words, this indicates that the scale effectively captures the underlying construct it is supposed to capture, which is the requirement of construct validity.

As mentioned before, we asked our participants in Study 1 to construct passwords in three different conditions. In addition to two types of mobile keypad/keyboard layouts, they were also asked to construct passwords by using a computer keyboard. The computer keyboard condition was added so that we would have a known "comfortable" condition. Constructing a strong password on a computer keyboard is easier than constructing it on a mobile keypad/keyboard due to the space constraints of the mobile device and the inconvenience of capitalizing letters and inserting digits or special characters. For example, on an iPhone, one additional click is required for each shift to and from digits, and this shift presents a different keypad view to the user. On the other hand, digits can be inserted in the same way as letters on a computer keyboard.

We compared aggregated means in the two mobile conditions to the computer condition via repeated-measure ANOVA. Mean scores for the combined mobile conditions ($M = 3.32$, $SD = .79$) were significantly lower than for the computer condition ($M = 4.39$, $SD = .68$), $F(1, 47) = 90.92$, $p < .05$. This established the construct validity of our scale.

3.6 Criterion Validity Assessment – Study 2

Criterion-related validity tests the relationship between a scale score and a particular outcome. For example, in the United States, SAT scores are used to determine whether a student will be successful in undergraduate studies. Here, the criterion for success for an undergraduate student may be her first-year GPA. If her SAT score correlates positively with her first-year GPA, it would indicate that her SAT score has effectively predicted her future performance in college, thus demonstrating an evidence of the criterion-related validity of the SAT.

In our case, in order to demonstrate evidence for criterion-related validity of our scale, we selected two outcomes that are potentially related to comfort of constructing a strong password when using a particular layout:

- The length of the constructed password
- The total number of uppercase letters, digits, and special characters in the constructed password

Although there exists no empirical evidence that the comfort of constructing a strong password is related to the total number of uppercase letters, digits, and special characters, the experimental results of Haque et al. provide the primary rationale for this proposition [14]. Their results demonstrate that if users are presented with a more comfortable mobile handset interface for entering digits and some special characters, they construct passwords that contain significantly more digits and special characters [14]. As for length, we implicitly assume that the more comfort a user feels when using a particular interface, the longer her typed password would be.

In order to observe the correlation between our construct of interest and the selected outcomes, we conducted a separate study.

Participants.

A total of 30 undergraduate students (17 male and 13 female) from UTA voluntarily participated in this study, and they were recruited from a course on computer literacy. The course is offered to majors from all departments and gets a diverse set of students. In exchange for their time, students were assigned extra course credits. Written informed consent was obtained from each student, and an alternative extra credit assignment was offered to the students who were not willing to participate in our study.

Material.

In this study, participants were asked to construct passwords by using one of the two layouts of our Motorola MILESTONE A853 mobile handset (see Section 3.3). Each participant was randomly assigned to one of the layouts to construct passwords. Since we collected the passwords of the participants for this study and analyzed them, we used deception in this study so that the participants would construct passwords just the way they do in real-life situations.

Procedure.

We designed this study so that it appeared to the participants as if they were opening a new bank account at Chase.com. They were asked to complete a set of tasks that resembled the usual steps of creating a new online bank account. Password construction was framed as one of these multiple tasks, not as the primary task.

The participants were first presented with the following instructions:

> "Chase is one of the largest banks in the US and it has an ATM on campus. Imagine that you are creating an account at Chase.com for online banking. Proceed to the next page to start creating your new bank account."

When a participant clicked OK, she was asked to enter dummy values given to her on a piece of paper for the following fields: name (containing both uppercase and lowercase letters), account number, address (containing multiple special characters), phone number, and email address. These tasks ensured that the participants were familiar with the

typing interface, including entering uppercase letters, digits, and special characters that would be needed for a strong password. Next the participant was asked to answer a few questions like "Do you want overdraft protection for your new account?" and "How much daily withdrawal limit do you want?". Finally, the participant was redirected to the password construction page where she was asked to construct a strong password for the new account. We did not enforce any requirement for length or the use of uppercase letters, digits, or special characters, though we did offer a hint for what a strong password is in our instructions:

> "Please create a strong password (a password that is long and contains uppercase letter and lowercase letter, digit, and special character) for your new account. Proceed to the next page to input your new password. Do not provide a password that you currently use or have previously used for any accounts. Also, do not use any confidential or personally identifiable information in your password."

After a participant finished all these steps, she was asked to evaluate the password construction experience on her assigned layout by using our 19-item scale. As with Study 1, the items were randomized and a five-point Likert scale was used to capture the responses. For each participant, we correlated the length of the constructed password and the total number of uppercase letters, digits, and special characters with the mean score from the scale.

Mean scale score vs. length

Mean scale score and length were strongly correlated, $r(28) = .51$, $p < .05$.

Mean scale score vs. total number of uppercase letters, digits, and special characters

The correlation between mean scale score and total number of uppercase letters, digits, and special characters was moderately strong, $r(28) = .41$, $p < .05$. Furthermore, we calculated mean scale scores by considering only the 12 items that are related to uppercase letter, digit, and special character (items 1-3, 5-7, 9-11, 16-18 in Table 1), and correlated these scores with the total numbers of uppercase letters, digits, and special characters. As expected, the correlation was stronger in this case, $r(28) = .47$, $p < .05$.

According to Cohen, a validity coefficient can be interpreted in this way: less than .1 is trivial; .1 to .3 is weak; .3 to .5 is moderate; and greater than .5 is strong [6]. Based on this guideline, our correlation coefficient values were satisfactory and evident of good criterion-related validity.

4. PROFILING POPULAR SMARTPHONE HANDSET INTERFACES – STUDY 3

In order to demonstrate the practical application of our scale, we evaluated the password construction interfaces (keyboard/keypad layouts) of popular smartphone handsets through our scale. We selected three handsets: BlackBerry Curve 9300, Motorola DROID 2 A955, and iPhone 4s. The iPhone handset was selected because of its touchscreen keypad layout, while the BlackBerry and Motorola handsets were representatives of QWERTY-type keyboard and slide-out physical keyboard, respectively.

We also implemented the custom touchsreen layout proposed and designed by Haque et al. as an Android app running in a Motorola MILESTONE A853 handset [14]. It involved adding two extra on-screen rows, one containing the ten digits and the other containing ten common special characters, in addition to the default Android touchscreen keypad.

Participants.

A total of 21 undergraduate (15 female and 6 male) students from UTA participated in this study. As with Study 1, we recruited participants from the research pool of the Department of Psychology (see Section 3.3). Written informed consent was obtained from each participant.

Procedure.

Since we did not require to collect any password constructed by the participants, we used exactly the same experimental design as Study 1 for this study. Participants were asked to type two passwords of their own and five fixed passwords by using each of the four layouts (see Section 3.3). After typing the passwords by using one layout, they evaluated that particular layout by using our scale.

For each layout, we calculated the mean scale score. The iPhone 4s touchscreen keypad layout was rated the most comfortable (mean = 4.19 out of 5), while Blackberry's layout was considered the least comfortable (mean = 2.78 out of 5). The Motorola layout received a moderate score (mean = 3.32 out of 5). The custom layout of Haque et al. obtained a slightly lower score (mean = 4.13 out of 5) than that of iPhone 4s.

We note that these findings should be interpreted with caution. We discuss this in detail in Section 6.2.

5. FACTOR ANALYSIS

Factor analysis is a statistical procedure that examines the correlations or covariances among items to discover clusters of related items. In psychometrics, factor analysis is often used to identify the underlying *subconstructs* that might reside in the construct of interest. These subconstructs are also referred to as *factors*, *components*, or *dimensions*. For example, in his classic paper, Spearman uses factor analysis to posit a two-factor theory for measuring human intelligence: the general intelligence factor and the specific intelligence factor [40].

Factor analysis comprises two different perspectives: exploratory factor analytic approaches and confirmatory factor analysis. Exploratory factor analysis is used when a researcher is uncertain about the theoretical conceptualization of her construct of interest. It provides a quick way to explore the underlying factors of the construct, thus providing an opportunity to refine the theory at an early stage of scale development. Confirmatory factor analysis (CFA), on the other hand, is used when the researcher has a more specific theory about the conceptualization of the construct of interest. Based on this theory, the researcher builds a model and gathers data to examine whether the data fits the hypothesized model.

Factor analyses can provide meaningful information regarding the overarching structure of the data, and can provide guidance on how best to aggregate the data after the factor. There are a number of ways to extract factors, in-

cluding principal component analysis (PCA), principal axis factoring (PAF), maximum likelihood, and more, but PCA and PAF are most frequently used. Factor rotation is an important consideration during a factor analysis. By maximizing high item loadings and minimizing low item loadings, rotation helps to produce a more interpretable factor analysis solution. There are several rotation techniques, varimax rotation is the one that is used most commonly.

We conducted a PCA with a varimax rotation (eigenvalues greater than 1) on our data set for Study 1 [20], and it was found that items loaded on one general component. The one component accounted for 62% of the variance, and item loadings ranged from 0.61 to 0.91. PCA tends to however identify one factor, and does not allow for examination of more complicated models as is possible in CFA.

Given the strength of relations obtained across items (demonstrated via both the Cronbach's alpha and the PCA), we decided that there was undue redundancy in the items and decided to cut unnecessary items. Upon careful examination of the current items and their respective relations, we were interested to examine the extent to which a higher order factor (comfort), and four corresponding second level factors (uppercase letter, numeric digit, special character, and general typing) would fit the data based on the eight retained items.

Our hypothesis was based on the observation that issues like ease of edit ("It was easy to make edits when typing in this layout"), ability to type by using both hands ("It was easy to type using both hands in this layout"), and ability to type the exact letter that the user wants to type ("I could easily type the exact letter that I wanted to type in this layout"), in turn, result in quick and easy typing of passwords. Items 13, 14, and 15, therefore, essentially capture the quickness and easiness of general typing when using a specific layout. Furthermore, the items related to frustration and restriction actually capture the cognitive and emotional hindrances of a user when constructing a strong password by using a particular layout. Intuitively, these hindrances should prevent users from typing quickly and influence their perceptions regarding the ease of using the layout.

We therefore focused exclusively on the quickness and easiness related items and posited the following four-factor theory regarding the comfort of constructing a strong password when using a particular layout:

Factor: Uppercase letter.

1. It was easy to type an uppercase letter in this layout.
2. I was able to quickly type an uppercase letter in this layout.

Factor: Numeric digit.

1. It was easy to insert a numeric digit in this layout.
2. I was able to quickly insert a numeric digit in this layout.

Factor: Special character.

1. It was easy to insert a special character in this layout.
2. I was able to quickly insert a special character in this layout.

Factor: General typing.

1. It was easy, overall, to type passwords using this layout.
2. I was able to quickly type passwords using this layout.

We used IBM SPSS Amos (version 21) to conduct CFA and evaluate the fit of our confirmatory model. The default settings (i.e. maximum likelihood) were used, with the raw data of Study 1 supplied as an input. It was found that our proposition was supported, the data fit the overarching model well ($\chi^2(16) = 24.952$, $p = .07$, $RMSEA = .06$, $PCLOSE = .31$). Contrary to other statistical models, the null hypothesis is that the model fits the data well. Thus, a chi square value that does not reach statistical significance (i.e., $>.05$) is considered indicative of good fit. Because of the extremely conservative nature of this particular statistic (i.e. rarely do arguably good fitting models meet this criteria), RMSEA and PCLOSE statistics are also typically reported. A small RMSEA value is an indicator of good fit as a value of 0.08 or less is often considered acceptable [2]. PCLOSE is a test of statistical significance for RMSEA, with the assumption that the RMSEA= .05 (i.e. close fit). A statistically significant difference again means that the theoretical model is significantly different from the actual relationships among variables (which is not in our case, hence a good fit). Thus, the statistical results demonstrate that our model is likely a good fit for our data.

These results suggest that there appears to be four highly related factors in the scale that collectively comprise our representation of user comfort. In turn, data can be averaged to the level of the scale for most purposes (or in the case of missing data, data should be averaged at the level of factor, and then the factors averaged for representative individual indicators). Similarly, if variations in comfort based on these factors need to be examined (e.g. "Is this particular layout more comfortable for typing uppercase letters?"), then the scale can effectively do so by specifically examining those specific factor values.

6. DISCUSSION

This is the first study to date that we are aware of which specifically applies psychometric principles to develop and test a scale designed to measure how well suited keyboard or keypad layouts are in the context of password construction. We have utilized numerous frameworks and conceptualizations, and extensively tested the scale in various ways to create the most accurate, useful scale possible. From extensive content validation efforts, to examination of construct validity and analysis of factor structure, to prediction of important meaningful criteria, this scale has demonstrated very promising initial evidence.

In the subsequent sections, we first discuss the ecological validity of our study and highlight the limitations of our work. Next, we discuss several issues related to scale development.

6.1 Ecological validity

As mentioned before, Study 1 and Study 3 did not involve deception, and we did not try to hide the motive of our study from the participants for these two studies. This was in accordance with the experimental methodology for scale development studies, where users are first explicitly subjected to a certain task and later asked to evaluate the experience

by using the scale. For our case, the task was to type a few passwords (five fixed, two of the users' own) by using different layouts. The experience of constructing a password was more important here, rather than the password itself. When the users were constructing their own passwords, we involved a simple role-play scenario, since prior work has shown that it is more effective than a survey scenario in motivating the users to construct passwords more seriously [24].

For Study 2, we collected passwords constructed by the participants. Ecological validity therefore was an important consideration for this study. The results of a recent work of Fahl et al. on the ecological validity of password study reveal that passwords collected during user studies closely resemble users' actual passwords [8]. We tried our best to frame Study 2 as an experiment that asks the users to perform a real-life online task, namely creating a new online bank account by using a mobile phone handset. Password construction was one of a series of steps for completing the primary task (i.e. creating the new account), just as it would be in real life. The word "password" was not used anywhere in the informed consent document. A debriefing session was arranged at the end of the study where the deception was revealed and the participants were provided with the opportunity to withdraw their consent to participate in the study. None of the participants decided to do so and we could use all of their passwords to test the criterion-related validity of our scale.

We note, however, that our participants were not required to return on a second day to re-enter their passwords, and as such, we were not completely able to emulate the real-life password construction behavior of users.

6.2 Limitations

For this work, we quantified password strength in terms of entropy, according to the recommendation of password researchers (see Section 3.1). We do not overlook the findings of Weir et al. or Kelley et al., which demonstrate that entropy is not the most appropriate measure of password strength [42, 22]. However, since our developed scale focuses on measuring the comfort of constructing a strong password when using a particular layout, we believe that entropy is a better approximation of password strength here because it effectively captures the layout-related aspects of a strong password. Alternative measures such as guessability are more dependent on the exact password choices of users and do not clearly capture aspects related to keyboard layout, such as the use of special characters. This approximation is consistent with Haque et al. [14], a related work on password strength and keypad/keyboard layout.

For all three studies, we recruited participants from university students, who may vary considerably from other populations in their smartphone usage behavior. We plan to test our scale by using a more diverse population group in future. Ultimately, scale development is a never-ending process in which developers continually strive to understand the intricacies of the conventions in regards to any meaningful variations (e.g. does my scale predict other meaningful criteria, does it behave differently in other contexts or for other sample compositions, etc.). However, in general, this scale has demonstrated solid initial evidence of its efficacy.

Our results of Study 3 should be interpreted carefully, particularly by considering the fact that we did not control for participants' previous familiarity with the interfaces. For example, if most of the participants were iPhone users, their familiarity with the iPhone layout would probably bias them towards that layout. We conducted Study 3 for demonstrating a practical application of our developed scale, not for a definitive comparison among the interfaces.

6.3 Aggregration and application

Our shortened item list has four factors, each of which contains two items (see Section 5). Since each of the underlying factors contains the same number of items, none of the factors is underestimated or overestimated when individual item scores are combined and averaged to form a final composite score. Subsequently, depending on the intended application (examination of individual level issues with comfort, or identification of problem areas with the layout), the scale score could also be computed in terms of each factor. As a result, the scale could be used to answer more specific questions like "Which layout is more comfortable for inserting a numeric digit when constructing a strong password?" or "Which layout is more comfortable for general password typing?". This provides an additional motivation for us to conduct further experiments with this shortened scale.

6.4 Norm development

After a sound scale (reliable and reasonably valid) has been developed, depending on the intended application of the scale, the researcher should continue to conduct further experiments. If the purpose of the scale is to compare different interfaces with respect to the construct of interest, then administering the scale to different users and profiling the interfaces based on scale scores should be sufficient. Our scale can currently be used in this way.

On the other hand, if the purpose of a user comfort scale is to answer the question of whether users are sufficiently comfortable with a particular user interface, then the researcher should also develop norms for her new scale. Developing norms involves setting up standard scores for a scale. Ideally, for a 5-point Likert-type scale, mean scale scores of 3 and 4 should imply neutral and positive attitudes, respectively. However, this might not be always true. For example, a mean score of 4 might represent the highest (or lowest) score ever achieved on that particular scale. To this end, the researcher should specify the benchmark scores for her new scale. This can be done by administering the scale over a large number of users to obtain a distribution of scores and subsequently characterizing the distribution by various statistical features such as mean and standard deviation. A detailed description about the norm development procedure can be found in [10].

We believe that this norm development technique could be used to specify a standard score that would represent "sufficient user comfort" in the context of a specific security system user interface. This would be helpful to precisely find out whether users are sufficiently comfortable with a particular security system user interface, which, according to the working definition of usable security in the seminal paper of Whitten and Tygar [43], is an important consideration for measuring the usability of that security system.

6.5 Revalidation study

We note that we did not prune any items during the reliability assessment stage because all of the items had a satisfactory corrected item-total correlation value and the Cron-

bach's alpha value of the overall scale was high (see Section 3.4). If items need to be pruned at this stage, a revalidation study is recommended to be conducted with the shortened scale. This involves administering the shortened scale to a new sample which is independent to the previous sample and assessing the reliability of the shortened scale.

For our scale, however, we needed to assess the criterion-related validity by using a separate study that involved deception and collection of participants' passwords. This provided us an opportunity to reassess the reliability of our scale by using a different sample. We calculated the Cronbach's alpha for this new data set. As before (0.96), the value was high enough (0.93). This provided further evidence for the reliability of our scale.

7. CONCLUSION AND FUTURE WORK

In this work, we adopted the techniques of psychometric theory to solve a specific usable security problem: measuring user comfort when using a specific interface to construct a strong password. We followed standard psychometric theory procedures to develop a questionnaire for this purpose. This involved consulting with subject-matter experts, testing an initial set of questions with a survey, statistical analysis to refine the set of questions and validate their consistency and accuracy, and conducting a separate study to demonstrate that the questionnaire is capable of predicting certain real-world outcomes. All these results establish the two essential psychometric properties of our questionnaire: reliability and validity. Thus, the questionnaire can be used to profile the password construction interfaces of popular smartphone handsets.

Based on our observations, we further shortened our questionnaire and attempted to build a specific theory about user comfort in the context of our work. We tested this theory and the shortened questionnaire by using confirmatory factor analysis, a widely used technique in psychometric theory. The results of confirmatory factor analysis align with our theory, which encourages us to conduct further studies in future to test this shortened questionnaire. We are interested to administer this version of the questionnaire with different participants and assess its psychometric properties. If the results are satisfactory, we would replace our current questionnaire with the shortened version, since the later one is more intuitive and capable of finer-grained comparisons among different interfaces across multiple factors or dimensions.

It is likely, for example, that older individuals (such as elderly populations) may report lower comfort scores when using our scale to evaluate their password construction experience. We have some preliminary information regarding the norms for our particular samples. However, we will continue administering the scale over a diverse group of users to obtain a representative distribution of scores that is generalizable to a much broader population.

To the best of our knowledge, the current work is the first to introduce the concepts of psychometric theory in usable security. In the future, we are interested to apply psychometric theory to develop reliable, valid and conceptually meaningful questionnaire for measuring user comfort when using other security system user interfaces (antivirus or encryption software user interfaces, for example). Also, we believe that the technique of factor analysis could be helpful in identifying the underlying factors or dimensions of usability in the context of a security system. We plan to work on this in future.

8. ACKNOWLEDGMENTS

We would like to thank our panelists and the anonymous participants in our user studies. We are also grateful to Mehdi Tanzeeb Hossain for reviewing the wordings of the items. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1117866.

9. REFERENCES

[1] J. D. Brown. *Testing in language programs*. Prentice Hall Regents, Upper Saddle River, NJ, 1996.

[2] M. W. Browne and R. Cudeck. Alternative ways of assessing model fit. In *Testing structural equation models*. Sage Publications, Newbury Park, CA, 1993.

[3] H.-Y. Chiang and S. Chiasson. Improving user authentication on mobile devices: A touchscreen graphical password. In *MobileHCI*, 2013.

[4] J. P. Chin, V. A. Diehl, and K. L. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *CHI*, 1988.

[5] G. A. Churchill. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73, February 1979.

[6] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum, Hillsdale, NJ, 1988.

[7] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrica*, 16(3):297–334, September 1951.

[8] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *SOUPS*, 2013.

[9] D. George and P. Mallery. *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Allyn & Bacon, Boston, 2003.

[10] E. E. Ghiselli. *Theory of psychological measurement*. McGraw-Hill, New York, 1964.

[11] J. A. Gliem and R. R. Gliem. Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. In *Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, 2003.

[12] J. S. Grant and L. L. Davis. Selection and use of content experts for instrument development. *Research in Nursing & Health*, 20(3):269–274, June 1997.

[13] R. M. Guion. On Trinitarian doctrines of validity. *Professional Psychology*, 11(3):385–398, June 1980.

[14] S. M. T. Haque, M. Wright, and S. Scielzo. Passwords and interfaces: Towards creating stronger passwords by using mobile phone handsets. In *SPSM*, 2013.

[15] J. Harry N. Boone and D. A. Boone. Analyzing likert data. *Journal of Extension*, 50(2), April 2012.

[16] J. Hattie and R. W. Cooksey. Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement*, 8(3):295–305, July 1984.

[17] P. Jaferian, K. Hawkey, A. Sotirakopoulos, M. Velez-Rojas, and K. Beznosov. Heuristics for

evaluating IT security management tools. In *SOUPS*, 2011.

[18] M. Jakobsson and R. Akapivat. Rethinking passwords to adapt to constrained keyboards. In *MoST*, 2012.

[19] M. Jakobsson, E. Shi, P. Golle, and R. Chow. Implicit authentication for mobile devices. In *HotSec*, 2009.

[20] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, 23(3):187–200, September 1958.

[21] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, March 1974.

[22] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *IEEE S&P*, 2012.

[23] J. Kirakowski and M. Corbett. SUMI: The software usability measurement inventory. *British Journal of Educational Technology*, 24(3):210–212, September 1993.

[24] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *CHI*, 2011.

[25] C. H. Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, December 1975.

[26] A. N. Leontev. *Activity, Consciousness, Personality*. Prentice Hall, Englewood Cliffs, NJ, 1978.

[27] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, January 1995.

[28] J. P. McIver and E. G. Carmines. *Unidimensional scaling*. Sage, Beverly Hills, CA, 1981.

[29] N. McNamara and J. Kirakowski. Defining usability: quality of use or quality of experience? In *IPCC*, 2005.

[30] J. C. Nunnally. *Psychometric theory (1st ed.)*. McGraw-Hill, New York, 1967.

[31] J. C. Nunnally. *Psychometric theory (2nd ed.)*. McGraw-Hill, New York, 1978.

[32] J. C. Nunnally and I. H. Bernstein. *Psychometric theory (3rd ed.)*. McGraw-Hill, New York, 1994.

[33] A. Parasuraman, V. A. Zeithaml, and L. L. Berry. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1):12–40, Spring 1988.

[34] D. Pavlas, F. Jentsch, E. Salas, S. M. Fiore, and V. Sims. The play experience scale: Development and validation of a measure of play. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(2):214–225, April 2012.

[35] J. R. Rossiter. The C-OAR-SE method and why it must replace psychometrics. *European Journal of Marketing*, 45(11):1561–1588, November 2011.

[36] Y. S. Ryu and T. L. Smith-Jackson. Reliability and validity of the Mobile Phone Usability Questionnaire (MPUQ). *Journal of Usability Studies*, 2(1):39–53, November 2006.

[37] J. Sauro and J. R. Lewis. Correlations among prototypical usability metrics: Evidence for the construct of usability. In *CHI*, 2009.

[38] F. Schaub, M. Walch, B. Konings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *SOUPS*, 2013.

[39] C. A. Schriesheim, K. J. Powers, T. A. Scandura, C. C. Gardiner, and M. J. Lankau. Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, 19(2):385–417, April 1993.

[40] C. Spearman. *The abilities of man*. Macmillan, New York, 1927.

[41] P. Spector. *Summated rating scale construction*. Sage, Thousand Oaks, CA, 1992.

[42] M. Weir, S. Aggarwal, M. Collins, and H. Stern. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *CCS*, 2010.

[43] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX*, 1999.