Journal of Usability Studies
Vol. 13, Issue 1, November 2017 pp. 17–37

SUPR-Qm: A Questionnaire to Measure the Mobile App User Experience

Jeff Sauro
MeasuringU
Principal
jeff@measuringu.com

Pareezad Zarolia
UX Researcher
MeasuringU
pzarolia@gmail.com

Abstract

In this paper, we present the SUPR-Qm, a 16-item instrument that assesses a user's experience of a mobile application. Rasch analysis was used to assess the psychometric properties of items collected from four independent surveys (N = 1,046) with ratings on 174 unique apps. For the final instrument, estimates of internal consistency reliability were high (alpha = .94), convergent validity was also high, with significant correlations with the SUPR-Q (.71), UMUX-Lite (.74), and likelihood-to-recommend (LTR) scores (.74). Scores on the SUPR-Qm correlated with the number of app reviews in the Google Play Store and Apple's App Store (r = .38) establishing adequate predictive validity. The SUPR-Qm, along with category specific questions, can be used to benchmark the user experience of mobile applications.

Keywords

user experience, Item Response Theory, Rasch, questionnaire, mobile apps, human factors, measurement

Introduction

The world continues to go mobile. Industry leaders predicted that mobile would overtake fixed Internet access by 2014. This prediction is now a certainty, with U.S. users spending 51% of their total time online accessing online content via mobile devices (Meeker, 2015). Global annual mobile app revenue exceeds $45B US (Dogtiev, 2015). To understand and improve the quality of the mobile app user experience, there is an increasing need for a valid and reliable measure of the mobile app user experience. The purpose of the present research was to develop a measure that would be equally effective at evaluating the user experience across all mobile app types (e.g., banking and gaming) and provide a benchmark as part of larger efforts to improve the app user experience.

Current Measures of Mobile App User Experience

The terms *usability* and *user experience* are used somewhat interchangeably in practice but represent different but related constructs (Tullis & Albert, 2013). Usability refers to the ability of participants to complete tasks effectively and efficiently and is embodied in an international standard (ISO, 1998). In contrast, user experience is a broader term that includes usability but also the constructs of beauty, hedonic, affective, or experiential aspects of a technology (Hassenzahl & Tractinsky, 2006). It can be applied to products where the goal is to have a good experience but effectiveness and efficiency are not a primary concern, such as with gaming apps. Under this distinction, an app can be usable but offer a poor user experience—usability is necessary although not sufficient for a good user experience. This makes the measurement of mobile app user experience a particularly challenging task, as many apps may provide a particularly pleasing user experience without conforming to traditional usability design standards.

Despite their differences, both usability and the user experience are primarily measured using similar evaluation methods (Sauro, 2016). The user experience for an app is most commonly measured using one of three methods: observing participants attempt tasks in a controlled setting (a usability test), collecting users' attitudes about an app experience in a survey, or having interface experts evaluate an app using guidelines and heuristics—often called inspection methods or expert reviews (Lewis, 2012).

One of the challenges of measuring the mobile app experience is the wide variety of apps and their disparate goals (gaming, social media, and productivity apps). For example, a particularly delightful and satisfying gaming app that outperforms competitors on meaningful metrics like number of unique downloads, may have very different characteristics than a similarly satisfying and successful mobile banking app. Perhaps for this reason, measures of the mobile app experience are often domain specific (Lin, Liu, Sadeh, & Hong, 2014; O'Malley, Dowdall, Burls, Perry, & Curran, 2014).

While these domain specific metrics are useful, a common instrument would allow developers the ability to compare the mobile app user experience across domains. Thus far, mobile phone user experience measures have focused on relatively high-level constructs such as ease of learning, efficiency, emotional responses, and memory load (Ryu & Smith-Jackson, 2005; Ryu & Smith-Jackson, 2006). On the other hand, scales such as the System Usability Scale (Brooke, 1996; Kortum & Sorber, 2015) assess lower-level, system-based usability, and a measure like the Technology Assessment Model (TAM) items only seem relevant to productivity apps (Davis, 1989).

The Standardized User Experience Percentile Rank Questionnaire (SUPR-Q) is a measure developed to capture both broad user experience constructs and more specific components of all types of full desktop websites. It does this by evaluating websites along the dimensions of usability, trust, appearance, and loyalty (Sauro, 2016). The SUPR-Q has been used to generate reliable scores in benchmarking websites and normed scores that are used to determine how well a website scores relative to others in the database. To date, there is no measure that serves a similar function as the SUPR-Q for mobile apps. The purpose of the present research was to design a measure similar to the SUPR-Q and other reliable user experience measures found in the literature to assess the user experience of mobile applications. Our goal was to create an instrument that was parsimonious while still exhibiting desirable psychometric properties.

Study Overviews

Four studies were conducted to generate a new measure of mobile application user experience. Study 1 was designed to generate items from a more exploratory and qualitative understanding of users' attitudes towards various mobile apps, the functionality of preferred apps, and the design features that makes some applications better than others. In Study 2, items generated in Study 1 were evaluated for clarity, reliability, and validity, and were modified or removed based on the results. Then, remaining items were retested by asking participants to respond to items while considering a particularly popular app, an app they felt ambivalent towards, or an app they used most recently. In Study 3, new items were tested with an additional sample and across two contexts (an app the participant used most recently and an app they had on their second screen). Finally, Study 4 established the reliability and validity of the scale with a subset of items.

Study 1: Item Identification

The purpose of Study 1 was to generate potential survey items based on previous literature and from a qualitative understanding of what makes a good mobile app experience.

Method

The initial set (selected from the literature and expected to apply to a broad range of mobile app categories) included 23 items associated with constructs of utility, usability, intended usage, reasons for deleting (Fakhruddin 2016; Varshneya, 2015), and future usage (Brooke, 1996; Lewis, Utesch, & Mayer, 2013; Ryu & Smith-Jackson, 2005; Sauro, 2016).

A total of 104 participants from Amazon's Mechanical Turk completed the Study 1 survey. Participants were located throughout the US, were a mix of gender (60% female), and varied in age (*M*= 34, *27* to 63) and device experience (59% Android, 41% iOS).

Initial data were collected in June 2016. Participants were paid $1.25 and asked to identify their favorite mobile app and answer the 23 candidate items in an online survey. Participants rated their level of agreement to each item using a 5-point Likert scale (*strongly disagree* = 1 to *strongly agree* = 5), except for the item "How likely are you to recommend the mobile app to a friend" which used an 11-point scale (Reichheld, 2003).

In addition to the 23 scale items, four free-response questions were asked to determine how participants thought and felt about their favorite mobile app. The questions were the following:

- Describe what you like most about the app and why.
- Describe your primary motivation for downloading the app.
- What features do you frequently use on the app? Please be specific.
- Under what circumstances do you use the app?

Results

A total of 104 responses were collected account for 59 unique apps. Participants were given an N/A option for the items. A high percentage of N/A responses indicated items that would not work well for a questionnaire designed to target all types of apps (e.g., a questionnaire equally applicable to gaming apps and banking apps). The percent of respondents (out of 104) who selected N/A for each item is shown in Table 1.

Table 1. Items Considered in Study 1

Abbreviation	Item	% N/A
Fun	It’s fun using the app.	2%
Enjoy	I enjoy using the app.	2%
Happy	Using the app makes me happy.	2%
Exciting	It’s exciting to use the app.	3%
Bored	I use the app when I am bored.	2%
LTR	How likely are you to recommend the app to a friend or colleague?	0%
MeetReq	The app’s capabilities meet my requirements.	0%
MobileSite	The app offers features its mobile website doesn’t.	22%
Crash	The app rarely crashes or causes problems on my phone.	0%
Bugs	The app runs without bugs or errors.	0%
Freq	I would like to use the app frequently.	0%
Misuse	The app does not misuse my information.	4%
Trust	I trust the app with my personal information.	2%
Appagain	I plan to use the app again soon.	1%
EasyNav	It is easy to navigate within the app.	1%
EasyUse	The app is easy to use.	0%
Attractive	I find the app to be attractive.	0%
Clean	The app has a clean and simple presentation.	1%
Discover	I like discovering new features on the app.	3%
Cantlive	I can’t live without the app on my phone.	1%
Delightful	The app is delightful.	0%
LearnFriends	I talk about things I do or learn on the app with my friends.	3%
CxFriends	I am able to connect or communicate with friends directly from the app.	9%

Two items had relatively high N/A rates: “The app offers features its mobile website doesn’t,” with 22% of respondents selecting it as not applicable, and “I am able to connect or communicate with friends directly from the app,” with 9% of respondents selecting it as not applicable.

Both items were flagged for removal from the subsequent studies for the general app assessment, as they only seemed to pertain to a specific type of app (i.e., apps with a corresponding mobile website or apps with a communication component). However, these items may be useful when assessing specific subtypes of mobile applications, though we did not explore this as a possibility in the present research.

Qualitative data from the four free-response questions were coded and summarized by an expert-coding strategy. An analyst read through each of the 104 responses, noting recurring themes to inform future question generation. The primary themes extracted from these responses were as follows: (a) Users liked the way their favorite app integrated with other apps or features on their phone, (b) Users enjoyed apps that allowed them to connect with their friends or colleagues, (c) Users frequently used communication features in their app, (d) Users enjoyed apps with features that integrated with real world products. These insights were used to generate more rating scale items designed specifically to probe the integration and social aspects of mobile app use.

Study 2: Item Refinement and Assessment in Three Contexts

A second study was conducted with the newly refined candidate items with responses on a 5-point Likert scale (1 = *strongly disagree* and 5 = *strongly agree*).

Method

Prior to launching the survey, a convenience sample of three participants (ages 28, 30, 44; 2 female) was used to conduct a cognitive test of item and response wording. Each participant was instructed to think out loud as they worked through the survey, describing any points of confusion along the way. Each session lasted approximately 30 minutes with the experimenter asking questions about the clarity of the survey options throughout the session. Confirming the findings from Study 1, the primary point of confusion across all three participants was questions asking for a comparison between the mobile app and the mobile website for a particular company. For example, Amazon can be accessed using a smart phone's mobile browser (e.g., Safari) or via the Amazon app which is downloaded to the phone. This difference was not immediately clear to participants. Even after the experimenter explained these two formats of mobile access, participants were not sure how to compare their experience on each platform. Therefore, questions asking participants to compare the mobile website to the mobile app were removed. One participant found the *strongly agree* and *strongly disagree* anchors confusing for questions that she wanted to answer with a simple yes or no. However, neither of the other two participants shared this concern. We therefore kept the anchors as *strongly agree* and *strongly disagree* but made a note of this concern for the analysis phase. Other comments concerning some awkward wording were considered and incorporated to the extent that they did not change the purpose of the item.

Participants were then recruited on Mechanical Turk, paid $.54, and spent on average 7 minutes completing the survey. There were three sampling groups that responded to the items. In Group 1, participants were presented with a list of 15 of the most popular apps as defined by the number of unique downloads in the last month (provided by the App Annie website, www.appannie.com) and asked to select the app they used the most. In Group 2, participants were asked to select an app they don't use very often but which is still on their phone. The Group 3 participants were asked to rate the app they used most recently.

Rasch Analysis

To evaluate the psychometric properties of the items that describe the mobile app user experience, a Rasch analysis (Rasch, 1960; Rasch, 1980) was employed. The Rasch model is considered a special case of Item Response Theory (IRT) and is seen as providing additional benefits over traditional methods—often called Classical Test Theory (CTT)—for creating standardized instruments. IRT and Rasch modeling are used increasingly in psychology and the social sciences to create standardized instruments (Bond & Fox, 2007; Wright & Stone, 2004) and even applied in the field of HCI (see Schmettow & Vietze, 2008 and Tezza, Bornia, & de Andrade, 2011 for examples).

Despite having been used for several decades in the creation of questionnaires and achievement tests, it's unlikely that most UX practitioners are familiar with both the theory and applications of IRT and Rasch modeling. We will use this section to contrast the major difference between CTT and IRT and highlight the advantages of using IRT.

Most questionnaires continue to be developed using Classical Test Theory, including the SUS (Brooke, 1996) and SUPR-Q (Sauro, 2016). From a high level, CTT tends to generate a set of items for a questionnaire that are optimized around the *average* level of a construct. For example, the 10 items in the SUS are highly correlated which results in high internal reliability (Lewis & Sauro, 2009). Items that tend to deviate from each other lead to lower reliability and get dropped. The result is essentially a redundant set of items that asks about the same concept in many subtly different ways. IRT, by contrast, optimizes around a questionnaire that reliably measures a fuller range of the construct, from low to high, not just around the average. As an example, a math exam using CTT would consist of a lot of items of average difficulty (e.g., Solve for x in the equation 2x = 10), whereas an exam developed using IRT would have easy items (e.g., What is 11 x 2?) to difficult items (What is the derivative of $2X^2$?). The analysis at the item level (hence the name Item Response Theory) allows for selecting the best items (reliable and valid) that best differentiate people's ability or attitudes toward a construct (Bond & Fox, 2007). Rasch modeling provides visual and numeric information about individual items (fit and difficulty). (The technical details of Rasch modeling are beyond the scope of this paper; for more information see Hambleton 1991.)

Criticisms of CTT include the treatment of ordinal raw scores as interval and sample dependence of person and item statistics. IRT accounts for some of the limitations of CTT. Because the calibrations used in IRT allow the measure to be generalized across samples (due to the theoretical independence of the items from the sample), the difficulty of the item and person ability are not confounded as they are in CTT (Bond & Fox, 2007). While CTT can serve as more of a "blunt" examination of a scale's psychometric viability, IRT allows detailed examination of individual items, person functioning on the scale, and relation of items to one another.

More practically, Rasch modeling provides additional information not included in CTT, including indices of model fit and visual and numeric information about individual items (fit and difficulty) (Hambleton, 1991). Rasch analysis allows researchers to evaluate the extent to which a unidimensional scale is created by the items in the measure. Dimensionality is assessed by determining whether each item meaningfully contributes to the measurement of a single construct. If data can be shown to fit the Rasch model, item and person estimates can be interpreted in terms of abstract, equal-interval units created by natural log transformations of raw data odds (Bond & Fox, 2007).

A Rasch analysis was used in the subsequent three studies with a polytomous rating scale model (Wright & Stone, 2004) using Winsteps 3.81 software (Linacre, 2012).

Results

There were 341 usable responses across the three groups. Respondents were from the US, were a mix of gender (54% male), 46% had a college-degree or higher, and the majority (82%) of the respondents were between the ages of 18 and 39.

Participants rated a total of 139 unique apps. The most commonly rated apps were Google Maps (28), Facebook Messenger (24), Snapchat (24), PayPal (20), Facebook (16), and Instagram (13).

Dimensionality

A fundamental requirement for using the Rasch model is that the items only measure one construct. This property is called unidimensionality. To assess the dimensionality of the candidate items, a technique called Principal Components Analysis (PCA) of residuals is used. It's a technical process that has some similarities to Factor Analysis (for example, see Christensen, Engelhard, & Salzberger, 2012). The PCA on the residuals was conducted and fit statistics were examined for the 34 items. There is poor evidence to support unidimensionality, as the first contrast eigenvalue of 6.2 is above the commonly acceptable threshold of 2-3 eigenvalues and higher than the recommended 5% (9.2%; Bond & Fox, 2007).

The raw variance explained by the items, 20.3%, is around twice the variance explained by the first contrast, 9.2%, so there is a noticeable secondary dimension in the items. The eigenvalue of the first contrast was 6.2; this indicates that it has the strength of about 6 items (6.2 rounded to 6, out of 34 items), bigger than the strength of two items (an eigenvalue of 2), the smallest amount that could be considered a "dimension."

Item Fit

The 34 items were examined to identify poorly fitting items. With adequate fit, easier to agree to items are endorsed by more people than are difficult to agree to items. Fit is assessed using mean square values (MNSQ), which are transformations of chi-square statistics, that have expected values of 1.0 if the data fit the Rasch model. Items and persons can underfit or overfit the model. Underfit occurs when there is too much noise in an estimate; overfit occurs when there is less than the model-expected noise in an estimate. Cut-offs for determining fitting items and persons from misfitting items and persons are flexible with Infit and Outfit MNSQ values between 0.5 and 1.5, which is considered “productive of measurement” (Lin, Liu, Sadeh, & Hong, 2014). Underfit or Infit/Outfit MNSQ values exceeding a cut-off (e.g., >1.4), occurs for items eliciting idiosyncratic responses or items that are less strongly related to the core measure. Underfit is generally considered a greater flaw than overfit. Overfit was not used as exclusion criteria in this study (Bond & Fox, 2007).

In the initial pass, six items were identified as having infit or outfit mean statistics greater than 1.4 (underfitting items) and were removed.

Table 2. Items and Infit/Outfit Mean Square Values

Item	Infit	Outfit	Label
NoCrash*	1.76	2.25	The app rarely crashes or causes problems on my phone.
WhenBored*	1.89	1.91	I use the app when I am bored.
LTR*	1.27	1.89	How likely are you to recommend this app to a friend?
PreInstalled*	1.70	1.74	The app should come pre-installed on phones.
NoBugs*	1.44	1.65	The app runs without bugs or errors.
Enhances*	1.56	1.55	This app enhances the other features of my mobile phone (e.g., a camera app that adds capabilities to your phone's camera).
CantLiveWo	1.39	1.37	I can't live without the app on my phone.
Addictive	1.25	1.35	This app is addictive.
UseAppOtherForms	1.22	1.19	I use this app even when there are other forms of entertainment or distraction around me.
Credible	1.04	1.16	The information on the app is credible.
TalkFriends	1.16	1.14	I talk about things I do or learn on the app with my friends.
NeverDelete	1.10	1.05	I would never delete the app.
Integrates	0.91	1.07	This app integrates well with the other features of my mobile phone.
Trustworthy	1.04	1.05	The information on the app is trustworthy.
ExtEasy	0.98	1.04	This app is extremely easy to use.
EasyNav	1.02	0.97	It is easy to navigate within the app.
AppBest	0.89	1.02	The app is the best app I've ever used.
UseFuture	1.00	0.80	I will likely use the app in the future.
EveryoneHave	0.97	0.96	Everyone should have the app.
Easy	0.94	0.92	The app is easy to use.
Clean	0.90	0.75	The app has a clean and simple presentation.
FindInfo	0.88	0.86	The design of this app makes it easy for me to find the information I'm looking for.
UseFreq	0.87	0.83	I like to use the app frequently.
DefFuture	0.87	0.78	I will definitely use this app many times in the future.
AppMeetsNeeds	0.87	0.83	The app's capabilities meet my requirements.
Exciting	0.77	0.86	It's exciting to use the app.
Discover	0.82	0.86	I like discovering new features on the app.
AppAttractive	0.67	0.70	I find the app to be attractive.
Happy	0.63	0.69	Using the app makes me happy.
Fun	0.68	0.66	It's fun using the app.

Item	Infit	Outfit	Label
ExtAttractive	0.57	0.56	I find the app to be extremely attractive.
Enjoy	0.51	0.54	I enjoy using the app.
Delightful	0.48	0.50	The app is delightful.
Lookforward	0.46	0.48	I look forward to using this app.

* Items were removed due to underfit.

The items were removed and the model was rerun until all poorly fitting items were removed (7 total). Next, the person-fit was examined to look for unusual response patterns. Respondents with outfit or infit mean squares values greater than 3 were removed (Bond & Fox, 2007). This removed 10 respondents (3% of the total sample). For example, one participant rated the Discover banking app low on likelihood to recommend but also high on it should be preinstalled and everyone should have it, suggesting possible inattentive responding.

After removing the poorly fitting items and persons, the dimensionality was reexamined with the remaining 27 items. There was still poor evidence to support unidimensionality as the unexplained variance in the first contrast was still above the threshold of 2–3 eigenvalues (5.3) and higher than the recommended 5% (8.4%) suggesting there was still a noticeable secondary dimension in the items.

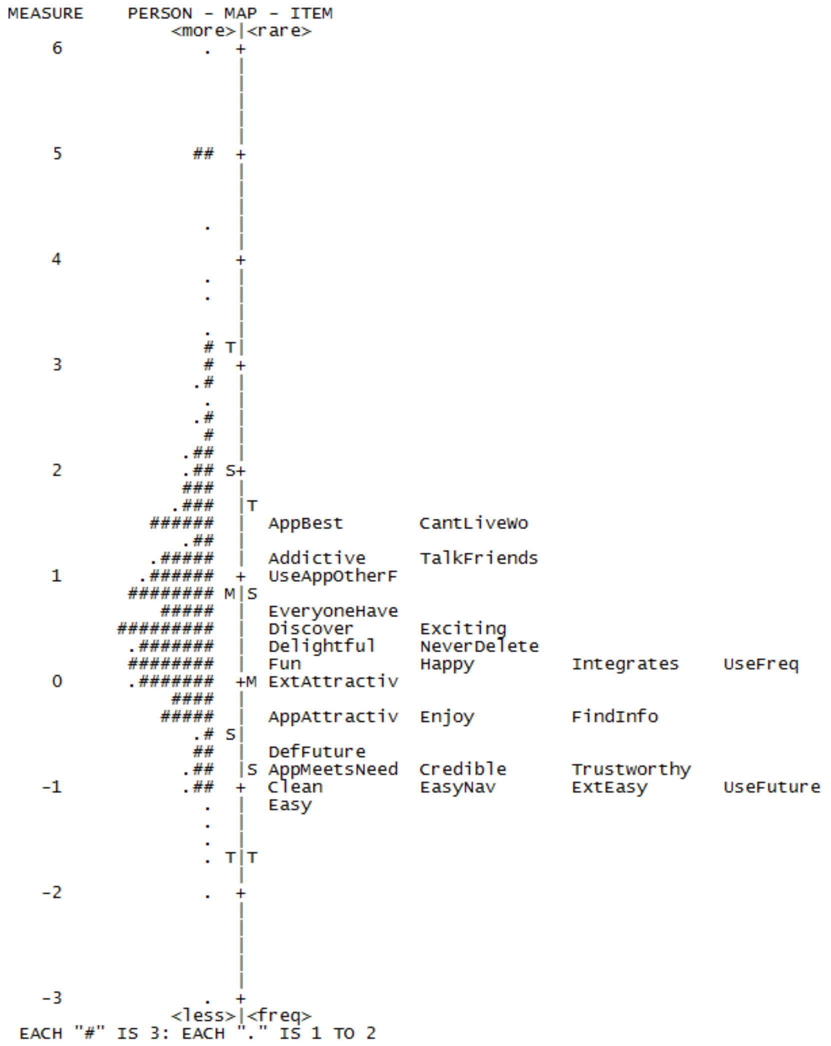
The 27 retained items are shown in Figure 1 in a visual display called an item-person map (also called a "Wright-Map" based on its creator Benjamin Wright, 2004). The item-person map places the difficulty of the items (how hard it was for respondents to agree with them) on the same measurement scale as the participants' ratings. It is another technique that Rasch analysis offers to help examine how well the items match the sentiments of the participants. Each "#" represents a participant on the left side and the label shows the item on the right side of the map.

The item-person map is organized as two vertical histograms with the items and respondents (persons) arranged from easiest on the bottom to most difficult on the top. For example, most participants agreed or strongly agreed (4s and 5s) to the items "Easy" (the app is easy to use) and "Clean" (the app has a clean and simple presentation). In contrast, fewer participants rated apps as "AppBest" (best app they've every used) as it is highest among the items.

On the left side, the item-person map shows the mean (M) and two standard deviation points (S = one SD and T = two SD) for measured participants tendency to agree. On the right side of the map, the mean difficulty of the items (M) and two standard deviation points (S = one SD and T = two SD) for the items are shown. Figure 1 shows that the mean (M) ability of the participants is approximately one standard deviation (S) above the mean (M) difficulty of the items.

The goal is to have the mean of the items and mean of the persons opposite each other on the map. The sample of respondents showed an above average score of .77 logits (where average is 0 logits). The items do a good job of covering the middle and lower scoring persons. For the respondents rating apps higher, however, there is poor coverage above 2 logits. This can be seen by all the # marks of participants scoring higher than any items. This suggests more difficult items are needed to differentiate at these higher levels.

There is redundancy in the middle levels of agreement, with several items (e.g., Fun, Happy, Integrates) having similar logit positions, meaning they discriminate at a similar level of app quality. While some redundancy is good for achieving reliability, when looking to reduce the burden on the respondent, items that discriminate at a similar level are good candidates for removal. The internal consistency reliability of the retained items was high (Cronbach alpha = .95; Nunnally, 1978).



```
MEASURE    PERSON - MAP - ITEM
                <more>|<rare>
    6              .  +
                      |
                      |
                      |
                      |
                      |
    5             ##  +
                      |
                      |
                      |
                   .  |
                      |
    4                 +
                   .  |
                   .  |
                      |
                   .  |
                   # T|
    3              #  +
                  .#  |
                   .  |
                  .#  |
                   #  |
                 .##  |
    2            .## S+
                 ###  |
                .###  |T
              ######  |  AppBest       CantLiveWo
                 .##  |
              .#####  |  Addictive     TalkFriends
    1        .######  +  UseAppOtherF
            ######## M|S
               #####  |  EveryoneHave
           #########  |  Discover      Exciting
            .#######  |  Delightful    NeverDelete
            ########  |  Fun           Happy         Integrates    UseFreq
    0       .#######  +M ExtAttractiv
                ####  |
               #####  |  AppAttractiv  Enjoy         FindInfo
                  .# S|
                  ##  |  DefFuture
                 .##  |S AppMeetsNeed  Credible      Trustworthy
   -1            .##  +  Clean         EasyNav       ExtEasy       UseFuture
                   .  |  Easy
                   .  |
                   .  |
                   . T|T
                      |
   -2              .  +
                      |
                      |
                      |
                      |
                      |
   -3              .  +
                <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 1. Item-person map.

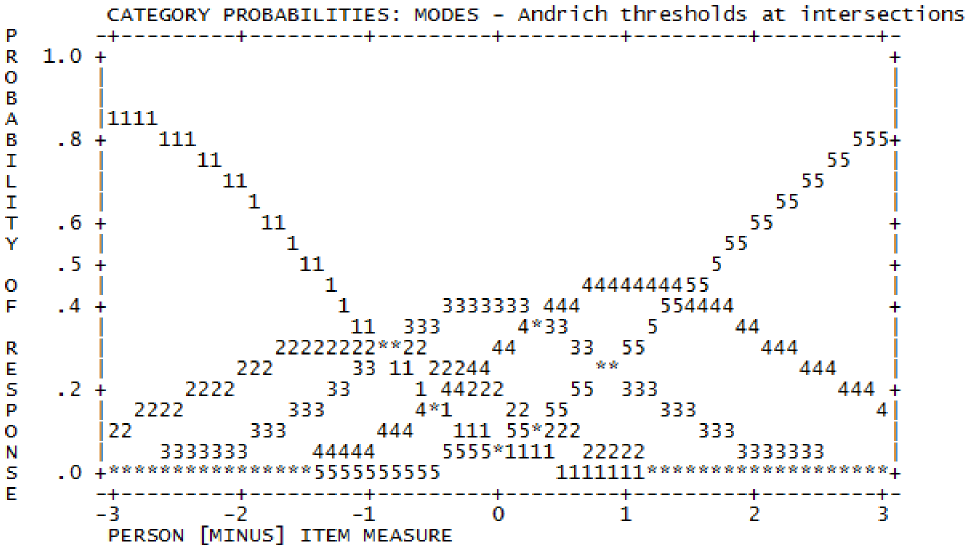
Scale Usage

The response scale was examined to ensure respondents were responding appropriately to the scale points. This can be seen by examining the Andrich Thresholds for each category, from 1 to 5. The category probability map displays values going from low to high as shown in Figure 2. The Andrich Thresholds also progress normally without any scale reversal. A scale reversal happens when respondents are, for example, using higher values (such as a 2 or 3) rather than a 1 when rating apps as low. The proper use of the scale can be seen in Figure 2 as progressing "waves" of responses (Bond & Fox, 2007). For more information on interpreting these graphs and Andrich Threshold values, see Linacre (2006).

However, there is some evidence that respondents were not endorsing all scale points equally. Point 2 was used less than expected. This can be seen in Figure 2 as respondents were more likely to pick 1 and then 3, not utilizing the 2 response option as much. The subsequent study will clarify the scale points by adding labels to all points. In this current study, only points 1 and 5 were labeled (*strongly disagree* and *strongly agree*, respectively). Table 3 presents a summary of how the categories are structured in Figure 2.

Table 3. Summary of Category Structure, Model = "R"

Category label	Score	Observed count	%	Observed average	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category measure
1	1	827	9	-1.12	-1.16	1.14	1.19	None	(-2.45)
2	2	959	11	-.42	-.46	1.00	.99	-.95	-1.10
3	3	1893	22	.17	.22	.95	.93	-.80	-.11
4	4	2507	29	.95	.97	.98	.90	.30	1.04
5	5	2562	29	2.11	2.08	1.01	1.00	1.45	(2.74)



```
        CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |1111                                                         |
B   .8 +    111                                                   555+
I      |       11                                               55   |
L      |         11                                           55     |
I      |           1                                        55       |
T   .6 +            11                                    55         +
Y      |              1                                 55           |
    .5 +               11                              5             +
O      |                 1                   4444444455              |
F   .4 +                  1       3333333 444      554444            +
       |                   11  333      4*33      5      44          |
R      |             22222222**22     44    33  55         444       |
E      |          222      33 11 22244        **              444    |
S   .2 +      2222       33     1 44222     55  333              444 +
P      |  2222        333       4*1    22 55       333              4|
O      |22         333       444   111 55*222         333            |
N      |    3333333     44444     5555*1111  22222       3333333     |
S   .0 +****************5555555555         1111111*******************+
E      -+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3
        PERSON [MINUS] ITEM MEASURE
```

Figure 2. Category probability curves for Study 2.

Discussion

Study 2 provided evidence for multiple dimensions with these data and item overlap, showing redundancy in the middle values. Items that were aimed at capturing users likelihood to delete an app ("The app runs without bugs or errors" and "The app rarely crashes or causes problems on my phone") were among the poorest fitting items. This may be due to a survivorship bias as participants were by definition rating apps still on their phone (and are most likely not crashing or buggy). Other poorly fitting items (e.g., Fun and Exciting) may not fit all app types well even though they fit the model. New items need to discriminate at the higher end of the scale. The scale use was acceptable but could be improved with labeling of all the values.

Study 3: Discriminating Between High Scoring Apps

A third study was conducted to assess more difficult to agree to items using a different app selection criteria and a fully labeled 5-point Likert scale.

Method

Participants were recruited again on Mechanical Turk, paid $.54, and spent on average 7 minutes completing a survey. There were two sampling groups that responded to the candidate items. In Group 1, participants were presented with a list of 15 popular apps and asked to select the app they used the most recently. If they did not use one of the listed apps, they were asked to identify and rate another app they used most recently. In the Group 2, participants were presented with the same group of 15 popular apps and asked to select an app that appeared on the “second screen” of their smartphone (indicating potentially lower usage) and rate it.

Participants responded to the same 34 items as in Study 2 and three additional items intended to discriminate between higher scoring apps:

- BestApp: <App name > is the best app I’ve ever used.
- All I Ever Want: <App name > has all the features and functions you could ever want.
- Can’t Imagine Better: I can’t imagine a better app than <app name>.

The same 5-point Likert response options were used; however, all points were labeled (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*).

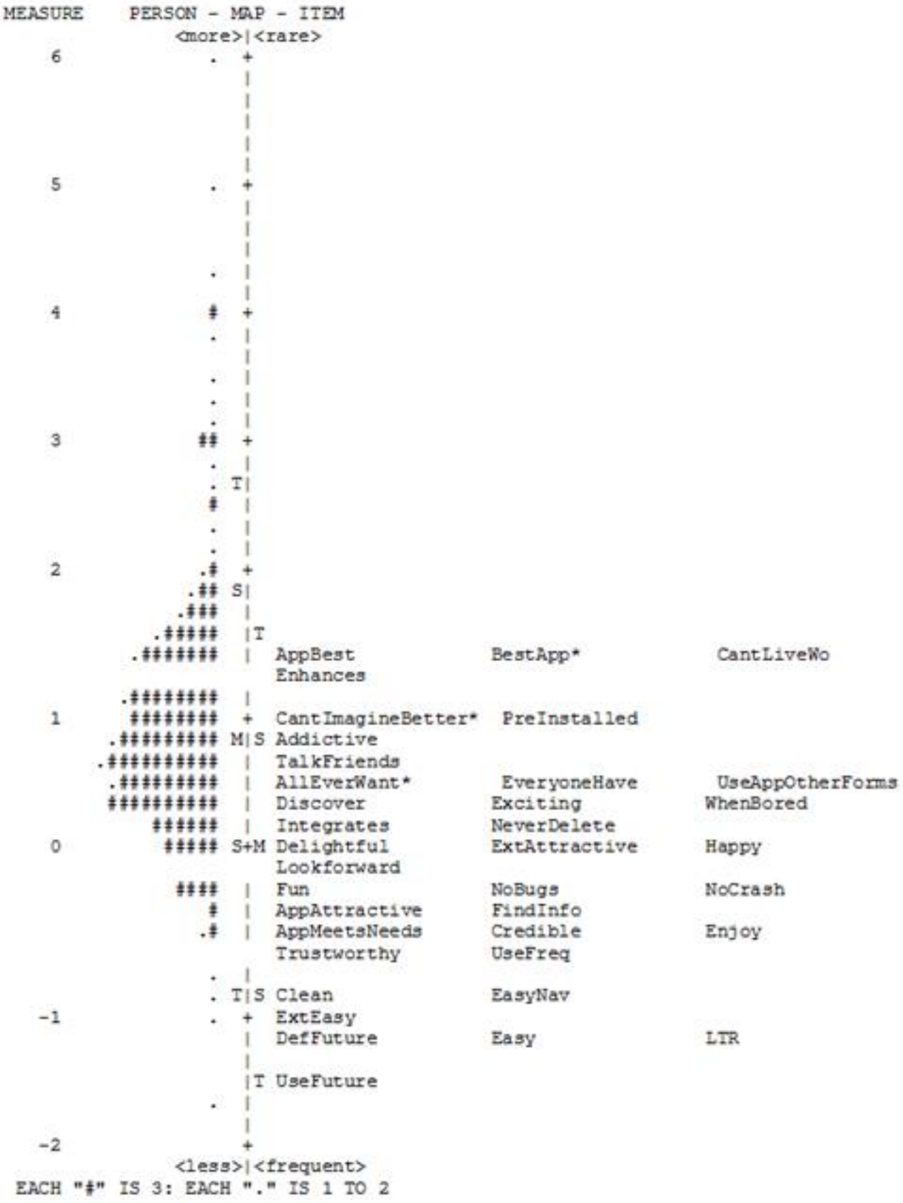
Results

There were 318 usable responses from both groups with respondents from around the US. The respondents were a mix of gender (51% male), 49% had a college-degree or higher, most (74%) were between the ages of 18 and 39, and had a mix of experience with iOS (40%) and Android (60%) device types. Participants rated a total of 55 unique apps. The most commonly rated apps were YouTube (48), Instagram (25), Pandora (23), Twitter (21), Skype (15) and Candy Crush (12).

Item and Person Fit

All items from Study 2 and the three candidate items are shown in the item-person map in Figure 3. The new items intended to discriminate at the highest level tended to perform similarly to the existing items. For example, the item “CantImagineBetter” has a similar logit position as “PreInstalled.”

The high rating experiences are still not well matched by these items with responses scored above 2 logit having no coverage and some items below -1 logits where no participants rated an app (Figure 3). The majority of the app experience is, however, covered well by the items. The average is .91 logits for this sample, slightly higher than the .71 for Study 2.



```
MEASURE    PERSON - MAP - ITEM
                 <more>|<rare>
    6                 .  +
                         |
                         |
                         |
                         |
                         |
    5                 .  +
                         |
                         |
                         |
                      .  |
                         |
    4                 #  +
                      .  |
                         |
                      .  |
                      .  |
                      .  |
    3                ##  +
                      .  |
                      . T|
                      #  |
                      .  |
                      .  |
    2                .#  +
                    .## S|
                   .###  |
                 .#####  |T
               .#######  |  AppBest            BestApp*           CantLiveWo
                            Enhances
              .########  |
    1          ########  +  CantImagineBetter*  PreInstalled
             .######### M|S Addictive
            .##########  |  TalkFriends
             .#########  |  AllEverWant*        EveryoneHave       UseAppOtherForms
             ##########  |  Discover           Exciting           WhenBored
                 ######  |  Integrates         NeverDelete
    0             ##### S+M Delightful         ExtAttractive      Happy
                            Lookforward
                   ####  |  Fun                NoBugs             NoCrash
                      #  |  AppAttractive      FindInfo
                     .#  |  AppMeetsNeeds      Credible           Enjoy
                            Trustworthy        UseFreq
                      .  |
                      . T|S Clean              EasyNav
   -1                 .  +  ExtEasy
                         |  DefFuture          Easy               LTR
                         |
                         |T UseFuture
                      .  |
                         |
   -2                    +
                 <less>|<frequent>
  EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 3. Item-person map for Study 3.

There is clear redundancy with many items again at the middle and lower end with similar logit positions (similar vertical positions in Figure 3). The 37 items were examined to identify poorly fitting items. The same six items identified as having infit or outfit mean statistics greater than 1.4 in Study 2 were also identified as poorly fitting (Bond & Fox, 2007) and removed. The hardest item for respondents to agree to, "CantLiveWithout," was the poorest fitting item (infit of 1.46 and outfit of 1.55) but it discriminated at the higher end and was retained. Items that were less relevant for productivity apps (e.g., Banking Apps), which also had similar logits positions, were removed (Fun, Exciting, Enjoy). Seven persons were removed for responses not fitting the model well (infit or outfit MNSQ values greater than 3). This iterative process retained 16 items.

Dimensionality

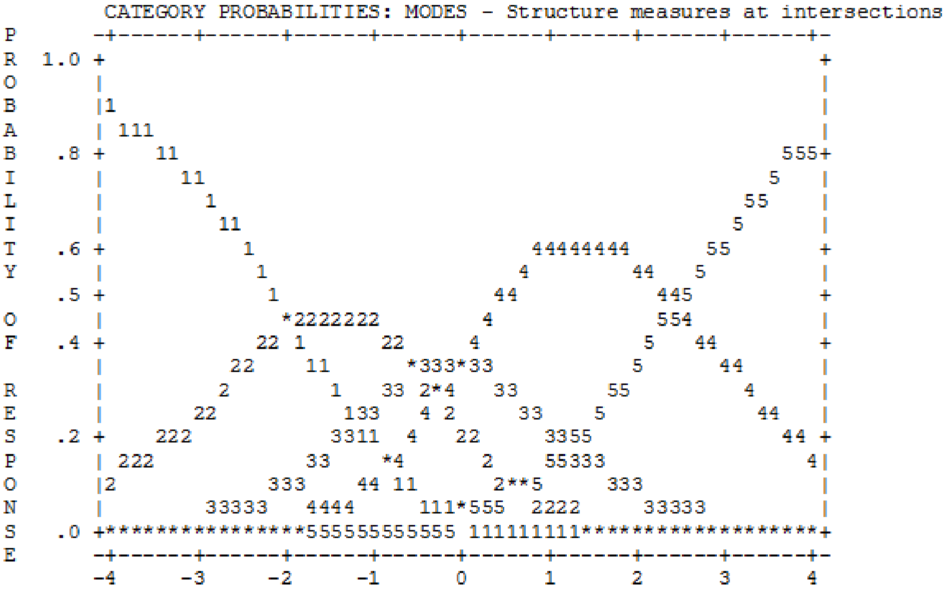
To assess the dimensionality of the reduced set of items, a principal components analysis of residuals was conducted and fit statistics were examined for the 16 items. There is some evidence to support unidimensionality, as the variance explained by measures at 52% (above the recommended 40%), and the unexplained variance in the first contrast is 2.8 eigenvalues. The internal consistency reliability of the retained items was also reasonably high (Cronbach alpha = .88; Nunnally, 1978).

In re-examining the scale usage after labeling all categories, the category probability map displays values going from low to high as shown in Figure 4. The Andrich Thresholds also progress normally without any scale reversal. This can also be seen in the figure of progressing "waves" of responses. There is evidence that labeling all points in the response options improved the response pattern. Point 3 is scoring close to the middle value (-.18 logits, 0 is the middle). The item-person map is shown in Figure 5; Table 4 gives a summary of the category probabilites in Figure 4.

Table 4. Summary of Category Structure, Model = "R"

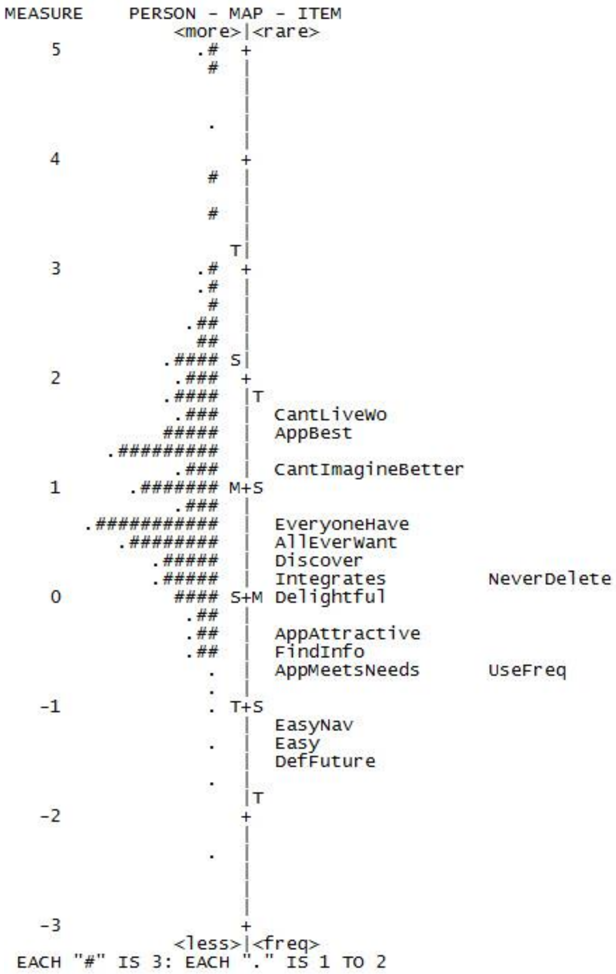
Category label	Score	Observed count	%	Observed average[a]	Sample expect	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category measure
1	1	184	4	-.97	-1.17	1.28	1.41	None	(-3.18)
2	2	574	12	-.41	-.42	1.04	1.14	-1.92	-1.43
3	3	933	19	.29	.37	.87	.91	-.52	-.18
4	4	2102	42	1.26	1.27	.93	.89	-.01	1.36
5	5	1167	24	2.47	2.42	.99	.99	2.45	(3.61)

[a] Observed average is a mean of measures in category. It is not a parameter estimate.



```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+------+------+------+------+------+------+------+------+-
R  1.0 +                                                         +
O      |                                                         |
B      |1                                                        |
A      | 111                                                     |
B   .8 +    11                                                555+
I      |      11                                             5   |
L      |        1                                          55    |
I      |         11                                       5      |
T   .6 +           1                      44444444      55       +
Y      |            1                    4        44   5         |
    .5 +             1                 44           445          +
O      |              *2222222        4             554          |
F   .4 +            22 1      22     4             5   44        +
       |          22    11      *333*33           5      44      |
R      |         2        1   33 2*4   33       55         4     |
E      |       22          133   4 2     33    5            44   |
S   .2 +    222           3311  4   22     3355               44 +
P      | 222            33    *4      2    55333                4|
O      |2            333    44 11      2**5     333              |
N      |        33333   4444     111*555  2222     33333         |
S   .0 +****************555555555555 111111111*******************+
E      -+------+------+------+------+------+------+------+------+-
       -4     -3     -2     -1      0      1      2      3      4
```

Figure 4. Category probability curves for scale usage.



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5            .#  +
                  #  |
                     |
                     |
                  .  |
                     |
    4                +
                  #  |
                     |
                  #  |
                     |
                    T|
    3            .#  +
                 .#  |
                  #  |
                .##  |
                 ##  |
              .#### S|
    2          .###  +
              .####  |T
               .###  |  CantLiveWo
              #####  |  AppBest
          .#########  |
               .###  |  CantImagineBetter
    1      .####### M+S
               .###  |
       .###########  |  EveryoneHave
          .########  |  AllEverWant
             .#####  |  Discover
             .#####  |  Integrates         NeverDelete
    0          #### S+M Delightful
                .##  |
                .##  |  AppAttractive
                .##  |  FindInfo
                  .  |  AppMeetsNeeds      UseFreq
                  .  |
   -1             . T+S
                     |  EasyNav
                  .  |  Easy
                     |  DefFuture
                  .  |
                     |T
   -2                +
                     |
                  .  |
                     |
                     |
                     |
   -3                +
               <less>|<freq>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 5. Updated item-person map for Study 3.

Discussion

Parsimony is an important practical consideration of a valid instrument. The retained items removed the redundancy while still maintaining both reliability and the ability to discriminate at different levels of the measure. There is reasonable evidence for unidimensionality. Items targeted for deletion were again poorly fitting and removed, as well as items that did not seem to apply to all apps (e.g., Fun and Exciting), even though they fit the model. The scale use improved with the addition of the value labels.

Study 4: Convergent and Predictive Validity

The purpose of Study 4 was to collect additional data for more apps with the same items, reassess the internal consistency reliability, assess the convergent validity with existing instruments, and assess predictive validity with published app ratings and reviews.

Method

Participants were recruited again on Mechanical Turk, paid $.59, and spent on average 9 minutes completing a survey. There were three sampling groups that responded to the candidate items. All groups were presented with a list of 15 popular apps. Group 1 was asked which app they had deleted from their phone in the last year. In Group 2 participants were asked to select an app they hadn’t used recently but is still on their phone (indicating potentially lower usage). In Group 3 participants were asked to identify an app they used most recently.

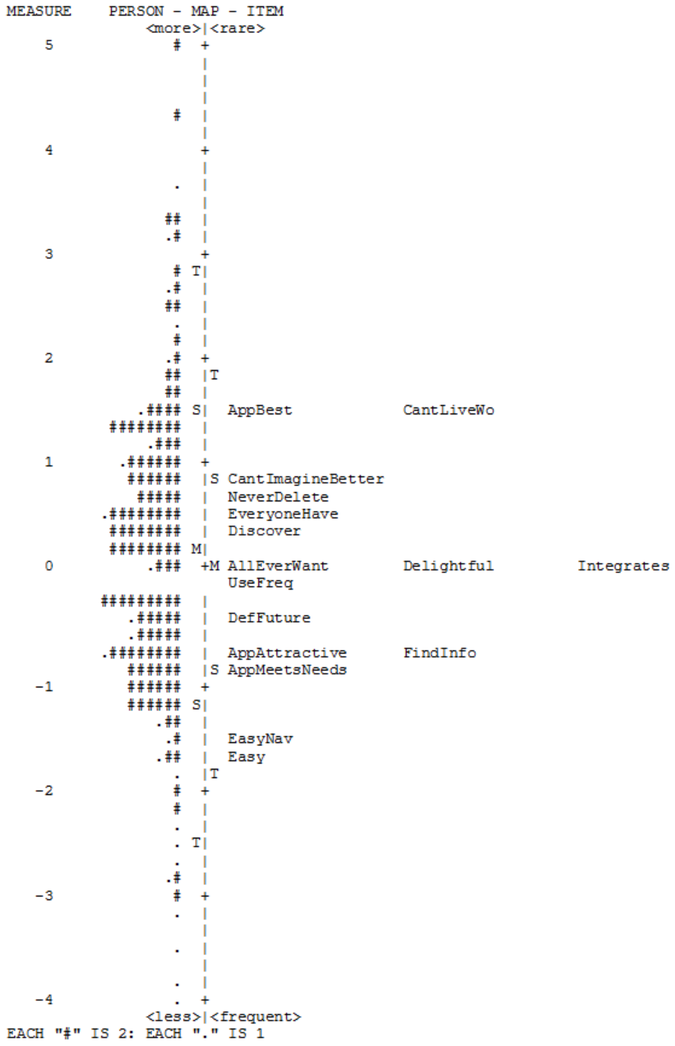
Participants responded to the same 16 items as identified in Study 3 along with items from the SUPR-Q (Sauro, 2016), LTR (basis for the Net Promoter Score; Reichheld, 2003), and a 5-point version of the UMUX-Lite (Lewis et al., 2013).

Results

There were 284 usable responses from the three groups with respondents from around the US. The respondents were a mix of gender (48% male), 46% had a college-degree or higher, and most (79%) were between the ages of 18 and 39, and had a mix of experience with iOS (40%) and Android (60%) device types. Participants rated a total of 44 unique apps. The most commonly rated apps were Spotify (25), Amazon (23), My Fitness Pal (23), Pinterest (23), Gmail (20), and Pokémon Go (16).

Item and Person Fit

Three respondents were removed that had MNSQ values above 3. All 16 items from Study 4 fit the Rasch model well, with infit/outfit MNSQ values below 1.4. There was good coverage from low to high as shown in the item-person map in Figure 6.



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5             #  +
                     |
                     |
                     |
                  #  |
                     |
    4                +
                     |
                  .  |
                     |
                 ##  |
                 .#  |
    3                +
                  # T|
                 .#  |
                 ##  |
                  .  |
                  #  |
    2            .#  +
                 ##  |T
                 ##  |
              .#### S|  AppBest            CantLiveWo
           ########  |
               .###  |
    1       .######  +
             ######  |S CantImagineBetter
              #####  |  NeverDelete
          .########  |  EveryoneHave
           ########  |  Discover
           ######## M|
    0          .###  +M AllEverWant        Delightful         Integrates
                        UseFreq
          #########  |
             .#####  |  DefFuture
             .#####  |
          .########  |  AppAttractive      FindInfo
             ######  |S AppMeetsNeeds
   -1        ######  +
             ###### S|
                .##  |
                 .#  |  EasyNav
                .##  |  Easy
                  .  |T
   -2             #  +
                  #  |
                  .  |
                  . T|
                  .  |
                 .#  |
   -3             #  +
                  .  |
                     |
                  .  |
                     |
                  .  |
   -4             .  +
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 6. Item-person map for Study 4.

Dimensionality and Scale Usage

To assess the dimensionality of the reduced set of items, a principal components analysis of residuals was conducted and fit statistics were examined for the 16 items. There is evidence to support unidimensionality, as the variance explained by measures at 61% (above the recommended 40%), and the unexplained variance in the first contrast is 2.9 eigenvalues. The internal consistency reliability of the retained items was high (Cronbach alpha = .94; Nunnally, 1978). The scale usage performed well and similar to Study 3, with the category probability map displaying values going from low to high. The person-separation of 3.34 indicates a good spread of items and persons (Bond & Fox, 2007).

Convergent and Predictive Validity

To assess convergent validity, the raw scores on the 16 items were aggregated across Studies 2, 3, and 4 (n = 914). In all three studies, items from the UMUX-Lite, SUPR-Q, and Likelihood to Recommend (LTR) were collected. The correlation was strong with the UMUX-Lite (r = .74), SUPR-Q (r = .71), and the single item LTR question (r = .74). All correlations were statistically significant at p < .01, providing good evidence for convergent validity.

To assess predictive validity, the raw scores of the 16 items were aggregated across Studies 2, 3, and 4 (n = 914) and averaged by specific apps. This provided 29 apps with 10 or more responses (Table 5). These scores were correlated with the average star rating for all versions of the apps (US only) and the total number of app reviews as provided by the website App Annie. This is a similar process as used by Nayebi, in which a machine learning algorithm was able to predict Apple's App Store ratings using the compliance to Apple's Human Interface Guidelines (2015).

The average App Store and Google Play Stores ratings on the items were weighted to match the sample (60% Android and 40% iOS). The correlations were low and not statistically significant between both the unweighted and weighted ratings (r = .13, p > .2). The correlation with the weighted total number of reviews by app was, however, statistically significant (r = .38, p = .02). It was also statistically significant for the unweighted number of Android app reviews from the Play Store (r = .38, p < .02) but not with the App Store number of ratings (r = .28, p = .14). While the total number of downloads is a good outcome measure (more downloads provides a good indicator of the success of an app), such data is not publicly available. Instead, the number of reviews written was used as a proxy for number of downloads.

Table 5. 29 Apps with 10+ Ratings Correlated with Downloads and Ratings

AppName	**Num**	**SUPR-Qm**	**Weighted Download #**	**App Store**	**Play Store**
YouTube	53	3.77	7327112	2.5	4.2
My Fitness Pal	41	3.57	1025064	4.5	4.6
Instagram	38	3.81	23046027	4.5	4.5
Pandora	38	3.81	1943050	4	4.4
Pinterest	36	3.36	1710202	4.5	4.6
Pokémon Go	33	3.00	3510841	3	4.1
Skype	29	3.11	5560820	3.5	4.1
Twitter	28	3.37	5022398	3.5	4.2
Facebook	28	3.83	30471735	3.5	4
Amazon	28	3.57	313464	3.5	4.3
Spotify	27	3.33	4029856	4.5	4.5
Google Maps	27	3.90	3935981	4.5	4.3
PayPal	26	3.62	302050	4	4.3
Snapchat	26	3.66	5130828	2.5	3.9
Hangouts	26	3.33	1261693	4	3.9
eBay	25	3.40	1178836	4	4.2
Messenger	25	3.72	18665806	3	3.9

AppName	Num	SUPR-Qm	Weighted Download #	App Store	Play Store
Gmail	23	3.79	1803610	4	4.3
Tinder	23	3.22	1121487	3.5	4
Duolingo	22	3.73	2015480	5	4.7
Bank of Amer.	21	3.42	218223	3.5	4.2
Groupon	21	3.08	789571	4.5	4.5
Candy Crush	20	3.54	9847789	4	4.3
Clash of Clans	18	3.47	18095767	4.5	4.6
Wells Fargo	18	3.43	113510	3	4.3
Waze	16	3.60	3286792	4.5	4.6
YELP	12	3.20	262826	3.5	4.3
Chase Bank	11	3.78	484106	4.5	4.6
Uber	10	2.97	656939	4	4.3

General Discussion

This research has identified a core set of 16 items that were found to be unidimensional, reliable, valid, and discriminated well across a number of app types and difficulty levels.

Reflecting its similar purpose to the website-focused SUPR-Q, we have named the questionnaire that uses the final 16 (shown in Table 6) items the Standardized User Experience Percentile Rank Questionnaire for Mobile Apps (SUPR-Qm).

Table 6. App Logit Position

Item	Logit Position	Full Item Wording
CantLiveWo	1.55	I can't live without the app on my phone.
AppBest	1.50	The app is the best app I've ever used.
CantImagineBetter	0.88	I can’t imagine a better app than this one.
NeverDelete	0.70	I would never delete the app.
EveryoneHave	0.50	Everyone should have the app.
Discover	0.32	I like discovering new features on the app.
AllEverWant	0.05	The app has all the features and functions you could ever want.
UseFreq	0.03	I like to use the app frequently.
Delightful	-0.04	The app is delightful.
Integrates	-0.04	This app integrates well with the other features of my mobile phone.
DefFuture	-0.30	I will definitely use this app many times in the future.
FindInfo	-0.59	The design of this app makes it easy for me to find the information I'm looking for.
AppAttractive	-0.71	I find the app to be attractive.
AppMeetsNeeds	-0.79	The app’s capabilities meet my requirements.
EasyNav	-1.44	It is easy to navigate within the app.
Easy	-1.63	The app is easy to use.

The items in Table 6 are ordered by difficulty (high to low). For example, respondents readily endorsed the ease of use items (Easy, EasyNav), suggesting that usability is necessary but not sufficient for a good user experience. In contrast, "CantLiveWo" and "AppBest" were the most difficult items to endorse. Even these rather extreme sentiments did not fully capture the highest rated apps, suggesting participants feel quite strongly about the importance and need for some apps (at least in the context of mobile usage).

The items intended to identify the common reasons for deletion (bugs and crashing) did not fit the model well. This could be due to a survivorship bias (participants rate apps they generally still have) as well as the particular apps rated for this study, which tended to be among the most popular. The average of the SUPR-Qm items correlated modestly with app ratings but still exceeded the common acceptable threshold of r = .3 (Nunnally, 1978).

The items can be administered to participants after using an app in a usability study or part of a retrospective survey. Present the items using a 5-point scale (1 = *strongly disagree* and 5 = *strongly agree*). To score the app, researchers can take the average of the items and use the average for benchmarking against future changes and compare these scores to the ones presented in this paper (mean values shown in Table 5).

Limitations and Future Research

This analysis identified a core set of items that span all app categories. The correlation with the number of app reviews, while statistically significant, shows that only around 10% of the variation in the number of reviews (and by a rough proxy the number of downloads) can be explained by this measure. There are likely two reasons for this lower correlation. First, app category specific issues likely account for much of an app's success. For example, the success of a game may have more to do with its ability to entertain and engage rather than accomplish tasks. Second, while the 29 apps included a range of type and quality, most were already very successful apps with millions of downloads. It's possible this set has a restriction of range that is attenuating the correlation. Complicating matters are the fickle nature and non-representativeness of app reviews. By number of reviews alone, many of these apps are quite successful, yet still had very low ratings (e.g., YouTube has a rating of 2.5 for App Store users despite millions of downloads).

Future research should investigate a broader range of app quality. Ideally, the research should correlate to the actual number of downloads and should extend beyond the U.S. market to assess the relationship between the number of downloads and an app's ratings. These results also suggest that there are many additional variables that underlie the mobile app user experience. There were a number of items that had similar logit positions and fit, meaning there may be multiple items to continue to assess in the future. The average app ratings and number of ratings are a crude proxy for both the number of downloads and success of the app user experience. Additional research can test more apps to assess the continued correlation with the number of app reviews. For example, future analysis may involve a closer examination of the items not included in the final measure as well as a good set of secondary items to see how responses to these items may differ by app type (e.g., productivity and gaming).

Finally, one of the advantages of using IRT and having items with known logit positions is the ability to maintain an "item-bank" and deliver items programmatically using an adaptive questionnaire. An adaptive questionnaire should reduce the number of items needed to obtain a stable estimate of the apps' user experience. Future analyses can explore how to administer the SUPR-Qm using an adaptive method and explore any advantages it may provide compared to a traditional administration where participants receive all items.

Conclusion

The user experience of mobile applications impacts the likelihood of deletion and future use. In this paper, we identified items that described the quality of the mobile application user experience that applied to all app categories. Rasch analysis was used to assess the psychometric properties of items collected from four independent surveys (N = 1,046) with ratings on 174 unique apps. Sixteen items were identified that fit the model well and comprise a new instrument called the SUPR-Qm. For the final instrument, estimates of internal consistency reliability were high (alpha = .94), convergent validity was also high, with significant correlations with the SUPR-Q (.71), UMUX-Lite (.74), and likelihood-to-recommend (LTR) scores (.74). Scores on the SUPR-Qm correlated with the number of app reviews in the Google Play Store and Apple's App Store (r = .38), establishing adequate predictive validity. The SUPR-Qm can be used to benchmark the user experience of mobile applications which is an essential step in understanding what works and what needs to be improved in the rapidly growing mobile app market.

Recommendations

We anticipate the future research on this topic will include the following:

- Data on additional mobile apps will be collected to continue to validate both the items and build a larger database of scores.
- An analysis on factors that affect SUPR-Qm scores (called a DIF analysis) is recommended. For example, it could be that certain mobile app users (iOS vs. Android) respond differently to different items.
- Items specific to app genres will be evaluated, for example, items that measure the more specific aspects of the gaming, social media, or productivity user experiences.
- Algorithms for using the logit positions for these 16 items in a calibrated item bank can be used for presenting the items in an adaptive questionnaire to reduce the number of items needing to be presented.

Tips for Usability Practitioners

When evaluating the mobile experience, consider the following:

- When assessing the quality of the mobile app user experience, practitioners should consider using a standardized questionnaire.
- Standardized questionnaires provide a more reliable and valid measure of the construct of interest compared to homegrown questionnaires.
- Standardized questionnaires, like the SUPR-Qm, can be administered during a usability test or outside of a usability test to have participants reflect on their most recent experience. The SUPR-Qm can be coupled with questions specific to a domain (e.g., a gaming experience or a hotel app experience).
- The advantage of the items selected for the SUPR-Qm is that they measure a wider range of the user experience by providing reliable estimates from items that are both easy and hard for participants to agree to.
- The SUPR-Qm can be administered before and after app changes to determine how much improvement, if any, was achieved.
- When measuring mobile apps, practitioners should consider a standardized questionnaire that measures more than just usability (such as the SUS does). While usability is a critical component of the mobile app user experience, it does not fully cover the spectrum of what differentiates excellent apps from poor ones.

Acknowledgements

We would like to thank Kathy Green, PhD and Jim Lewis, PhD for commenting on earlier drafts of this manuscript.

References

Bond, T. G., & Fox, C. M. (2007*). Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed). Mahwah, NJ: Erlbaum, Associates Inc.

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

Christensen, K., Engelhard, G., & Salzberger, T. (2012). Ask the experts: Rasch vs. factor analysis. *Rasch Measurement Transactions, 26*(3), 1373–1378.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Dogtiev, A. (2015, November 16). App Revenue Statistics 2015, Business of Apps. Retrieved September 20, 2016, from http://www.businessofapps.com/app-revenue-statistics/2015/

Fakhruddin, H. (January 30, 2016). Top 12 reasons why users frequently uninstall mobile apps. Retrieved September 20, 2016, from https://www.linkedin.com/pulse/top-12-reasons-why-users-frequently-uninstall-mobile-apps-fakhruddin

Hambleton, R. (1991). *Fundamentals of item response theory (measurement methods for the social science)*. Newbury Park, CA: SAGE Publications, Inc.

Hassenzahl, M., & Tractinsky, N. (2006). User experience - A research agenda. *Behaviour & Information Technology, 25*(2), 91–97.

ISO 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs), part 11, Guidance on Usability. Retrieved September 20, 2016 from https://www.iso.org/standard/16883.html

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human–Computer Interaction, 31*(8), 518–529.

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (1267–1312). Hoboken, NJ: John Wiley & Sons, Inc.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102). New York, NY: ACM.

Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), *Human Centered Design*, HCII 2009 (pp. 94–103). Heidelberg, Germany: Springer-Verlag.

Lin, J., Liu, B., Sadeh, N., & Hong, J. I. (2014, July). Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In Symposium on Usable Privacy and Security (pp. 199–212). Menlo Park, CA: Usenix.

Linacre, J. M. (2006). Demarcating category intervals: Where are the category boundaries on the latent variable? *Rasch Measurement Transactions, 19*(3), 10341–10343. Available at https://www.rasch.org/rmt/rmt194f.htm

Linacre J. M. (2012). *A user's guide to Winsteps ministep 3.70.0: Rasch model computer programs*. Chicago, IL: Winsteps.

Meeker, M. (2015, May). *Internet Trends 2015*. In Code Conference. Rancho Palos Verdes, CA: Recode.

Nayebi, F. (2015). iOS application user rating prediction using usability evaluation and machine learning. (Doctoral dissertation). Retrieved from Research Gate doi: 10.13140/RG.2.1.3217.9681

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Malley, G., Dowdall, G., Burls, A., Perry, I. J., & Curran, N. (2014). Exploring the usability of a mobile app for adolescent obesity management. *JMIR Mhealth and Uhealth, 2*(2), e29.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded ed.). Chicago, IL: University of Chicago.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–54.

Ryu, Y. S., & Smith-Jackson, T. L. (2005, July). Development of usability questionnaire items for mobile products and content validity. In *Proceedings of HCI International* (pp. 22–27). Las Vegas, NV: HCI.

Ryu, Y. S., & Smith-Jackson, T. L. (2006). Reliability and validity of the mobile phone usability questionnaire (MPUQ). *Journal of Usability Studies,2*(1), 39–53.

Sauro, J. (2016). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies, 10*(2), 68–86.

Schmettow, M. & Vietze, W. (2008, April). Introducing item response theory for measuring usability inspection processes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 893-902). Florence, Italy: ACM.

Tezza, R., Bornia, A. C., & de Andrade, D. F. (2011). Measuring web usability using item response theory: Principles, features and opportunities. *Interacting With Computers, 23*(2), 167–175.

Tullis, T., & Albert, B. (2013). Measuring the user experience: Collecting, analyzing, and presenting usability metrics (2nd ed.). Boston, MA: Elsevier Inc.

Varshneya, R. (June 4, 2015). 7 reasons why users delete your mobile app. Retrieved September 20, 2016, from http://www.inc.com/rahul-varshneya/7-reasons-why-users-delete-your-mobile-app.html .

Wright, B. D., & Stone, M. H. (2004). *Making measures*. Chicago, IL: The Phaneron Press.

About the Authors

Jeff Sauro
Dr. Sauro is the founding principal of MeasuringU, a UX research firm based in Denver. He has published over 25 peer-reviewed research articles and 5 books, including *Quantifying the User Experience* and *Customer Analytics for Dummies*.

Paree Zarolia
Dr. Zarolia is a cognitive psychologist and user experience researcher. She has conducted research in decision-making, consumer choice, and user experience. She is currently a User Experience Researcher at Google.