JUX Journal of User Experience

Vol. 19, Issue 2, February 2024 pp. 75-103

Generalized User Experience Questionnaire (UEQ-G): Holistic Measurement of Multimodal UX

Chase S. Boothe
Sr. Director of Product Research
BeyondTrust Corporation
775 Woodlands Parkway
Suite 200
Ridgeland, MS 39157
cboothe@beyondtrust.com

Lesley Strawderman
Professor
Department of Industrial and Systems Engineering
Mississippi State University

Reuben F. Burch V
Associate Professor
Department of Industrial and Systems Engineering
Mississippi State University

Brian K. Smith
Associate Professor
Department of Industrial and Systems Engineering
Mississippi State University

Cindy L. Bethel
Professor
Department of Computer Science and Engineering
Mississippi State University

Kate Holmes
Lead UX Researcher
BeyondTrust Corporation

Abstract

The User Experience Questionnaire (UEQ) is a commonly used tool for measuring product experience. This study covers extending the UEQ to measure multimodal experiences that include both product and service experiences. Currently, no questionnaires measure holistic user experiences, including pragmatic and hedonic qualities, for both product and service experiences. Through three study phases, we created and tested the Generalized User Experience Questionnaire (UEQ-G). First, we generalized and tested language from the UEQ's original, product experience context. Second, the UEQ-G was applied to controlled service experiences in which conditions were artificially manipulated across traditional UEQ factors. Third, we applied the UEQ-G in the field to experiences that contained both product and service experiences within the same scenario.

No significant differences were observed between the UEQ and UEQ-G during the first phase, but the UEQ-G detected differences between high and low conditions for all expected factors except one during the second phase. During the third phase, many expected correlations were found among UEQ-G factors and those from other well-established tools; however, a few expected correlations were not observed. This study found the UEQ-G to be as valid and reliable as its predecessor, UEQ, in product experience scenarios, and although additional study is required, the UEQ-G showed great potential in evaluating service experience scenarios and for evaluating multimodal experiences in the field. With additional study, the UEQ-G tool could be the first tool of its type for assessing holistic user experience across various multimodal experiences.

Keywords

User Experience Questionnaire, UEQ, UEQ-G, experience evaluation, holistic experience, multimodal experience

Introduction

Purpose

In an age in which users' expectations of their digital and non-digital experiences are paramount, providing top-notch experiences is important to retain current customers, capture competitors' unsatisfied customers, and maximize employee morale and productivity; not to mention, it is a way to make the world a little more pleasant. Whether the experience is digital, non-digital, or multimodal, a thoughtful and effective user experience is a basic requirement for any user-centric design. But how can improvements be made without effective measurement?

Defining UX

ISO 9241-210 defines user experience as the "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" and specifically identifies brand image, presentation, functionality, system performance, interactive behavior, and assistive capabilities as factors determining UX (ISO 9241-210:2019, 2019).

ISO 9241-210 goes on to define user interface (UI) as "all components of an interactive system (software or hardware) that provide information and controls for the user to accomplish specific tasks with the interactive system" (ISO 9241-210:2019, 2019). Because the ISO definition of UX specifically identifies interactive behavior—a factor that seemingly subsumes the entire UI definition—as just one of many UX factors, it is reasonable to conclude that a UI is but one factor in determining the resulting UX.

Another important distinction between UX and a UI is that while a UI can be directly designed, a UX cannot. UX is rather the resulting perception a user has based on a series of interactions with a system, product, or service, and it is only those points of interaction that can be designed (Rogers et al., 2011-b). Therefore, the term "user experience design" can be misleading as it really describes the process of designing those points of interaction that impact the experience rather than the experience itself.

Confusion can be further exacerbated by the often-misunderstood relationship between UX and usability. Usability is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-210:2019, 2019). Usability only accounts for a portion of the factors that determine a resulting UX. However, unlike a UI, which can also be a factor in determining a resulting UX, it is virtually impossible to have a UX if usability is absent. Usability is "fundamental" to UX, and UX is "inextricably linked" to usability; therefore, the two should be evaluated in concert (Rogers et al., 2011-b). Those elements included in UX that are not related to usability are captured within the concepts of hedonic quality (Laugwitz et al., 2008).

Measuring UX

UX can be measured using various techniques such as interviews, focus groups, questionnaires, direct observation in the field, direct observation in a controlled environment, and indirect observation (Rogers et al., 2011-a). This study focuses on the single technique of standardized questionnaires, which have the benefits of being widely distributable and highly consistent. These benefits grant the ability to collect a large number of responses and compare results from one test to those of other user experiences, to benchmarks, or to previous versions of the same UX. However, questionnaires do not allow clarification and often do not lead to deep understanding. For research, it may be advantageous to employ a methodological triangulation strategy, utilizing the questionnaire in conjunction with other techniques (Rogers et al., 2011-a).

As discussed previously, the established, complete definition of UX includes system, product, and service experiences and includes elements of usability and hedonic quality. However, there are no questionnaires that attempt to assess the complete UX. As shown in Table 1, we reviewed 31 UX questionnaires on two dimensions: scope (product/system versus service) and assessed qualities (usability versus hedonic). As Table 1 shows, 15 of the 31 assessment tools apply to service experience. Of those 15, only one, the Customer Experience Index (CXi or CX

Index™), assesses both usability and hedonic quality. However, the CXi is only applicable to an overarching service or brand experience and does not include specific product or service experience assessments. Furthermore, CXi is administered by Forrester Research, which does not allow complete visibility into their process (Forrester Research, Inc., n.d.).

In addition to the CXi, five other assessment tools measure both usability and hedonic quality but only for product or system experiences and not for service experiences. Those tools include the AttrakDiff2™ (Hassenzahl et al., 2003), Software Usability Measurement Inventory (SUMI) (Kirakowski & Corbett, 1993), Standardized User Experience Percentile Rank Questionnaire (SUPR-Q®) (Sauro, 2015), User Experience Questionnaire (UEQ) (Lund, 2001), and Website Quality (WEBQUAL™) (Wang & Senecal, 2007). Both the SUPR-Q and WEBQUAL are highly specialized for evaluating websites, so expanding their use to assess other types of products, not to mention service experiences, would be challenging. The AttrakDiff2, SUMI, and UEQ could all be modified to include service experience assessment with seemingly little effort. However, with three usability factors and three hedonic quality factors, the UEQ has a more balanced approach to assessing traditional usability and hedonic qualities than the AttrakDiff2, which includes only one usability factor, and the SUMI, which includes only one hedonic quality factor (Hassenzahl et al., 2003; Kirakowski & Corbett, 1993; Laugwitz et al., 2008). Therefore, the UEQ provided the best opportunity to create an assessment tool that can measure product, system, and service experiences on dimensions of usability and hedonic quality.

Table 1. Inventory and Brief Evaluation of UX Questionnaire Assessment Tools (Based on Tool Scope and Qualities Assessed)

Tool	Applies to Product/ System	Applies to Service	Assesses Usability	Assesses Hedonic Quality	Source
After-Scenario Questionnaire	Yes	Yes	Yes	No	Lewis, 1995
American Customer Satisfaction Index	Yes	Yes	No	Yes	American Customer Satisfaction Index LLC, n.d.
AttrakDiff2	Yes	No	Yes	Yes	Hassenzahl et al., 2003
Computer System Usability Questionnaire	Yes	No	Yes	No	Lewis, 1995
Customer Effort Score	Yes	Yes	Yes	No	Dixon et al., 2010
Customer Experience Index	No	Yes	Yes	Yes	Forrester Research, Inc., n.d.
Customer Satisfaction	Yes	Yes	No	Yes	Bendle et al., 2016
Emotional Metric Outcomes	Yes	Yes	No	Yes	Lewis & Mayes, 2014
Information Satisfaction	Yes	No	No	Yes	Lascu & Clow, 2008
Intranet Satisfaction Questionnaire	Yes	No	Yes	No	Bargas-Avila et al., 2009
NASA Task Load Index	Yes	Yes	No	No	Hart & Staveland, 1988
Net Promoter Score®	Yes	Yes	No	No	Reichheld, 2003
Post-Study Usability Questionnaire	Yes	No	Yes	No	Lewis, 1992

Tool	Applies to Product/ System	Applies to Service	Assesses Usability	Assesses Hedonic Quality	Source
Practical Usability Rating by Experts	Yes	Yes	Yes	No	Rohrer et al., 2016
Questionnaire for User Interaction Satisfaction (QUIS™)	Yes	No	Yes	No	Chin et al., 1988
SERVPERF	No	Yes	No	No	Cronin Jr. & Taylor, 1994
SERVQUAL	No	Yes	No	No	Parasuraman et al., 1988
Single Ease Question	Yes	Yes	Yes	No	Sauro & Dumas, 2009
Software Usability Measurement Inventory	Yes	No	Yes	Yes	Kirakowski & Corbett, 1993
Standardized User Experience Percentile Rank Questionnaire	Yes	No	Yes	Yes	Sauro, 2015
Subjective Mental Effort Question	Yes	Yes	Yes	No	Zijlstra & van Doorn, 1985
System Usability Scale	Yes	Yes	Yes	No	Brooke, 1996
Usability Magnitude Estimation	Yes	Yes	Yes	No	McGee, 2003
Usability Metric for User Experience	Yes	No	Yes	No	Finstad, 2010
Usability Metric for User Experience Lite	Yes	No	Yes	No	Lewis, 2013
Usefulness, Satisfaction, and Ease-of-Use	Yes	Yes	Yes	No	Lund, 2001
User Experience Questionnaire	Yes	No	Yes	Yes	Laugwitz et al., 2008
Web Quality	Yes	No	Yes	No	Aladwani & Palvia, 2002
Website Analysis and Measurement Inventory	Yes	No	Yes	No	Kirakowski & Cierlik, 1998
Website Quality	Yes	No	Yes	Yes	Loiacono et al., 2002
Website Usability	Yes	No	Yes	No	Wang & Senecal, 2007

User Experience Questionnaire (UEQ)

The UEQ is an efficient product experience evaluation tool to supplement traditional expert evaluations and usability testing. It was originally created in German in 2006 (Laugwitz et al., 2006) and translated into English in 2008 (Laugwitz et al., 2008). Based on the UX framework proposed by Hassenzahl (2001), the aim of the UEQ was to measure the holistic UX including the aspects of perceived ergonomic and hedonic quality as well as overall perceived attractiveness (Laugwitz et al., 2008).

UEQ Method

The UEQ can be administered via paper and pencil or online survey. Participants are presented with 26 seven-point semantic differentials, with opposite attributes on either end of the scale. Based on their initial perceptions, participants rate their opinion for a product by checking a value within a scale. For half of the questions, the positive attribute is presented at the beginning of the differential, and for the other half it is flipped, providing the negative attribute first (Schrepp, 2019).

Raw data from the UEQ is analyzed by first transforming the data to a -3 to +3 scale, with -3 indicating that the participant associated the product experience most closely with the negative attribute and +3 indicating the opposite. For each participant, scores comprising each of the six UEQ factors are averaged. Each average factor score is then averaged across all participants to obtain six separate factor scores for the product of interest. Those six scores can be rolled up into attractiveness, pragmatic quality, and hedonic quality as shown in Figure 1. Table 2 shows each of the UEQ adjective pairs along with the factor to which each pair contributes. A single UEQ score cannot be attained without following a separate method, which we described earlier in the KPI approach (Hinderks et al., 2019). Of course, additional analyses of correlation of individual items with each factor, as well as variance of responses within the participant pool, are suggested for a standard UEQ analysis. There are also recommended methods for eliminating data from analysis based on inconsistent participant responses (Schrepp, 2019).

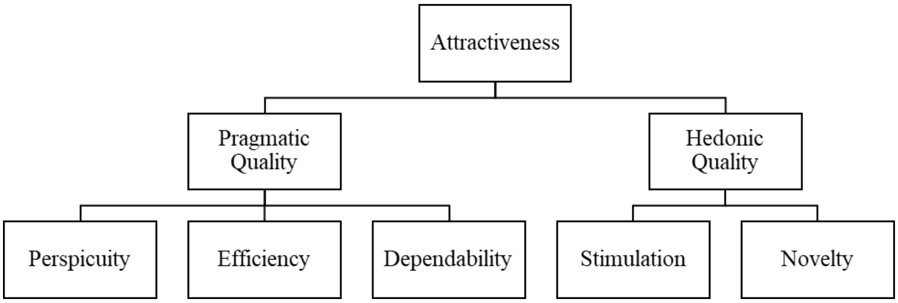


Figure 1. The assumed underlying structure of the UEQ factors (Schrepp et al., 2017).

Table 2. UEQ Individual Adjective Pairs and Their Respective Contributing Factors (Laugwitz et al., 2008)

UEQ Adjective Pairs	UEQ Factor
annoying/enjoyable	Attractiveness
attractive/unattractive	Attractiveness
friendly/unfriendly	Attractiveness
good/bad	Attractiveness
unlikable/pleasing	Attractiveness
unpleasant/pleasant	Attractiveness
meets expectations/does not meet expectations	Dependability
obstructive/supportive	Dependability
secure/not secure	Dependability
unpredictable/predictable	Dependability
fast/slow	Efficiency
impractical/practical	Efficiency
inefficient/efficient	Efficiency

UEQ Adjective Pairs	UEQ Factor
organized/cluttered	Efficiency
conservative/innovative	Novelty
creative/dull	Novelty
inventive/conventional	Novelty
usual/leading edge	Novelty
clear/confusing	Perspicuity
complicated/easy	Perspicuity
easy to learn/difficult to learn	Perspicuity
not understandable/understandable	Perspicuity
boring/exciting	Stimulation
motivating/demotivating	Stimulation
not interesting/interesting	Stimulation
valuable/inferior	Stimulation

Research Overview

The next three sections describe three separate studies designed to develop and validate a robust questionnaire that measures both pragmatic and hedonic user experience qualities across product and service experiences. The first study, phase 1, takes the initial step by generalizing product-centric language from the UEQ so that the language is applicable to both product and service experiences. Furthermore, the first study examines the new, generalized version of the UEQ—the Generalized User Experience Questionnaire (UEQ-G)—within the original, product experience context for which the UEQ was originally designed. Study two then tests the UEQ-G in a series of controlled service scenarios which were designed to test each factor of the UEQ-G separately. Finally, study 3 applies the UEQ-G in the field to scenarios that include both product and service modalities within the same experience. Each of the following studies complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Mississippi State University (IRB-20-312 & IRB-20-315). Informed consent was obtained from each participant.

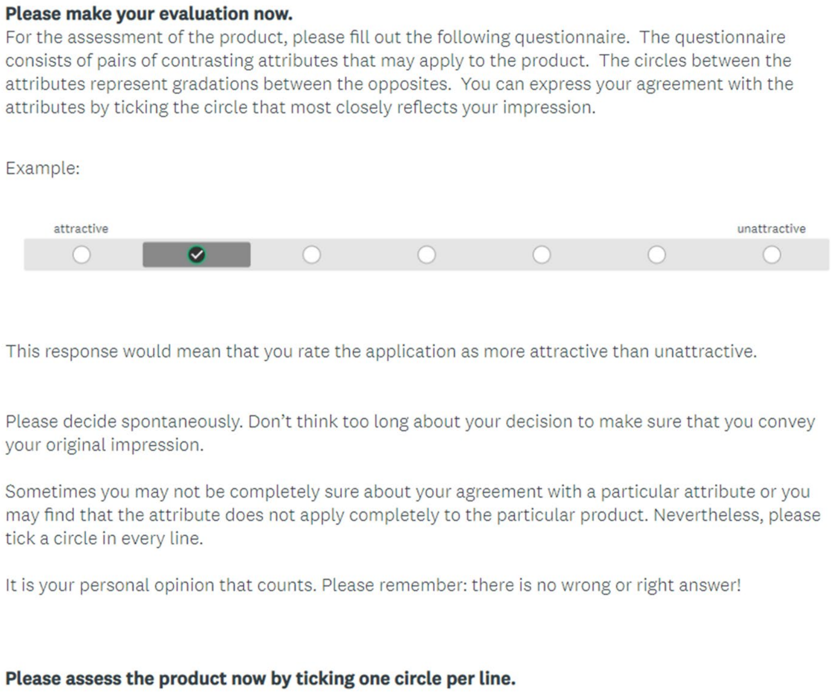
Phase 1: Generalizing the UEQ to Measure Product and Service Experience

This study sought to generalize the language used in the UEQ, thereby creating a generalized user experience questionnaire (UEQ-G). Additionally, this study tested the new UEQ-G alongside the original UEQ in product scenarios in which the UEQ is known to be effective. The goal of this study was not to test the UEQ-G in a novel situation but to first confirm that it continued to perform in its original context after being generalized.

Methodology

Expert Revision Process

We formed a panel of nine expert UX professionals including researchers, designers, and information architects, all with significant industry experience. We asked the panel to identify which words or phrases in the current UEQ were product-oriented, rather than generalizable to a more holistic UX. Each professional independently reviewed the UEQ and highlighted specific words or phrases that they believed fit that description. Then, the panel discussed all the highlighted portions and collectively agreed on generalized replacement terms or phrases. The instructions for the original UEQ (format modified for this study) are shown in Figure 2; the modified instructions are shown using italics and red font to indicate changes in Figure 3. Additionally, the "easy to learn/difficult to learn" adjective pair was revised to "easy to grasp/difficult to grasp." The expert revision process only resulted in a couple small changes, so there were early expectations that the revised UEQ would be successful when tested in its original context.



Please make your evaluation now.

For the assessment of the product, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:

attractive ○ ● ○ ○ ○ ○ ○ unattractive

This response would mean that you rate the application as more attractive than unattractive.

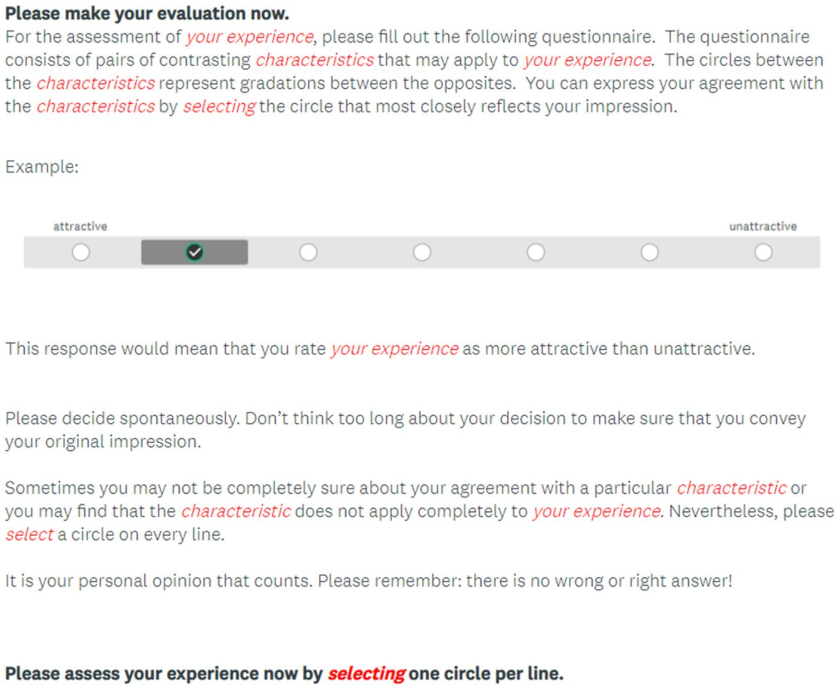
Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Please assess the product now by ticking one circle per line.

Figure 2. Screenshot of original UEQ instructions (formatted specifically for this study).



Please make your evaluation now.

For the assessment of *your experience*, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting *characteristics* that may apply to *your experience*. The circles between the *characteristics* represent gradations between the opposites. You can express your agreement with the *characteristics* by *selecting* the circle that most closely reflects your impression.

Example:

attractive ○ ✓ ○ ○ ○ ○ ○ unattractive

This response would mean that you rate *your experience* as more attractive than unattractive.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular *characteristic* or you may find that the *characteristic* does not apply completely to *your experience*. Nevertheless, please *select* a circle on every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Please assess your experience now by *selecting* one circle per line.

Figure 3. Screenshot of revised UEQ-G instructions (formatted specifically for this study) using italics and red font to show the changes from the original UEQ.

Participants

A human intelligence task (HIT) was posted on Amazon® Mechanical Turk™ with an incentive of $5.00 for successful HIT completion. To be eligible, Mechanical Turk workers had to be in the United States, have a HIT approval rating of at least 50%, have completed at least 50 HITs, have normal or corrected to normal vision, have full-color vision, be planning to use an Apple® iPhone® for the study, and have no experience with the Weather Puppy™ or Kelley Blue Book® mobile apps. An estimated 20 minutes were required for each participant to complete the HIT. As a first step, participants had to pass a screener questionnaire. The screener was attempted by 2,956 workers. Of those, 913 workers successfully passed, and 408 workers chose to begin the study. The 408 participants were comprised of 234 females (57.35%), 172 males (42.16%), and 2 individuals (0.49%) who preferred not to answer the question regarding gender. Participant ages ranged from 18 to 70 years of age with an average age of 33.48 (*SD* = 10.31). Only 268 participants successfully completed the study. Furthermore, for each participant, the difference between the highest and lowest individual adjective pair values in each factor was calculated. Forty participants had more than a difference of three for more than two factors, and they were removed from analysis due to a suspected lack of participant thoughtfulness, in alignment with previous recommendations (Schrepp, 2019). Another 4 participants had data that were less than 1.5 times the interquartile range below the first quartile; therefore, they were identified as outliers and removed as well. In all, data from 224 participants were used in this study. Based on an a priori power analysis with $\alpha = 0.05$, $\mathrm{d} = 0.50$, and $\beta = 0.10$, a sample size of 99 for each population was necessary to detect a mid-sized effect for a Mann-Whitney U Test. With a sample size greatly exceeding that recommended by the power analysis, a failure to detect a difference was unlikely to be due to an insufficient sample size.

Procedure

After consenting to participate and completing demographic questionnaires, participants were asked to complete four study tasks within one of the two randomly assigned apps (Weather Puppy or Kelley Blue Book) and then complete either the UEQ or UEQ-G as well as the AttrakDiff2 in a SurveyMonkey® survey.

For the Weather Puppy app, participants were asked to complete the following tasks:

1. Download and open the "Weather Puppy Forecast + Radar" app from the App Store.
2. Using the app, find and review the chance of rain over the next few hours in your current city.
3. Using the app, find the current temperature in London, England.
4. Change your puppy theme within the app.

For the Kelley Blue Book app, participants were asked to complete the following tasks:

1. Download and open the Kelley Blue Book app from the App Store.
2. Using the app, find the fair market value of any used car you want.
3. Using the app, find the nearest Honda™ car dealer to your current location.
4. Using the app, find copyright and trademark information for Kelley Blue Book.

Mobile apps were chosen for testing due to the apps' availability and recent studies confirming the effectiveness of the UEQ for evaluating mobile apps (Sabukunze & Arakaza, 2021; Hartono et al., 2022). Then, participants were asked to complete four more study tasks in the other app and complete the same questionnaire set as they did for the first app. Aside from the screening process via Mechanical Turk, all studies were conducted via participants' mobile phones.

Results

Data Preparation

To begin, raw data from the UEQ, UEQ-G, and AttrakDiff2 were analyzed by first transforming the data to a -3 to +3 scale, with -3 indicating the participant associated the product experience most closely with the negative attribute and +3 indicating the opposite. Factor scores were calculated for each participant by averaging question scores from each UEQ, UEQ-G, and AttrakDiff2 respective factor. For each participant, the difference between the highest and lowest individual adjective pair values in each factor were calculated. In addition to the outliers previously mentioned, any participants who incorrectly answered an attention check question were removed from the analysis. Attention check questions were embedded toward the middle of the UEQ, UEQ-G, and AttrakDiff2. The questions simply asked the participant to "select the 3rd option" and were formatted identically to the surrounding items. Finally, any participant who had a factor score greater than 1.5 times the interquartile range above the third quartile, or less than 1.5 times the interquartile range below the first quartile for a given factor, was labeled as an outlier and removed from analysis.

UEQ and UEQ-G Individual Adjective Pair Score Comparisons

Table 3 shows the individual adjective pair score means and standard deviations for each UEQ version.

Table 3. Mean UEQ/UEQ-G Individual Adjective Pair Scores Organized by UEQ Version

Individual UEQ/UEQ-G Adjective Pairs	**UEQ**			**UEQ-G**		
	$\bar{x}$	s	n	$\bar{x}$	s	n
annoying/enjoyable	0.64	1.82	268	0.77	1.80	180
not understandable/understandable	1.46	1.62	268	1.40	1.68	180
dull/creative	0.68	1.70	268	0.66	1.74	180
difficult to learn/easy to learn (changed to “difficult to grasp/easy to grasp” for the UEQ-G)	1.44	1.65	268	1.37	1.59	180
inferior/valuable	0.84	1.60	268	1.02	1.59	180
boring/exciting	0.31	1.62	268	0.52	1.63	180
not interesting/interesting	0.77	1.70	268	0.88	1.70	180
unpredictable/predictable	0.85	1.47	268	0.78	1.49	180
slow/fast	0.88	1.67	268	1.12	1.51	180
conventional/inventive	0.08	1.78	268	0.10	1.71	180
obstructive/supportive	0.84	1.55	268	0.79	1.60	180
bad/good	1.20	1.69	268	1.28	1.64	180
complicated/easy	1.00	1.72	268	1.06	1.64	180
unlikable/pleasing	0.98	1.70	268	1.04	1.71	180
usual/leading edge	-0.32	1.64	268	-0.23	1.58	180
unpleasant/pleasant	1.11	1.69	268	1.13	1.69	180
not secure/secure	0.85	1.44	268	0.94	1.41	180
demotivating/motivating	0.64	1.50	268	0.86	1.48	180
does not meet expectations/meets expectations	1.14	1.86	268	1.26	1.74	180
inefficient/efficient	0.91	1.77	268	0.97	1.76	180
confusing/clear	1.03	1.82	268	1.08	1.82	180
impractical/practical	1.13	1.63	268	1.18	1.66	180
organized/cluttered	0.86	1.86	268	0.92	1.80	180
unattractive/attractive	0.92	1.76	268	0.82	1.82	180
unfriendly/friendly	1.21	1.63	268	1.42	1.56	180
conservative/innovative	0.19	1.56	268	0.22	1.63	180

A series of Mann-Whitney U Tests with a Bonferroni-corrected alpha of 0.002 indicated that there was no statistically significant difference between any of the individual adjective pair scores due to UEQ versions. Table 4 shows the results of the Mann-Whitney U Tests.

Table 4. Mann-Whitney Results from Testing the Effect of UEQ Version on UEQ/UEQ-G Individual Adjective Pair Scores

Individual UEQ/UEQ-G Adjective Pairs	***U***	***z***	***p***
annoying/enjoyable	23119.00	-0.76	.450
not understandable/understandable	23947.00	-0.13	.894
dull/creative	24082.50	-0.03	.977
difficult to learn/easy to learn (changed to "difficult to grasp/easy to grasp" for the UEQ-G)	23115.00	-0.77	.441
inferior/valuable	22211.00	-1.45	.147
boring/exciting	22286.00	-1.39	.165
not interesting/interesting	23213.50	-0.69	.493
unpredictable/predictable	23809.00	-0.24	.813
slow/fast	22435.50	-1.28	.200
conventional/inventive	23997.00	-0.09	.926
obstructive/supportive	23769.50	-0.27	.791
bad/good	23612.00	-0.39	.699
complicated/easy	23876.00	-0.19	.853
unlikable/pleasing	23411.50	-0.54	.591
usual/leading edge	23850.50	-0.20	.838
unpleasant/pleasant	23954.50	-0.13	.900
not secure/secure	23223.50	-0.69	.493
demotivating/motivating	22263.50	-1.42	.157
does not meet expectations/meets expectations	23558.50	-0.43	.668
inefficient/efficient	23640.50	-0.36	.716
confusing/clear	23641.50	-0.36	.716
impractical/practical	23434.00	-0.52	.602
organized/cluttered	23876.00	-0.18	.853
unattractive/attractive	23465.50	-0.50	.620
unfriendly/friendly	22351.50	-1.35	.177
conservative/innovative	23977.50	-0.11	.914

UEQ and UEQ-G Factor Score Comparisons

Table 5 shows the factor score means and standard deviations for each UEQ version.

Table 5. Mean UEQ/UEQ-G Factor Scores Organized by UEQ Version

UEQ/UEQ-G Factors	**UEQ**			**UEQ-G**		
	$\bar{x}$	***s***	***n***	**$\bar{x}$**	***s***	***d***
Attractiveness	1.01	1.56	268	1.08	1.54	180
Perspicuity	1.23	1.56	268	1.23	1.55	180
Efficiency	0.95	1.45	268	1.05	1.48	180
Dependability	0.92	1.26	268	0.94	1.27	180
Stimulation	0.64	1.41	268	0.82	1.38	180
Novelty	0.16	1.40	268	0.18	1.42	180

A series of Mann-Whitney U Tests with a Bonferroni-corrected alpha of 0.008 indicated that there was no statistically significant difference between any of the factor scores due to UEQ versions. Table 6 shows the results of the Mann-Whitney U Tests.

Table 6. Mann-Whitney Results from Testing the Effect of UEQ Version on UEQ/UEQ-G Factor Scores

UEQ/UEQ-G Factors	***U***	***z***	***p***
Attractiveness	23553.00	-0.42	.673
Perspicuity	24037.50	-0.06	.951
Efficiency	22948.00	-0.87	.382
Dependability	23690.00	-0.32	.748
Stimulation	22191.00	-1.44	.150
Novelty	24032.50	-0.07	.948

Internal Consistency of the UEQ Factors

Cronbach's alpha (Cortina, 1993) was calculated for each of the UEQ factors. Alpha values of 0.70 and above are generally thought to show a high level of internal consistency (Landauer, 1997). Each factor was found to be highly internally consistent including attractiveness (6 items, α = .959), perspicuity (4 items, α = .935), efficiency (4 items, α = .854), dependability (4 items, α = .802), stimulation (4 items, α = .899), and novelty (4 items, α = .862).

Internal Consistency of the UEQ-G Factors

Cronbach's alpha (Cortina, 1993) was calculated for each of the UEQ-G factors. Again, a value of 0.70 is thought to indicate a high level of internal consistency (Landauer, 1997). Each factor was found to be highly internally consistent including attractiveness (6 items, α = .954), perspicuity (4 items, α = .942), efficiency (4 items, α = .899), dependability (4 items, α = .831), stimulation (4 items, α = .883), and novelty (4 items, α = .876).

Correlations Among UEQ/UEQ-G and AttrakDiff2 Factors

As no significant difference was found between UEQ and UEQ-G individual or factor scores, data from both were combined, and a Spearman correlation was performed to identify if there were significant relationships among combined UEQ/UEQ-G factors and AttrakDiff2 factors. Statistically significant relationships were found among all UEQ/UEQ-G and AttrakDiff2 factors with 446 degrees of freedom and $p < .001$ for all factor combinations. Table 7 shows the correlation coefficients for each factor combination. Except for the relationship between novelty and pragmatic quality, each of the correlation coefficients was greater than 0.50 and can be classified as strong positive correlations. Additionally, the novelty and pragmatic quality correlation coefficient was greater than 0.30 and can be classified as a moderate positive correlation (Cohen, 1977).

Table 7. Correlations Coefficients Among UEQ/UEQ-G Factor Scores and AttrakDiff2 Factor Scores

UEQ/UEQ-G Factors	**AttrakDiff2 Factors**				
	Attractiveness	**Pragmatic Quality**	**Identification**	**Stimulation**	**Hedonic Quality**
Attractiveness	.940	.769	.829	.728	.841
Perspicuity	.741	.872	.689	.484	.624
Efficiency	.777	.845	.789	.499	.681
Dependability	.770	.840	.757	.481	.655
Stimulation	.888	.694	.812	.725	.831
Novelty	.724	.434	.669	.861	.840

Conclusion

In this study, participants were asked to interact with two mobile apps, and after experiencing each, they completed either the UEQ or the UEQ-G and the AttrakDiff2. Upon analyzing the participants' responses to the questionnaires, no difference could be found between the UEQ and UEQ-G in either the individual component scores or the calculated factor scores. The inability to detect a difference between the two questionnaire versions, despite having sufficient sample size, indicates that the survey versions are likely to elicit the same participant responses for the same product experience. As there were only minor changes between the UEQ and UEQ-G, this finding is not surprising.

When the UEQ was originally created, correlations were found between some of its factors and the AttrakDiff2's factors (Laugwitz et al., 2008). This study also explored those correlations as well, and the same significant relationships with the AttrakDiff2 factors were found. The findings of this study strengthen the confidence in the original UEQ findings as well as indicate the similarity of the new UEQ-G to its predecessor. Using Cronbach's alpha, this study also explored the internal consistency of the factors contained in both the UEQ and UEQ-G. Each factor across both versions of the UEQ was found to be quite high. Again, these findings both strengthen confidence in the existing UEQ as well as the new UEQ-G.

This study sought to introduce and begin to qualify the UEQ-G, a new, slightly modified version of the UEQ with language revisions that support service-based experience evaluations in addition to the product experience evaluations to which the legacy UEQ was purposed. However, rather than immediately testing the UEQ-G in a novel, service experience scenario, this study sought to test it in a traditional, product experience scenario alongside the original UEQ. Results from this study indicate that the UEQ-G is an appropriate evaluation tool for the traditional scenarios for which the original UEQ was created.

Phase 2: Using the UEQ-G to Evaluate Controlled Non-Digital Service Experiences

Delivering thoughtful, refined experiences is important, but experiences can be complex, having both pragmatic and hedonic qualities including product experience and service experience modalities within the same scenario. The UEQ-G has the potential to aid in this challenge by measuring those complex experiences. However, a critical question must be answered to ensure the UEQ-G is able to meet these expectations: Is the UEQ-G sensitive, valid, and reliable when evaluating service experiences? The purpose of this study is to answer this question.

Methodology

Participants

A series of human intelligence tasks (HITs) was posted on Amazon Mechanical Turk with incentives ranging from $2.50 to $4.75 depending on the estimated time required to complete the HIT. To be eligible for the HITs, Mechanical Turk workers had to be in the United States, have a HIT approval rating of at least 50%, and have completed at least 50 HITs. Depending on the scenario, an estimated 10 to 19 minutes were required for each participant to complete their HIT. A total of 636 workers agreed to participate on the informed consent page before moving to the demographic questionnaire which was completed by 632 participants. After the demographic questionnaire, 608 participants accessed their scenario videos; however, on the following page, only 325 participants correctly answered attention-check questions about the videos to move further into the study. Out of those 325 participants, 47 did not complete the study, 30 did not answer embedded attention-check questions correctly, and 7 completed the study twice, leaving only 241 participants. Furthermore, using the methods described by Lewis (1995) and Schrepp (2019), data from an additional 44 participants were removed from analysis due to a suspected lack of thoughtfulness. There was a total of 197 participants whose data was used in this study. The 197 participants were comprised of 116 males (58.88%), 79 females (40.10%), and 2 (1.02%) who preferred not to answer the question regarding gender. Participant ages ranged from 18 to 69 years of age with an average age of 34.14 years (*SD* = 10.06).

Procedure

After consenting to participate and completing a brief demographic questionnaire, participants were asked to watch one of six scenarios. Each scenario was experienced in either a high or low condition in which the UEQ-G factor being tested was designed to be, respectively, high or low. Participants' only tasks were to watch the experience recording, answer the attention-check question after the video, and complete the UEQ-G that followed. Experience videos were either already created and publicly available prior to this study, or they were created specifically for this study and made available publicly on YouTube™. Brief scenario descriptions can be found in the following section. Each experience video was embedded and viewed within a SurveyMonkey survey. On the survey page immediately following the video, participants were asked two multiple choice questions about the video to ensure they had watched it. After answering the video questions correctly, participants completed the UEQ-G.

Classroom Lecture with Varied Stimulation Level

Participants were asked to watch a lecture video and imagine that they had just entered a college classroom. In the low condition, a reduced stimulation level was achieved by using a lecture video on basic addition, a topic that any of the adult participants should have found dull. In the high condition, an elevated stimulation level was achieved by using a TED Talk™ video viewed nearly 42 million times, "How to speak so that people want to listen" (TED, 2013). With proven popularity, a universally applicable topic, and even an interactive portion of the lecture, the TED Talk video was much more stimulating for participants than the basic addition video.

Convenience Store Shopping with Varied Novelty Level

Participants were asked to watch recorded walkthroughs of convenience store shopping experiences. In the low condition, a reduced novelty level was achieved by showing a walkthrough video of a traditional convenience store. Participants watched as a customer entered the store and shopped. In the high condition, participants watched a walkthrough of an Amazon Go™ experience where customers were able to scan a QR code on their phone, select items to buy in the store, and leave without any additional interaction. Before watching the walkthrough, participants were asked to view a brief introductory video about Amazon Go.

Facilitated Group Discussion with Varied Dependability Level

Participants were asked to watch recordings of facilitated group discussions in which dependability was varied in low and high conditions by modifications to the facilitator's behavior.

In the low condition, the facilitator behaved erratically, providing no clear order or direction, changing topics frequently, staring at her phone during participant responses, and even leaving the room to take a phone call in the middle of the session. In the high condition, the facilitator introduced a single topic, presented an agenda at the beginning, and stayed focused on that topic the entire time. Additionally, she actively listened to what participants said and encouraged collaborative discussion among the group.

Photo Booth with Varied Efficiency Level

Participants were asked to watch recorded photo booth session experiences. For the low condition, reduced efficiency was achieved by utilizing an inefficient process in which photo booth props were cluttered across the room, the photographer provided little direction, and she was overly chatty. The photographer would take a single picture of a group on her phone, walk to the opposite side of a larger room to a Polaroid™ Lab Instant Printer, print the individual picture, and return to the photo-taking location across the room before repeating the process again for each group in the recording. Furthermore, the photographer stopped to check her email during the recorded scenario. For the high condition, increased efficiency was achieved by using a process which included the photographer taking each person's picture with a Polaroid Now™ i-Type Instant Camera which immediately printed the photo. Additionally, props were organized and set up in a space beside the picture-taking backdrop, and the photographer gave clear direction and organized the groups into a line.

Boardgame Overview with Varied Perspicuity Level

Participants were asked to watch recorded boardgame overview videos. For the low condition, reduced perspicuity was achieved by having the participants watch an 8-minute video on how to play Kanban™ Automotive Revolution: Driver's Edition, a highly complex boardgame with 89.13% (418/469) of boardgamegeek.com voters giving it a heavy complexity score, and 1.49% (7/469) giving it a light complexity score (Board Game Geek, n.d.-b). In the high condition, increased perspicuity was achieved by having the participants watch an overview video on Candy Land®, a much simpler boardgame with 1.29% (4/309) of boardgamegeek.com voters giving it a heavy complexity score, and 97.09% (300/309) giving it a light complexity score (Board Game Geek, n.d.-a).

Airport Lounge with Varied Attractiveness Level

Participants were asked to watch first-person recordings of airport lounges. For the low condition, reduced attractiveness was accomplished by having the participants watch a video of a barebones airport lounge in Uganda. The lounge featured tightly spaced seating, dated furniture, and minimal refreshments. In the high condition, increased attractiveness was accomplished by having the participants watch a video from a first-class lounge in Paris. The video featured upscale, modern styling and food as well as premium beverages, and private restrooms and relaxation areas.

Results

Data Preparation

Data from the UEQ-G were processed by converting each raw item score to a -3 to +3 scale, with -3 indicating the participant associated the experience most closely with the negative attribute and +3 indicating the opposite. Scores were then averaged within each factor for each individual participant.

Detecting Differences in Stimulation

Table 8 shows the results of the series of Mann-Whitney U Tests that were performed to determine if the UEQ-G's stimulation factor was significantly different between high and low conditions for any of the scenarios. Table 9 shows the means, standard deviations, and medians for stimulation scores organized by scenarios.

The classroom lecture scenario was designed to be more stimulating in the high scenario than in the low scenario. However, as seen in the results table below, no difference was detected in the classroom scenario. There were significant differences in stimulation levels in the convenience store shopping, facilitated group discussion, and photo booth scenarios, though.

Table 8. Mann-Whitney Test Results for High and Low Condition Stimulation Score Comparisons

Scenario	N_H	N_L	***U***	***z***	***p***
Classroom Lecture	19	13	110.50	-0.503	.312[a]
Convenience Store Shopping	14	19	46.00	-3.184	.001
Facilitated Group Discussion	16	15	34.00	-3.405	< .001
Photo Booth	17	18	91.00	-2.055	.041
Boardgame Overview	18	18	157.50	-0.143	.888
Airport Lounge	14	16	68.00	-1.841	.070

[a] One-tailed Mann-Whitney test

Table 9. Mean, Standard Deviation, and Median Stimulation Scores Organized by Scenario Condition

Scenario	**High Condition**			**Low Condition**			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	***s***	$\tilde{x}_H$	$\bar{x}_L$	***s***	$\tilde{x}_L$	
Classroom Lecture	1.51	1.07	1.75	0.94	1.85	1.75	0.57
Convenience Store Shopping	1.80	1.19	1.88	0.12	1.47	0.00	1.68
Facilitated Group Discussion	0.88	1.35	0.88	-1.02	1.28	-1.25	1.90
Photo Booth	1.07	1.46	1.00	0.10	1.23	0.25	0.97
Boardgame Overview	0.76	1.15	0.50	0.69	1.37	0.63	0.07
Airport Lounge	1.68	1.30	2.00	0.84	1.16	1.00	0.84

Detecting Differences in Novelty

Table 10 shows the results of the series of Mann-Whitney tests that were performed to determine if UEQ-G's novelty factor was significantly different between high and low conditions for any of the scenarios. Table 11 shows the means, standard deviations, and medians for novelty scores organized by scenarios.

The convenience store shopping scenario was designed to be more novel in the high scenario than in the low scenario, and the high condition was found to be significantly higher than the low condition. There was also a significant difference in novelty level in the facilitated group discussion scenario.

Table 10. Mann-Whitney Test Results for High and Low Condition Novelty Score Comparisons

Scenario	N_H	N_L	***U***	***z***	***p***
Classroom Lecture	19	13	113.00	-0.405	.705
Convenience Store Shopping	14	19	40.50	-3.381	< .001[a]
Facilitated Group Discussion	16	15	70.50	-1.967	.049
Photo Booth	17	18	127.00	-0.862	.405
Boardgame Overview	18	18	102.00	-1.912	.059
Airport Lounge	14	16	101.50	-0.440	.667

[a] One-tailed Mann-Whitney test

Table 11. Mean, Standard Deviation, and Median Novelty Scores Organized by Scenario Condition

Scenario	High Condition			Low Condition			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	s	$\tilde{x}_H$	$\bar{x}_L$	s	$\tilde{x}_L$	
Classroom Lecture	0.50	1.05	0.50	0.21	1.47	0.25	0.29
Convenience Store Shopping	1.09	1.50	0.50	-0.86	1.39	-1.00	1.95
Facilitated Group Discussion	-0.05	0.46	-0.13	-0.85	1.40	-0.75	0.80
Photo Booth	-0.03	1.20	0.25	-0.38	1.11	0.00	0.35
Boardgame Overview	-0.29	1.06	-0.38	0.40	1.02	0.00	-0.69
Airport Lounge	0.14	1.39	0.13	-0.11	0.69	0.00	0.25

Detecting Differences in Dependability

Table 12 shows the results of the series of Mann-Whitney tests that were performed to determine if UEQ-G’s dependability factor was significantly different between high and low conditions for any of the scenarios. Table 13 shows the means, standard deviations, and medians for dependability scores organized by scenarios.

The facilitated group discussion scenario was designed to be more dependable in the high scenario than in the low scenario, and the high condition was found to be significantly higher than the low condition.

Table 12. Mann-Whitney Test Results for High and Low Condition Dependability Score Comparisons

Scenario	N_H	N_L	U	z	p
Classroom Lecture	19	13	100.50	-0.887	.383
Convenience Store Shopping	14	19	103.00	-1.099	.287
Facilitated Group Discussion	16	15	31.50	-3.512	< .001[a]
Photo Booth	17	18	109.00	-1.456	.153
Boardgame Overview	18	18	129.00	-1.048	.308
Airport Lounge	14	16	66.50	-1.906	.058

[a]One-tailed Mann-Whitney test

Table 13. Mean, Standard Deviation, and Median Dependability Scores Organized by Scenario Condition

Scenario	High Condition			Low Condition			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	s	$\tilde{x}_H$	$\bar{x}_L$	s	$\tilde{x}_L$	
Classroom Lecture	1.24	0.90	1.25	1.50	1.12	2.00	-0.26
Convenience Store Shopping	1.55	1.07	1.63	1.12	0.81	1.25	0.43
Facilitated Group Discussion	1.00	1.09	1.00	-0.77	1.26	-0.50	1.77
Photo Booth	1.32	1.25	1.00	0.68	1.00	0.75	0.64
Boardgame Overview	1.26	0.99	1.13	0.83	1.25	0.88	0.43
Airport Lounge	1.66	0.72	1.63	0.97	0.91	1.00	0.69

Detecting Differences in Efficiency

Table 14 shows the results of the series of Mann-Whitney tests that were performed to determine if the UEQ-G's efficiency factor was significantly different between high and low conditions for any of the scenarios. Table 15 shows the means, standard deviations, and medians for efficiency scores organized by scenarios.

The photo booth scenario was designed to be more efficient in the high scenario than in the low scenario, and the high condition was found to be significantly higher than the low condition. There were also significant differences in efficiency levels in the facilitated group discussion and boardgame overview scenarios.

Table 14. Mann-Whitney Test Results for High and Low Condition Efficiency Score Comparisons

Scenario	N_H	N_L	*U*	*z*	*p*
Classroom Lecture	19	13	119.50	-0.154	.880
Convenience Store Shopping	14	19	99.00	-1.245	.226
Facilitated Group Discussion	16	15	32.50	-3.469	< .001
Photo Booth	17	18	91.50	-2.036	.021[a]
Boardgame Overview	18	18	93.50	-2.174	.029
Airport Lounge	14	16	70.00	-1.760	.085

[a] One-tailed Mann-Whitney test

Table 15. Mean, Standard Deviation, and Median Efficiency Scores Organized by Scenario Condition

Scenario	High Condition			Low Condition			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	*s*	$\tilde{x}_H$	$\bar{x}_L$	*s*	$\tilde{x}_L$	
Classroom Lecture	1.32	1.15	1.75	1.35	1.12	1.50	-0.03
Convenience Store Shopping	1.54	1.13	1.50	1.01	1.08	1.00	0.53
Facilitated Group Discussion	1.19	0.96	1.50	-0.42	1.23	-0.25	1.61
Photo Booth	1.10	0.92	1.25	0.22	1.44	0.13	0.88
Boardgame Overview	1.40	0.93	1.50	0.44	1.45	0.25	0.96
Airport Lounge	1.50	1.02	1.88	0.69	1.09	0.50	0.81

Detecting Differences in Perspicuity

Table 16 shows the results of the series of Mann-Whitney tests that were performed to determine if UEQ-G's perspicuity factor was significantly different between high and low conditions for any of the scenarios. Table 17 shows the means, standard deviations, and medians for perspicuity scores organized by scenarios.

The boardgame overview scenario was designed to be more perspicuous in the high scenario than in the low scenario, and the high condition was found to be significantly higher than the low condition. There were also significant differences in perspicuity levels in the facilitated group discussion and airport lounge scenarios.

Table 16. Mann-Whitney Test Results for High and Low Condition Perspicuity Score Comparisons

Scenario	N_H	N_L	U	z	p
Classroom Lecture	19	13	98.50	-0.971	.343
Convenience Store Shopping	14	19	130.00	-0.110	.928
Facilitated Group Discussion	16	15	41.00	-3.130	.001
Photo Booth	17	18	120.50	-1.078	.287
Boardgame Overview	18	18	93.00	-2.191	.015[a]
Airport Lounge	14	16	40.00	-3.004	.002

[a] One-tailed Mann-Whitney test

Table 17. Mean, Standard Deviation, and Median Perspicuity Scores Organized by Scenario Condition

Scenario	**High Condition**			**Low Condition**			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	s	$\tilde{x}_H$	$\bar{x}_L$	s	$\tilde{x}_L$	
Classroom Lecture	1.71	1.09	2.00	1.94	1.30	2.75	-0.23
Convenience Store Shopping	1.61	1.18	1.75	1.64	1.06	1.75	-0.03
Facilitated Group Discussion	1.39	1.16	1.88	-0.07	1.34	0.00	1.46
Photo Booth	1.93	0.91	2.00	1.61	0.91	1.88	0.32
Boardgame Overview	1.57	1.22	1.50	0.39	1.54	0.25	1.18
Airport Lounge	2.18	0.68	2.38	1.02	1.03	0.75	1.16

Detecting Differences in Attractiveness

Table 18 shows the results of the series of Mann-Whitney tests that were performed to determine if UEQ-G's attractiveness factor was significantly different between high and low conditions for any of the scenarios. Table 19 shows the means, standard deviations, and medians for attractiveness scores organized by scenarios.

The airport lounge scenario was designed to be more attractive in the high scenario than in the low scenario, and the high condition was found to be significantly higher than the low condition. There were also significant differences in attractiveness levels in the convenience store shopping, facilitated group discussion, and photo booth scenarios.

Table 18. Mann-Whitney Test Results for High and Low Condition Attractiveness Score Comparisons

Scenario	N_H	N_L	U	z	p
Classroom Lecture	19	13	98.50	-0.963	.343
Convenience Store Shopping	14	19	58.50	-2.722	.005
Facilitated Group Discussion	16	15	20.00	-3.960	< .001
Photo Booth	17	18	93.50	-1.970	.049
Boardgame Overview	18	18	135.00	-0.856	.406
Airport Lounge	14	16	56.50	-2.315	.010[a]

[a] One-tailed Mann-Whitney test

Table 19. Mean, Standard Deviation, and Median Attractiveness Scores Organized by Scenario Condition

Scenario	High Condition			Low Condition			$\bar{x}_H - \bar{x}_L$
	$\bar{x}_H$	s	$\tilde{x}_H$	$\bar{x}_L$	s	$\tilde{x}_L$	
Classroom Lecture	1.72	1.08	2.00	1.17	1.47	1.17	0.55
Convenience Store Shopping	1.89	1.06	2.42	0.58	1.23	0.00	1.31
Facilitated Group Discussion	1.27	1.21	1.42	-1.09	1.33	-1.33	2.36
Photo Booth	1.55	1.18	1.83	0.81	1.12	1.00	0.74
Boardgame Overview	1.26	1.05	1.17	0.79	1.29	0.83	0.47
Airport Lounge	1.98	1.16	2.25	0.97	1.12	1.09	1.01

Internal Consistency of UEQ-G Factors

Cronbach’s alpha was calculated for each of the UEQ-G factors. Each factor was found to be highly internally consistent including attractiveness (6 items, α = .935), perspicuity (4 items, α = .857), efficiency (4 items, α = .775), dependability (4 items, α = .776), stimulation (4 items, α = .908), and novelty (4 items, α = .674).

Conclusion

Indicative of the sensitivity of all, and the validity of most, of the UEQ-G factors, the UEQ-G was able to detect significant differences for the tested factor in all but one scenario; and even the factor from that one scenario was unexpectedly found to be significantly different in several other scenarios. In five of the six scenarios, the UEQ-G was able to detect a difference in the factor being tested (novelty, dependability, efficiency, perspicuity, and attractiveness) which indicates that those factors are both sensitive and valid. However, in one scenario (classroom lecture), the UEQ-G was not able to detect a difference in the tested factor (stimulation). Although the scenario conditions were designed to be different and even validated with a small pilot group, there was no guarantee that the high and low conditions were sufficiently distinct to be detected in this study. Regardless, since a difference was not detected in the scenario in which stimulation was the tested factor, a conclusion cannot be reached about its validity. However, significantly different levels in stimulation were observed in three other scenarios, which indicates the stimulation factor does show some sensitivity. Additionally, significantly different levels were seen outside of the tested scenarios for most of the other factors too. Significant differences were seen for novelty in one additional scenario, efficiency in two additional scenarios, perspicuity in two additional scenarios, and attractiveness in three additional scenarios. The results from the stimulation scenario were unexpected, yet each of the UEQ-G factors was sensitive enough to detect a difference in at least one scenario with sample sizes that ranged from only 30 to 36 participants, depending on the scenario.

The reliability of the UEQ-G is also supported by the results of this study. As with the original UEQ study (Laugwitz et al., 2008), UEQ-G factor scores were found to be highly internally consistent, as indicated by Cronbach’s alpha.

This study tested the newly introduced UEQ-G, a questionnaire built to test the holistic user experience using language suitable to product and service experiences as well as novel service experiences, and it was shown to be an effective tool for those situations. Each of the UEQ-G factors was found to be reliable and sensitive enough to detect differences in controlled service experience scenarios, but only attractiveness, perspicuity, efficiency, dependability, and novelty could be shown as valid for measuring what each was intended to measure. Additional research is necessary to ensure the stimulation factor is indeed valid in service experiences.

Phase 3: Using the UEQ-G to Evaluate Multimodal Experiences in the Field

Having shown that the UEQ-G is a reliable method for measuring experiences in both product and non-product experiences, this study took the final step in introducing the UEQ-G by exploring its use in experiences that include product and non-product modalities within the same scenario. Furthermore, this study explores the UEQ-G outside of controlled scenarios, exposing it to real situations in the field.

Methodology

Participants

Forty Mississippi State University (MSU) student participants were recruited from the MSU campus by posting and passing out flyers that offered a $25 Chick-fil-A® (CFA) credit as reimbursement and payment for participation in an approximately 1-hour study. As advertised on the flyers, participants were required to be at least 18 years of age, have a smartphone capable of downloading the CFA app, be licensed drivers, have access to a vehicle, and be willing to pick up food from CFA during the COVID-19 Pandemic. Qualifications based on these criteria were validated during study scheduling as well. Out of the 40 participants who completed the study, 5 did not answer an embedded attention check question correctly, leaving only 35 participants. Furthermore, using the methods described by Lewis (1995) and Schrepp (2019), data from 1 additional participant was removed from analysis due to a suspected lack of thoughtfulness. Therefore, there was a total of 34 participants whose data was used in this study. Those 34 participants were comprised of 20 females (58.82%) and 14 males (41.18%). All participants answered the question regarding gender. Participant ages ranged from 18 to 28 years of age with an average age of 20.91 years (*SD* = 2.45).

Procedure

After responding to the flyer and scheduling a time to participate in the study, participants were asked to join a Zoom call at their scheduled session time. During the Zoom call, each participant was asked to complete a SurveyMonkey survey that included both informed consent and a brief demographic questionnaire. Then, each participant downloaded the CFA mobile app and received a credit of $10 on the app. Next, participants were asked to complete three tasks:

1. Take the necessary steps to use the CFA mobile app to order food of your choosing for pickup.
2. Drive to the restaurant and pick up the food you ordered.
3. Return to debrief and receive CFA credit as reimbursement and payment.

Upon their return from picking up their orders, participants were asked to rejoin the Zoom call; at which point the researcher confirmed order retrieval and instructed each participant to complete a set of questionnaires in SurveyMonkey including the UEQ-G, After-Scenario Questionnaire (ASQ), Customer Satisfaction (CSAT), Emotional Metrics Outcome (EMO), and Likelihood to Recommend (LTR). Once the questionnaires were complete, each participant was given an additional $15 credit on their CFA app.

Results

Data Preparation

Raw data from the UEQ-G were analyzed by first transforming the data to a -3 to +3 scale, with -3 indicating the participant associated the experience most closely with the negative attribute and +3 indicating the opposite. UEQ-G factor scores were calculated for each participant by averaging question scores from the respective factor. Additionally, for each participant, perspicuity, efficiency, and dependability scores were averaged together to determine an overall pragmatic quality score. Stimulation and novelty scores were averaged together to determine an overall hedonic quality score. For each participant, the difference between the highest and lowest value in each factor was calculated, and any participant with more than a difference of three for more than three factors was removed from analysis due to a suspected lack of participant thoughtfulness per the previous recommendations (Schrepp, 2019). Overall ASQ scores were calculated for each participant by averaging each individual item score that was not marked as "not applicable." As recommended, overall ASQ scores were still calculated for participants with up to one item marked as "not applicable" (Lewis, 1995). However, participants with more than one item marked as "not applicable" or any items skipped would have been removed from ASQ analysis, but no participants were removed for this reason.

In addition to maintaining ordinal CSAT responses, individual participants were also placed into one of two categories: satisfied (participants who responded as "somewhat satisfied" or "extremely satisfied") and not satisfied (participants who responded as "extremely dissatisfied," "somewhat dissatisfied," or "neither satisfied nor dissatisfied"). However, upon inspection of the data, all respondents were coded as "satisfied" based on their responses (100% CSAT); therefore, raw CSAT scores had to be used for correlation analysis to produce a meaningful result. EMO scores from the PRA and PPA sections were recorded directly, whereas scores from the NRA and NPA sections were reversed. LTR values were recorded directly as well, but individual participants were also placed into traditional NPS categories of detractor (such as participants scoring between 0 and 6), passive (participants scoring 7 or 8), and promoter (participants scoring 9 or 10).

Descriptive Statistics

Means and standard deviations across the UEQ-G, ASQ, CSAT, EMO, and LTR scores are shown in Table 20.

Table 20. Mean Scores for the UEQ-G, ASQ, CSAT, EMO, and LTR

Scores	**Possible Range**	$\bar{x}$	s	***n***
Attractiveness (UEQ-G)	-3 to 3	2.56	0.45	34
Perspicuity (UEQ-G)	-3 to 3	2.74	0.47	34
Efficiency (UEQ-G)	-3 to 3	2.59	0.50	34
Dependability (UEQ-G)	-3 to 3	2.30	0.64	34
Stimulation (UEQ-G)	-3 to 3	1.65	0.97	34
Novelty (UEQ-G)	-3 to 3	0.92	1.20	34
Pragmatic Quality (UEQ-G)	-3 to 3	2.54	0.39	34
Hedonic Quality (UEQ-G)	-3 to 3	1.28	0.94	34
ASQ	1 to 7	6.80	0.39	34
CSAT	1 to 5	4.97	0.17	34
PRA (EMO)	0 to 10	8.27	1.63	34
NRA (EMO)	0 to 10	7.54	1.47	34
PPA (EMO)	0 to 10	9.21	1.01	34
NPA (EMO)	0 to 10	8.70	0.57	34
Overall (EMO)	0 to 10	8.43	0.99	34
LTR	0 to 10	9.38	1.10	34

Correlations Between Hedonic Quality Factor and ASQ, CSAT, EMO, and LTR

Table 21 provides a summarized, consolidated view of the previous tables showing only those correlations that are both statistically significant and have a value of at least 0.4.

Table 21. Significant Correlations Among UEQ-G's Factors and ASQ, CSAT, EMO, and LTR

Scores	UEQ-G Factors							
	A	**P**	**E**	**D**	**S**	**N**	**PQ**	**HQ**
ASQ	.488	.500	-	.451	.451	-	.563	-
CSAT	-	-	-	-	-	-	-	-
PRA (EMO)	.569	-	-	-	.728	.495	.424	.675
NRA (EMO)	.482	-	.470	.426	.471	-	.470	-
PPA (EMO)	.765	.431	.443	.463	.676	-	.593	.539
NPA (EMO)	.678	-	-	-	.596	.439	.496	.579
Overall (EMO)	.654	-	.522	.428	.669	-	.551	.589
LTR	.566	-	.419	-	.536	-	-	.476

A = attractiveness; P = perspicuity; E = efficiency; D = dependability; S = stimulation; N = novelty; PQ = pragmatic quality; HQ = hedonic quality; PRA = positive relationship affect; NRA = negative relationship affect; PPA = positive personal affect; NPA = negative personal affect

Conclusion

The strongest correlations among UEQ-G factor scores and scores from the other questionnaires are shown in Table 21. These correlations provide clues about the implications of UEQ-G's results. UEQ-G's attractiveness, efficiency, and stimulation factors are moderately to strongly correlated with LTR, meaning that those organizations interested in tracking and improving NPS should pay special attention to these UEQ-G factor scores. Furthermore, periodically running regular UEQ-G initiatives for key experiences in addition to regular NPS rounds may provide some diagnostic information for NPS responses. For example, an organization with decreasing NPS scores may also see a decrease in UEQ-G efficiency scores for key experiences and understand that the reduction in perceived efficiency is leading to lower NPS scores. Furthermore, even greater diagnostic capabilities may be achieved by pairing the UEQ-G with an importance-performance analysis as described in recent literature (Hinderks et al., 2019).

UEQ-G's attractiveness and stimulation factors are moderately to strongly correlated with each of the EMO factors. The consistent link between these UEQ-G factors and those of the EMO indicate the fundamental nature of these UEQ-G factors to assess the emotional impact of an experience on an individual. Each factor can impact either positively or negatively how individuals feel about their relationships with organizations as well as how they feel about themselves. As with LTR, the UEQ-G can serve as a diagnostic tool when used in conjunction with the EMO. Should an organization see a drop in EMO scores, the organization may look to trends in UEQ-G factor scores to identify the potential cause for that EMO drop.

The correlation between UEQ-G's perspicuity factor and EMO's PPA factor indicates that an experience that is clearer will lead to a more positive personal affect. This finding supports the phenomenon UX professionals often observe in usability testing: Participants judge themselves rather than the organization based on how well they understand or do not understand what they are testing (Anderson, 1981). Although there was also a correlation between perspicuity and NPA, it was weaker, indicating that while perspicuity can have either a positive or negative impact on personal affect, it has more potential to positively impact individuals. Remarkably, no significant relationship was indicated between perspicuity and PRA or NRA, again supporting the idea that individuals judge themselves rather than the organization when it comes to clarity and understanding.

UEQ-G's novelty factor and EMO's PRA and NPA scores were also both found to have moderate, positive correlations. This finding indicates that there is an opportunity for novelty to positively impact the relationships individuals have with organizations; otherwise, a lack of novelty can negatively impact how people see themselves.

UEQ-G’s efficiency and dependability factors were both found to be moderately correlated with EMO’s NRA and PPA. These relationships suggest that a lack of efficiency and dependability can negatively impact an individual’s view of an organization while having little impact on that individual personally; the presence of both factors will positively impact that individual while having little impact on his or her view of the organization.

UEQ-G’s attractiveness factor was found to have a moderate to strong positive correlation with all the factors (except CSAT). This finding supports Hassenzahl’s framework (2001), which identifies attractiveness as the highest-level factor to which all other factors contribute.

Evidence was found that the UEQ-G measures holistic UX, including pragmatic and hedonic qualities, based on the relationships identified between UEQ-G factors and ASQ and EMO scores. UEQ-G’s pragmatic quality score was found to be positively correlated with the ASQ, which measures traditional usability elements (Tullis & Albert, 2013). And, UEQ-G’s hedonic quality score was found to be positively correlated with the EMO, which measures emotional elements (Sauro & Lewis, 2016).

No statistically significant correlation was found between raw CSAT and any of the UEQ-G factors. CSAT responses were overwhelmingly positive with a calculated CSAT of 100%, meaning that every participant marked either “somewhat satisfied” or “very satisfied.” Even when analyzing the raw CSAT score, the scores still averaged 4.97 (*SD* = 0.17) out of 5.00, indicating that the CSAT scale was largely maxed out by participants’ CFA experiences. With so little variety in responses, detecting a significant correlation was unlikely.

This study explored using the UEQ-G in the field for an experience that included product and service experiences within the same extended experience. The relationships identified between the UEQ-G and other questionnaire factors indicate that the UEQ-G is capable of measuring the holistic experience including pragmatic and hedonic factors within a multimodal experience, something no other questionnaire has demonstrated to date.

Additionally, several other important relationships between the UEQ-G and the established questionnaires were found. These additional relationships demonstrate the potential for the UEQ-G to supplement widely used tools such as the NPS to provide additional color to observed trends.

Recommendations

There are several directions for future UEQ-G research. As mentioned, additional research is necessary to ensure the stimulation factor is valid in these types of situations. Additional work could also be done to identify the relationship between the UEQ-G and other existing questionnaires or user experience metrics. Furthermore, the UEQ-G could be tested in additional service experience, or even traditional product experience, scenarios. One step that is necessary to accomplish the UEQ-G’s original purpose is to test the UEQ-G in a scenario that includes product and service experiences within the same scenario.

Many valuable opportunities are available for future UEQ-G research. Future research could explore the relationship between the UEQ-G and a larger variety of existing questionnaires and metrics. It could explore the same questionnaires from this study in a larger variety of situations. Additionally, future research could and should focus on exploring experiences that cross modalities with a variety of anticipated values for each of the UEQ-G factors. Research comparing multimodal experiences would also be valuable.

Tips for Usability Practitioners

- Consider using the UEQ-G to assess product user experiences. The UEQ-G performs as well as the original UEQ but has more generalized language.
- Test the UEQ-G in service and multi-modal user experience evaluations. The UEQ-G shows great potential in those scenarios and is the first tool of its type to measure extended multi-modal experiences.

- Try the UEQ-G alongside the NPS to gain greater insights into potential NPS score variations. The UEQ-G has a few factors that correlate with NPS scores.
- Share UEQ-G findings and lessons learned through reputable, peer-reviewed publications such as the *Journal of User Experience*.

Acknowledgements

I would like to acknowledge the contributions of several people who helped make this research possible. Without hesitation, I must thank each of my co-authors who have, of course, provided tremendous support and contribution to this work. I also give thanks to Bailey Jose and her team of undergraduate researchers for leading on-campus data collection and participating in scenario creation. To my colleagues at BeyondTrust, I thank you for working with me to accomplish the critical task of developing the Generalized User Experience Questionnaire. Thank you also to my employer, BeyondTrust (formerly Bomgar), who has supported my research from the very beginning. Finally, I cannot thank my wife, Emily, enough for the support she has given throughout this process. From proofreading and editing countless revisions to taking care of the kids and keeping the house running while I focused, she is just as responsible for the success of this work as me.

References

Aladwani, A. M., & Palvia, P. C. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information & Management*, *39*(6), 467–476.

American Customer Satisfaction Index LLC. (n.d.). *The Science of Customer Satisfaction*. Retrieved December 7, 2019, from https://www.theacsi.org/about-acsi/the-science-of-customer-satisfaction

Anderson, R. E. (1981, January). Some effects of considerate and inconsiderate systems. *ACM SIGSOC Bulletin*, 11–16.

Bargas-Avila, J. A., Lötscher, J., Orsini, S., & Opwis, K. (2009). Intranet satisfaction questionnaire: Development and validation of a questionnaire to measure user satisfaction with the intranet. *Computers in Human Behavior*, *25*(6), 1241–1250.

Bendle, N. T., Farris, P. W., Pfeifer, P. E., & Reibstein, D. J. (2016). Share of hearts, minds, and markets. In *Marketing metrics: The manager's guide to measuring marketing performance* (pp. 17–66). Pearson.

Board Game Geek. (n.d.-a.) *Candy Land*. Retrieved June 5, 2020, from https://boardgamegeek.com/boardgame/5048/candy-land

Board Game Geek. (n.d.-b.) *Kanban: Driver's Edition*. Retrieved June 5, 2020, from https://boardgamegeek.com/boardgame/109276/kanban-drivers-edition

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the Human-Computer Interface. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 213–218.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Elsevier Science.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 98–104.

Cronin Jr., J. J., & Taylor, S. A. (1994). SERVPERF versus SERVQUAL: Reconciling performance-based and perceptions-minus-expectations measurement of service quality. *Journal of Marketing*, *58*(1), 125–131.

Dixon, M., Freeman, K., & Toman, N. (2010, July-August). Stop trying to delight your customers. *Harvard Business Review*, *88*(7/8), 116–122.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, *22*(5), 323–327.

Forrester Research, Inc. (n.d.). *CX Index*. Retrieved December 7, 2019, from https://go.forrester.com/analytics/cx-index/

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Hartono, S., Shahudin, F., Widagdo, H. H., & Hendrawan, T. (2022). The analysis of online order mobile application using User Experience Questionnaire (a case study approach). *2022 International Conference on Information Management and Technology (ICIMTech)*, *Semarang*, *Indonesia*, 682–686. IEEE.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481–499.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In Szwillus, G., Ziegler, J. (Eds.), *Mensch & computer* (pp. 187–196). Vieweg+Teubner Verlag.

Hinderks, A., Meiners, A.-L., Domínguez-Mayo, F. J., & Thomaschewski, J. (2019). Applying Importance-Performance Analysis (IPA) to interpret the results of the User Expreience

Questionnaire (UEQ). *WEBIST 2019: 15th International Conference on Web Information Systems and Technologies*, 388–395.

Hinderks, A., Schrepp, M., Mayo, F. J., Escalona, M. J., & Thomaschewski, J. (2019, July). Developing a UX KPI based on the user experience questionnaire. *Computer Standards & Interfaces*, *65*, 38–44.

International Organization for Standardization. (2019). *Ergonomics of human-system interaction - Part 210: Human-centered design for interactive systems* (ISO Standard No. 9241-210:2019). Retrieved November 23, 2019, from https://www.iso.org/standard/77520.html?browse=tc

Kirakowski, J., & Cierlik, B. (1998). Measuring the usability of web sites. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(4), 424–428.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212.

Landauer, T. K. (1997). Behavioral research methods in Human-Computer Interaction. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbood of Human-Computer Interaction* (pp. 203–227). North-Holland.

Lascu, D.-N., & Clow, K. E. (2008). Web site interaction satisfaction: Scale development considerations. *Journal of Internet Commerce*, *7*(3), 359–378.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a User Experience Questionnaire. *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society*, *Graz*, *Austria*.

Laugwitz, B., Schrepp, M., & Held, T. (2006). Konstruktion eines fragebogens zur messung der user experience von softwareprodukten. In *Mensch und computer 2006: Mensch und Computer im Strukturwandel* (pp. 125–134). Oldenbourg Verlag.

Lewis, J. R. (1992). Psychometric evaluation of the Post-Study System Usability Questionnaire: The PSSUQ. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *36*(16), 1259–1260.

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

Lewis, J. R. (2013). Critical review of 'The usability metric for user experience'. *Interacting with Computers*, *25*(4), 320–324.

Lewis, J. R., & Mayes, D. K. (2014). Development and psychometric evaluation of the Emotional Metric Outcomes (EMO) questionnaire. *International Journal of Human-Computer Interaction*, *30*(9), 685–702.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WebQual: A measure of website quality. *Marketing Theory and Applications*, *13*(3), 432–438.

Lund, A. M. (2001). Measuring usability with the USE questionnaire. *Usability Interface*, *8*(2), 3–6.

McGee, M. (2003). Usability magnitude estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(4), 691–695.

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

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46–55.

Rogers, Y., Sharp, H., & Preece, J. (2011-a). Data gathering. In *Interaction design: Beyond Human-Computer Interaction* (pp. 222–268). Wiley.

Rogers, Y., Sharp, H., & Preece, J. (2011-b). What is interaction design. In *Interaction design: Beyond Human-Computer Interaction* (pp. 1–34). Wiley.

Rohrer, C. P., Wendt, J., Sauro, J., Boyle, F., & Cole, S. (2016). Practical Usability Rating by Experts (PURE): A pragmatic approach for scoring product usability. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 786–795.

Sabukunze, I. D., & Arakaza, A. (2021). User experience analysis on mobile application design using user experience questionnaire. *Indonesian Journal of Information Systems (IJIS)*, 15–26.

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86.

Sauro, J., & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1599–1608.

Sauro, J., & Lewis, J. R. (2016). Standardized usability questionnaires. In *Quantifying the user experience: Practical statistics for user research* (pp. 185–248). Morgan Kaufmann.

Schrepp, M. (2019, August 2). *User experience questionnaire handbook*. User Experience Questionnaire. Retrieved November 1, 2019, from ueq-online.org

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Construction of a benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(4), 40–44.

TED. (2013, June 27). *Julian Treasure: How to speak so that people want to listen* [Video]. YouTube. https://www.youtube.com/watch?v=eIho2S0ZahI

Tullis, T., & Albert, B. (2013). Self-reported metrics. In *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (pp. 121–162). Morgan Kaufmann.

Wang, J., & Senecal, S. (2007). Measuring perceived website usability. *Journal of Internet Commerce*, *6*(4), 97–112.

Zijlstra, F., & van Doorn, L. (1985). *The construction of a scale to measure subjective effort*. Delft University of Technology.

About the Authors

Chase S. Boothe, PhD
Dr. Boothe is the Sr. Director of Product Research at BeyondTrust Corporation where he and his team are responsible for all user-related product research within the organization.

Lesley Strawderman, PhD
Dr. Strawderman is a Professor and International Paper Endowed Chair in the Department of Industrial and Systems Engineering at Mississippi State University.

Reuben F. Burch V, PhD
Dr. Burch is an Associate Professor and Jack Hatcher Chair in the Department of Industrial and Systems Engineering at Mississippi State University as well as the Associate Director of Human Factors & Athlete Engineering at the Center for Advanced Vehicular Systems.

Brian K. Smith, PhD
Dr. Smith is an Associate Professor and Undergraduate Coordinator in the Department of Industrial and Systems Engineering at Mississippi State University.

Cindy L. Bethel, PhD
Dr. Bethel is a Professor in the Computer Science and Engineering Department and holds the Billie J. Ball Endowed Professorship in Engineering at Mississippi State University. Additionally, she is currently an IPA Program Director for the National Science Foundation.

Kate Holmes, PhD
Dr. Holmes is a Lead User Experience Researcher at BeyondTrust where she leads UX research efforts for next generation products.