Assessing Perceived Experience with Magnitude Estimation

Mick McGee[1], Misha Vaughan[2], and Joseph Dumas[3]

[1] EchoUser, San Francisco, USA
mick.mcgee@echouser.com
[2] Oracle Corporation, Applications User Experience, Redwood Shores, USA
misha.vaughan@oracle.com
[3] Dumas Consulting, Yarmouth Port, USA
joe.dumas99@gmail.com

Abstract. Professionals who develop and evaluate the interaction between people and systems have broadened their interests beyond ease of use and learning to higher-order concepts, such as "user experience." "Excellence," "delight" and other emotion-driven experiences are becoming more central to product and company success. In three case studies, we explore and demonstrate how the psychophysical Magnitude Estimation Technique (MET) can be used to quantify complex subjective experiences. We hypothesize that MET can be used to assess *any* user experience that can be defined. We describe studies that apply MET to three different contexts and perceived experience definitions: (1) the riding experience in a public transit system, (2) the effectiveness of a sales presentation, presented online vs. live, and (3) the safety and usability of cancer radiation equipment. In all three situations, participants were able to comprehend the definitions of and assign numeric values to the intensity of their experience. Those judgments were used in combination with other measures to assess the strengths and weaknesses of the overarching user experiences.

Keywords: user experience, usability, magnitude estimation, measurement.

1 Introduction

As technologies mature and business strategies adapt, the demand for differentiating user experience has increased. Making products simple and easy is now the minimum threshold; qualities such as excellence and delight are now expected [1]. Competitive advantage comes from well-designed experiences that have an engaging emotional impact. Synthesizing these engaging experiences into products requires that we enhance the way that we assess such qualities.

Usability measures of task completion rate, assists, time, and errors are still extremely useful and remain a fundamental part of our product development toolkit. Additionally, subjective ratings remain an effective way to assess usability "satisfaction". Our expanded interest goes beyond usability to other aspects of user experience such as 'fun,' 'delight,' and 'engagement,' across devices, formats, and contexts.

A. Marcus (Ed.): DUXU/HCII 2013, Part I, LNCS 8012, pp. 349–358, 2013.

The quantification of the “user experience” construct is still in a formative stage. “The key argument hinges on the meaningfulness, validity and usefulness of reducing fuzzy experiential qualities such as fun, challenge and trust to numbers [2].” Recently there have been a number of studies on the partitioning of user experience into some of its components, especially the interaction between aesthetics and usability [3,4]. These studies have used a variety of traditional measurement techniques, though most have used Likert scales. While Likert scales have the advantages of being both quick to administer and familiar to most end users, they have at least two limitations: they are closed ended, which results and floor and ceiling effects, and their equal-interval scale properties have been questioned [5].

In recent projects, to meet changing business demands and keep users at the center of design processes, we have explored methods to capture experiences and emotion-driven responses. We pursued measures that (1) assess a holistic experience rather than its components, (2) are open ended and have equal interval properties, (3) present the results in an easily interpretable metric, and (4) are flexible enough to accommodate widely differing contexts. We have had success applying a technique from psychophysics, the Magnitude Estimation Technique (MET) [6].

In previous studies, MET has proven flexible and robust enough to scale multifaceted perceptions with complex underlying physical stimuli. This capability is particularly compelling with perceptions that do not have a physical analog, especially when produced from multidimensional stimuli that are difficult to measure. For example, Gescheider [7] cites successful uses of magnitude estimation in a variety of contexts: trial evidence (physical stimulus) with guilt (perception); life events with emotional stress; and psychiatric symptoms with judgments of the severity of mental disorder. McGee [8] demonstrated that MET can measure usability on a variety of platforms: desktop browsers, handheld devices (PDAs and cellular phones), and interactive-voice applications. Rich and McGee [9] further demonstrated that MET is effective at assessing both expectations and actual usability, which can allow practitioners to more meaningfully prioritize usability issues.

2 Our Objective

We set out to explore the hypothesis that MET can be used to measure any novel user experience that can be clearly defined to participants. For example, McGee, Rich and Dumas [10] created an empirical definition of the concept of “usability.” They asked 46 respondents to fill out a survey containing adjectives that described 64 potential usability characteristics. The respondents rated how integral each characteristic was to their concept of usability. A cluster and factor analysis was used to create a definition:

> *Usability is your perception of how consistent, efficient, productive, organized, easy to use, intuitive, and straightforward it is to accomplish tasks within a system.*

Currently in our profession there is no clear context-free definition of “user experience” [2]. Consequently, we wanted to start by sampling a wide variety of contexts in which we could create tailored definitions of an experience that end-users could then use to judge the quality of the experience. We report in this paper on case studies of perceived experience for:

- The quality of the public transit riding experience.
- The effectiveness of a sales presentation using two different formats.
- The safety and usability of medical equipment.

In each of the studies we followed a similar basic MET procedure but with some modifications to explore variations of the method.

3 MET Mechanics

Historically MET began as a way to determine simple relationships between physical intensity, such as decibels, and psychological intensity, such as perceived loudness. Participants in those studies were asked to assign a number to a tone that represented its perceived loudness. Through an initial training session participants were taught to make ratio judgments about loudness. For example, a tone assigned the value of 50 should be perceived as twice as loud as a tone assigned the value of 25 [11]. While each participant is allowed to assign his or her own numbers, magnitude estimation scaling transforms them into a common scale.

As described above, researchers have learned that people can assign reliable judgments to much more complex qualities and experiences. MET can be used within any evaluation that assesses tasks, settings, environments, etc. Participants perform a number assignment procedure across items of interest based on their subjective perception of the defined experience. The outcome is a ratio scale that can be used to make a variety of summary judgments about whatever is being evaluated.

In our three studies, we followed a similar basic procedure. Participants were:

- Given an introduction to the study and that they would be rating an experience.
- Given a definition of the quality of the experience, and provided with an example of a ratio judgment. The definition of quality was developed empirically in some studies and by user experience and subject matter experts in others.
- Given a short reference exercise in which they were asked to assign ratio values to contextually comparable experiences such as the usability of sample web pages.
- Asked to assign ratio values to the primary subject of the study such as the quality of urban transportation experiences, or the quality of a sales presentation, or the usability and safety of medical hardware and software.

These case studies explore magnitude estimation as a metric for measuring users' holistic experience in a variety of different settings and complex user experiences. In each case, we were able to create custom definitions of the quality of experience that participants could understand. Furthermore, by including reference tasks, we were able to assess the quality of experiences in comparison with relevant benchmarks.

4 Case Study I: The Quality of an Urban Transportation Riding Experience

The purpose of this study was to provide the management of the Bay Area Rapid Transit (BART) system with insight into user experience issues through a series of "ride-along" sessions. BART was particularly interested in riders that had a "choice," such as tourists, shoppers, and occasional riders.

We began this research by observing people at BART transit stations and bus stops. The actions of the riders were observed and recorded, especially positive and negative experiences. Once we recorded the main rider activities and notable events, several riders were approached (with a permission letter created by the BART board) for feedback on what characteristics contributed to a positive transit riding experience. The participants either answered brief interview questions or completed a similar survey to that described in [10]. Through that work the following definition of "riding experience" was created:

> *A good public transportation experience is a cost-effective way of reliably, conveniently, and safely getting me to my intended destination on time.*

For the main study, fifteen BART riders were recruited and accompanied through their entire ride. They rode a subway train and/or bus. Upon meeting the participant at their desired location, a facilitator explained the purpose of the study and provided them with training on how to make ratio judgments for the riding experience. They were then given the above definition of rider experience and asked if they understood it or if there were any questions.

For the reference experience, they were given a description of a poor experience that included difficulties buying a ticket and having to stand in a crowded vehicle. They were told to assign that experience a value of 10 and to use that value as a baseline to rate their following ride-along experiences. We used a negative experience because we wanted to set the same lower limit for each rider, which in our experience, makes it easier to use the scale.

We could not control the order of tasks that each participant performed in their specific trip; each rider executed their trip as they normally would. Instead, we had a pre-made set of possible events and, as they occurred, participants provided ratings for that specific event. This "event-based" procedure had seven different possible activities, including waiting for the vehicle (train or bus), riding it, and getting off. Participants were also asked to think aloud to the facilitator about the experience as it was happening. Additional broad questions were asked between events as time permitted, such as rating seating and ride comfort. At the conclusion of the trip, participants made one final rating of the overall experience.

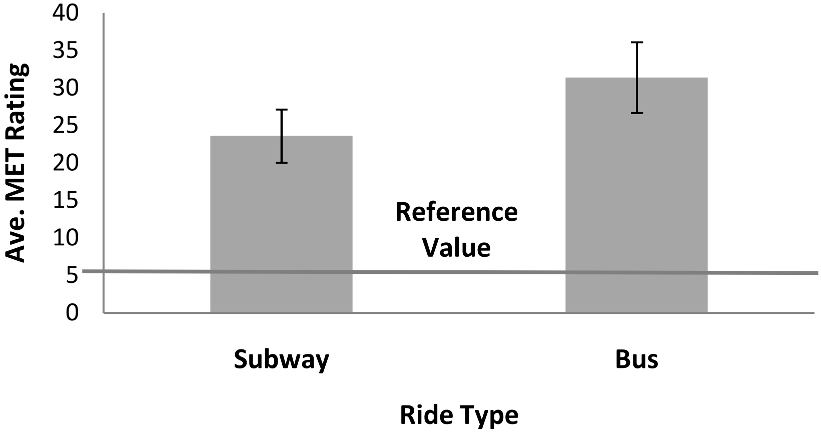

Fig. 1. Average MET ratings for 15 Bart riders

The results of the magnitude estimation ratings of the overall experience for bus and subway rides are shown in Figure 1. The y-axis shows the average rating value provided by the riders.

The scale on that axis had no upper limit as riders where allowed to assign any positive number to their experience. The reference value line represents the poor transit scenario, set at a value of 10. Both the BART bus and subway ride experiences were rated better than the reference experience and the bus rides were rated higher than the subway rides. We did not compute inferential statistics on the averages, but the confidence interval bars indicate a good deal of variability. The think aloud protocol indicated that factors such as the crowding, lack of cleanliness, and higher cost lowered the ratings for the subway.

Participants were also asked to rate their expectations for the experience before each of the events. Figure 2 shows the expected versus observed ratings for four bus ride (AC) events. The results are plotted by scaled expected versus actual user experience per event. For example, the "getting off AC" value shows the average expected usability (x-axis) against the average actual usability (y-axis) for all riders of the event of getting off the bus. The diagonal line (from bottom left to top right) represents where expectations are exactly met. The line from the bottom right to the top left represents where satisfaction is met exactly. The graph shows that the time waiting for the bus fell below expectations (a Fix it issue), but that getting off the bus was well above expectations (a Promote it opportunity). In contrast, settling into the bus seat and riding the bus approximately met expectations (either an opportunity to try and push actual experience above expectations, or information that resource investment may not be worthwhile, compared to other more urgent concerns).

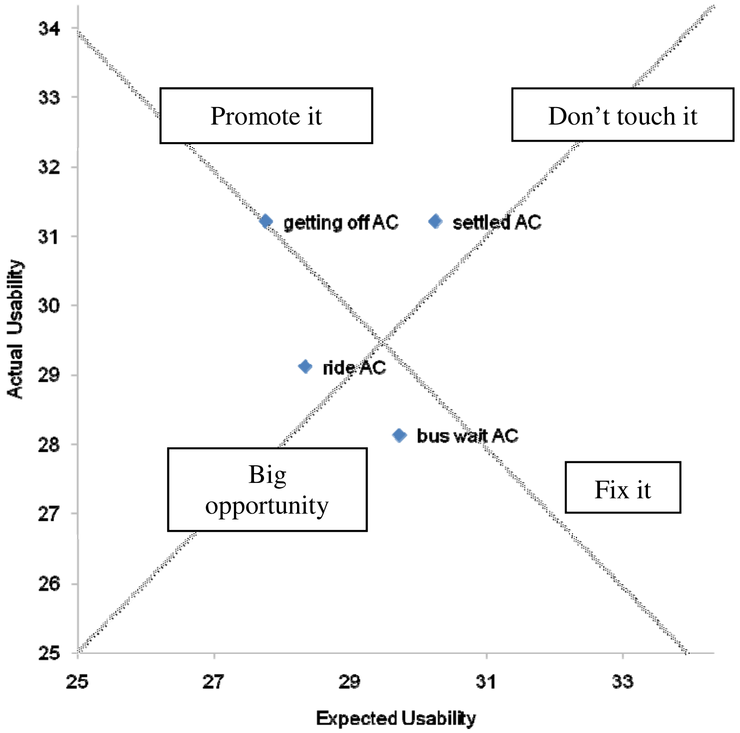

Fig. 2. Bus (AC) user experience expected vs. actual usability

5 Case Study II: The Quality of a Presentation Format

The objective of this study was to determine whether the quality of a presentation about a new product would be perceived as more effective live or in an online video.

We created a presentation about Oracle's next generation of applications. The participants were Oracle customers. Our goal was to make both presentation formats high quality. The presentations were given by the same Oracle VP. The slides for the live version were created with professional graphics and animations. The online presentation was in the form of a movie, which was professional quality, the same presenter's voice and slides.

We ran the session with eight Oracle customers in a group, a new variation of MET. We wanted to see if we could increase the sample size by running group sessions and we wanted to use MET in a setting in which participants normally experience a stimulus in groups. The session began with an introduction to the study and explanation of magnitude estimation. The customers were then asked if they understood the definition and it was discussed further when there were questions.

To establish a baseline for comparison, the participants were given three divergent multimedia scenarios to rate: negative, positive, and neutral:

- Think about an online video that is hard to hear, dimly lit, and boring (negative)
- Think about a podcast that is short, lively, easy to follow, with a clear concept and easy-to-hear voices (positive)
- Think about a blog post that is long and a bit boring, with too many ideas, but with a clear concept and clear and relevant pictures (neutral)

The live presentation was given in a room by the VP. The pre-recorded online movie was then shown projected on a screen. The participants rated the overall quality of each section of the presentation and, at the end, the overall presentation. Because this was a group session, we could not counterbalance the order of the formats.

The results are shown in Figure 3. The three lines represent the average ratings for the baseline definitions. The y-axis is the average quality rating for the eight participants for each format. The scale is open ended so there is no upper limit. Both formats were rated as higher quality than the neutral baseline, which suggests that our goal to create professional presentations was at least partially met. The fact that both formats fell below the positive experience indicates that they could do more to meet customers' expectations.

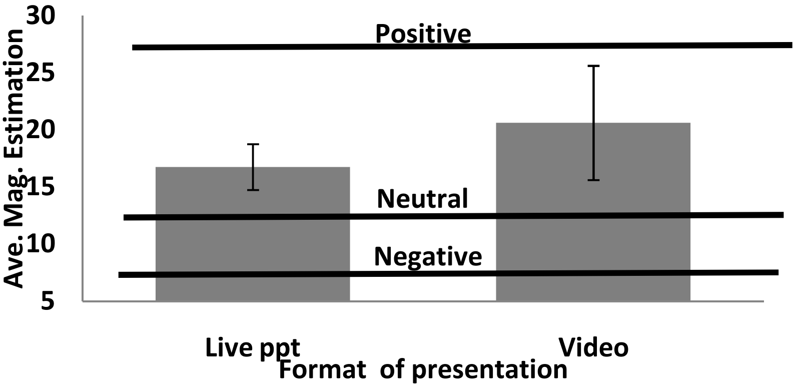

Fig. 3. Average MET ratings for live and video formats by eight participants

Collecting the data in a group setting presented no difficulties. The video presentation format was rated higher than the live presentation in perceived quality, which suggests that that format may be more effective for Oracle customers. The confidence intervals, however, show a large amount of variability. We also realize that we had no independent measure of quality and the live format was seen first. We do see the potential for having two different presenters give the same presentation and measuring the difference in the quality of the experience for the audience.

6 Case Study III: The Perceived Usability and Safety of Medical Equipment

In response to new FDA requirements, a medical equipment manufacturer requested help with creating a method for assessing both the safety and usability of complex hardware and software. Incorporating safety with usability in one unified scale was a challenge. We devised a definition with medical experts:

> *You will be rating user and safety experience. This is your perception of how easy to use, well designed, productive and safe the interface is for conducting tasks. "Safe" is how free an environment (including devices, software, facilities, people, etc.) is from danger, risk, and injury.*

Participants were asked if they understood the definition and if there were any questions. The test participants were physicians and technicians who did not have experience with the models being tested. The participants appeared to perceive safety as critical and integral to the overall 'user experience' and had no issues with the definition.

Reference comparisons for these studies were made against both generic and product specific safety-related tasks, from a common household "safety" task (e.g., using cleaning supplies) to representative tasks related to the products tested (e.g., a patient-management task).

The study consisted of evaluations of two products using this methodology. One product was a hardware medical treatment device; the second was a software application suite of patient management tools.

Comparing MET results between the two products, participants who used the hardware system rated most of the actual task experiences better than they expected, while participants of the software-only test rated the majority of the actual experiences worse than expected. Figures 4 and 5 show the average expected and actual ratings for each task.

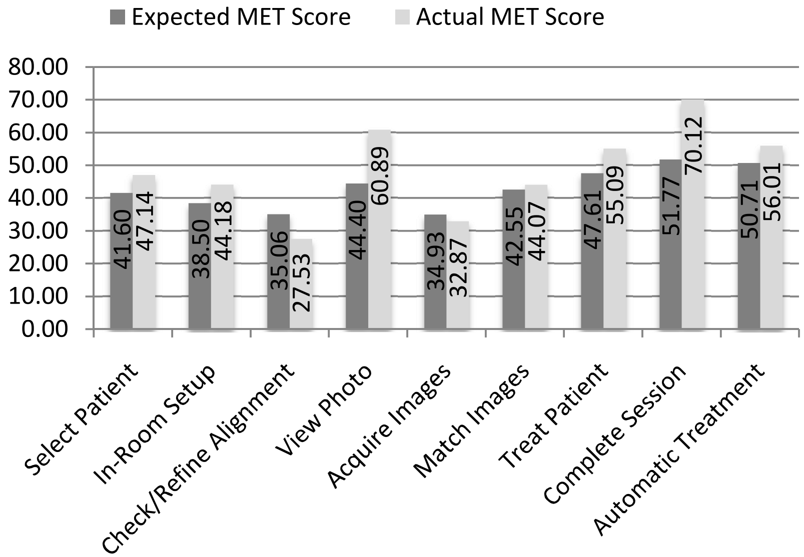

Fig. 4. Expected vs. actual ratings for tasks with the hardware-related product

The rating was open-ended so the y-axis has no upper limit. Notice that in Figure 4, most of the lighter bars (actual rating) are higher than the darker bars (expected rating). In Figure 5 the order is reversed.

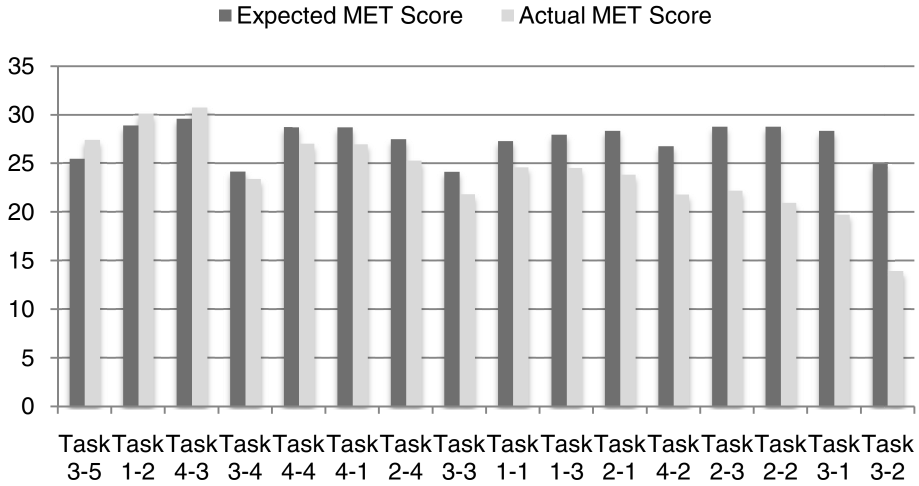

Fig. 5. Expected vs. actual ratings for tasks with the software-related product

Participants using the hardware system did spontaneously comment on safety mitigators that were in place, which they may not have expected. However, issues identified by participants using the software system, which, in our view, had safety concerns far less critical than the hardware system, seemed to focus on traditional usability problems. Perhaps for software-based products, the usability component of the definition becomes dominant. Objective data collected across the two studies did not show the same dichotomy as the holistic, subjective MET results.

Overall, testing for the two products showed successful integration of safety and usability into overall user experience. The results show that even experiences that are not inherently similar can be successfully assessed the same way as long as the measureable construct is the same and the definition of quality is clear to the raters.

7 Discussion

These case studies explored magnitude estimation as a metric for measuring users' holistic experience in a variety of different settings. In each case, we were able to create custom definitions of the quality of experience that participants could understand. Furthermore, by including reference tasks, we were able to assess the quality of experiences in comparison with relevant benchmarks. Additional method variations that we explored included, (1) event-based tasks across an extended transit trip, (2) using MET in a group setting, and (3) combining two perceptual concepts, safety and usability, into one definition.

Once a definition of the quality of the experience was created, the method was relatively easy to administer to participants with a few minutes of training. Participants could make their ratings without interfering with the experience itself. The concept of rating a holistic experience seemed to be easy for participants to grasp. Furthermore, unlike a traditional Likert scale, the MET scale is not bounded at its upper end, avoiding ceiling effects.

Because these case studies were done in industry settings, they were not controlled experiments and they used relatively small samples. They do, however, suggest interesting possibilities. For example, our study of the industry presentations provides a method for comparing the difference in quality of two presenters of the same material, offering the potential to provide feedback to improve presentation skills.

Our study of the perceived usability and safety of medical equipment shows that MET ratings could be added to clinical trials and a single definition can be used for a variety of product types.

Finally, more broadly applicable to the profession, MET provides the opportunity to explore alternative definitions of what is meant by holistic concepts. As we broaden our interest in user experience, MET can provide one tool to help us to begin to quantify that complex and intriguing construct.

References

1. Calacanis, J.: The age of excellence, http://www.launch.co/blog/the-age-of-excellence.html
2. Law, E.: The Measurability and Predictability of User Experience. In: Proceedings of the 3rd ACM SIGCHI Symposium Engineering Interactive Computing Systems EICS 2011, pp. 13–16. ACM, Pisa (2011)
3. Mahlke, S.: Understanding Users' Experience of Interaction. In: EACE 2005 Proceedings of the Annual Conference of the European Association of Cognitive Ergonomics, pp. 251–254 (2005)
4. Hartmann, J.: Assessing the Attractiveness of Interactive Systems. In: Proceedings of CHI 2006, Doctoral Consortium, Montréal, Québec, Canada, pp. 755–1758 (2006)
5. Badia, P., Runyon, R.P.: Fundamentals of Behavioral Research. Random House, New York (1982)
6. Stevens, S.S.: Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. John Wiley, New York (1975)
7. Gescheider, G.A.: Psychophysics: The Fundamentals, 3rd edn. Lawrence Erlbaum Associates, Publishers, Mahwah (1997)
8. McGee, M.: Usability Magnitude Estimation. In: Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, pp. 691–695 (2003)
9. Rich, A., McGee, M.: Expected Usability Magnitude Estimation. In: Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, pp. 912–916 (2004)
10. McGee, M., Rich, A., Dumas, J.: Understanding the Usability Construct: User-perceived Usability. In: Proceedings of Human Factors and Ergonomics Society 48th Annual Meeting, pp. 907–911 (2004)
11. Stevens, S.S.: The Direct Estimation of Sensory Magnitudes—Loudness. The American Journal of Psychology 69, 1–15 (1956)