Mixed Reality Doesn't Need Standardized Evaluation Methods

RICHARD SKARBEZ, La Trobe University, Australia
MARY C. WHITTON, University of North Carolina at Chapel Hill, USA
MISSIE SMITH, Independent Researcher, USA

In this position paper, we argue that *standardized* assessment methods for mixed reality are unachievable and undesirable. In fact, we argue for a future in which there is a greater diversity of purpose-specific measurement tools, rather than increased standardization. However, we recognize the value and encourage the use and development of *standard* evaluation methods, those externally validated by, accepted by, and frequently used by the community.

CCS Concepts: • **Computing methodologies → Mixed / augmented reality**; **Virtual reality**; • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: user experience evaluation

ACM Reference Format:
Richard Skarbez, Mary C. Whitton, and Missie Smith. 2018. Mixed Reality Doesn't Need Standardized Evaluation Methods. 1, 1 (March 2018), 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Let us begin with a clarification. The call for this workshop stated the following: "[O]ur workshop launches a discussion of research methods that should lead to standardizing assessment methods in MR user studies." This is the sense in which we will use the term *standardized* throughout this position paper: As something which should be applied widely, if not universally, as an assessment method. (If this is a misinterpretation, we apologize.) This is in contrast to *standard*, which we will use in the sense of something that is accepted as suitable by the community. We consider standard assessment methods to be an unalloyed good; researchers should, whenever possible, use instruments that have themselves been validated by the research community. However, we argue that standardized assessment methods are impossible to achieve in practice, and undesirable in any event.

2 STANDARDIZED UX EVALUATION IN MIXED REALITY

In this section we present something like a proof by induction. We argue that user experience evaluation in virtual reality is simpler than user experience evaluation in mixed reality more generally, then we argue that even in virtual reality, user experience evaluation is not standardizable. If it is not standardizable in the comparatively simple case, then surely it cannot be standardizable in the more complex general case.

Authors' addresses: Richard Skarbez, r.skarbez@latrobe.edu.au, La Trobe University, Melbourne, Victoria, Australia; Mary C. Whitton, whitton@cs.unc.edu, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; Missie Smith, Independent Researcher, Michigan, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
Manuscript submitted to ACM

2.1 Standardized UX Evaluation in Virtual Reality

We have argued elsewhere that virtual reality should be considered a subset of mixed reality, rather than a separate medium [13]. However, even if one does not accept this argument, it seems clear that the numbers of different application classes and technological implementations that fall under the heading of virtual reality are substantially smaller than the equivalent numbers for mixed reality. If this is true, the identification of suitable standardized assessment techniques should be easier in virtual reality than in mixed reality, if only because the application domain itself is more uniform.

2.1.1 Presence. There is perhaps a perception that for virtual reality, there is such a standardized evaluation metric: *presence*. It may be argued that this is one of the goals of presence measurement; that if we can measure how present a user is in a virtual environment A, and then measure how present that user is in some other virtual environment B, we could simply compare their presence scores and this would tell us whether A or B is the superior virtual environment. That said, the situation isn't nearly as rosy as this description makes it seem.

First, presence can be operationalized in many different ways. (For a recent survey outlining the many different definitions of and measures for presence, see [12].) To summarize, presence can be measured by post-experience questionnaires (of which there are many; the previously mentioned survey outlines fourteen different ones) or by behavioral and physiological means, which are inherently non-standardizable, due to their reliance on specific stimuli occurring in the virtual environment of interest. Per Freeman, Lessiter, and IJsselsteijn: "[C]ontent-dependency makes the development of a general behavioural metric unlikely" [5].

Second, presence is a quale; that is, an internal and individual feeling. Different users will feel different levels of presence in the same virtual environment. Moreover, the same user experiencing the same environment multiple times will likely feel different levels of presence each time. Presence measures, then, must be aggregated to be of much use.

Third, and perhaps most importantly, presence is rarely the key desideratum when designing and implementing a virtual environment. Consider the many different application classes that can be implemented in virtual reality: games and entertainment, psychological treatment, training of surgeons and soldiers, immersive visualization and analytics. Each of these has different criteria by which they should be judged, and it can be argued that presence is not the most important criterion for any of them.

Even in virtual reality, a relatively small and uniform slice of the mixed reality space, there is no standardized measurement. The best contender, presence, is not fit for purpose. For mixed reality more generally, what hope is there?

3 THE SEARCH FOR METRIC X

In the previous section, we argued that there can be no standardized measure of user experience in mixed reality. In this section, we make the case that this isn't necessarily a bad thing. The existence of a single standardized measure that can be broadly applied would tempt us to think that we can meaningfully compare mixed reality experiences that have widely varying technology stacks and intended uses. This, we believe, is a much greater danger to the mixed reality research community than the lack of standardized measurement tools.

Assume for the moment that such a standardized tool does exist; call it *Metric X*. Metric X has all the properties one could want in a measure: It is valid, relevant, sensitive, convenient, nonintrusive, reliable, and objective [7] [9]. Furthermore, it is defined in such a way that it produces meaningful results for any mixed reality application. Question: Is Metric X a good or a bad thing for the mixed reality research community?

We argue that Metric X would, in fact, harm mixed reality research. One reason is that it would enable comparisons between systems that are not productively compared. For example, if Metric X tells us that a virtual reality scenario for training laparoscopic surgery is "better" than an automotive augmented reality head-up display for navigation, what productive conclusions can be drawn from this? Should the community give up on AR HUD research and focus all its energy on surgical simulation, or vice versa? We believe that both projects would likely continue just as they would in the absence of Metric X. This is because they have different technological needs, different user bases, and different purposes, so it only makes sense that their conditions for success would also be different.

The second reason it would harm research is related to the first: It could hinder development of appropriate standard measures that would actually be more productive. Why develop and validate a measure that would only be applicable to a small slice of mixed reality—for example, automotive AR HUDs—when you could simply use Metric X? The answer, we feel, is that such a measure would actually be more useful than Metric X: more predictive, more diagnostic, and more informative.

4 HOPE IS THE THING WITH MEASURES

This, then, is how we would encourage the field to devote its effort. Not to the development of standardized UX measurement tools, which we have argued are improbably difficult to develop and less effective than non-standardized tools. Rather to the identification of constructs of broad interest, the development and validation of instruments to measure those constructs, and the effective communication of such validated instruments.

These constructs and the associated instruments do not have to be specifically created for mixed reality in order to be useful for evaluation of mixed reality experiences. Some will be, certainly: presence [12], Place Illusion/spatial presence [10], Plausibility Illusion [14], and self-presence [1], to name a few. (It is perhaps worth noting that not one of these constructs has a validated measurement instrument that we would consider "standard" as of this writing.) Many others will come from other areas, most notably psychology: these include co-presence [6], social presence [11], sense of embodiment [8], body ownership [4], and flow [3]. We can imagine many others that would apply in specific application domains. For safety-critical environments such as driving, piloting, and surgery, constructs such as *safety*, *perceived safety*, and *attention/distraction* might be most important. For entertainment and gaming applications, perhaps *engagement* and *enjoyment*. For a psychological therapy application, *stressfulness* (for the patient), and *controllability* (for the clinician). All of this is to say that our vision for the future of UX evaluation in mixed reality is actually an increased diversity of purpose-specific measurement tools, rather than an increase in standardization.

5 CONCLUSION

We agree that, whenever possible, researchers should use standard measures rather than creating their own. However, this does not mean that the evaluation itself should be standardized. We clarify with an example: If one thinks that the level of social presence created by interaction with an mixed reality application is of interest, one should absolutely measure social presence using a widely-used instrument, such as the Networked Minds questionnaire [2], rather than an ad-hoc one. However, one should not take this to mean that social presence is a construct relevant to every mixed reality experience, nor should one assume that an experience that elicits more social presence than another is superior. The constructs that determine the effectiveness of a given application or experience are intrinsically linked to its purpose.

REFERENCES

[1] Frank Biocca. 1997. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication* 3, 2 (1997). https://doi.org/10.1111/j.1083-6101.1997.tb00070.x

[2] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *PRESENCE 2001 - 4th Annual International Workshop on Presence*. Philadelphia, PA, 1–9.

[3] Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience.* Harper and Row, New York.

[4] Frédérique de Vignemont. 2011. Embodiment, ownership and disownership. *Consciousness and cognition* 20, 1 (2011), 82–93.

[5] Jonathan Freeman, Jane Lessiter, and Wijnand IJsselsteijn. 2001. An introduction to presence: A sense of being there in a mediated environment. *The Psychologist* 14 (2001), 190–194.

[6] Erving Goffman. 1963. *Behavior in public places: Notes on the social organization of gatherings.* The Free Press, New York.

[7] Claudia Hendrix and Woodrow Barfield. 1996. Presence within virtual environments as a function of visual display parameters. *Presence: Teleoperators and Virtual Environments* 5, 3 (Summer 1996), 274–289.

[8] Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments* 21, 4 (September 2012), 373–387. https://doi.org/10.1162/PRES{_}a{_}00124

[9] Michael John Meehan. 2001. *Physiological reaction as an objective measure of presence in virtual environments.* Ph.D. Dissertation. The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[10] Thomas W. Schubert. 2009. A new conception of spatial presence: Once again, with feeling. *Communication Theory* 19, 2 (2009), 161–187. https://doi.org/10.1111/j.1468-2885.2009.01340.x

[11] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications.* John Wiley & Sons, Ltd., Hoboken, NJ.

[12] Richard Skarbez, Frederick P. Brooks, Jr., and Mary C. Whitton. 2017. A Survey of Presence and Related Concepts. *Comput. Surveys* 50, 6, Article 96 (2017), 39 pages. https://doi.org/10.1145/3134301

[13] Richard Skarbez, Missie Smith, and Mary C. Whitton. 2021. Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality* 2 (March 2021), 647997:1–647997:8.

[14] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behavior in immersive virtual environments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364 (2009), 3549–3557.