

# A Framework for Usability Evaluation of Mobile Mixed Reality Games

Charley Gielkens<sup>1,2</sup> and Richard Wetzel<sup>2</sup>

<sup>1</sup> Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands,  
[charley.gielkens@gmail.com](mailto:charley.gielkens@gmail.com),

<sup>2</sup> Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

**Abstract.** This research presents a framework that supports usability experts in determining which method to use when evaluating Mobile Mixed Reality Games (MMRGs). These are games that combine the real and virtual world by means of e.g. a smartphone and require the player to change their geographical location. As some different styles of MMRGs exist, e.g. running versus cunning or multiplayer versus single player, not every method is suitable for each style. The results of the methods are benchmarked against a heuristic evaluation and it is shown that using Instant Data Analysis (IDA), Diary, interaction logs combined with audio diary and retrospective think aloud combined with IDA perform statistically comparable, but that the latter is not favorable based on qualitative merits.

**Keywords:** mixed reality, augmented reality, games, usability evaluation

## 1 Introduction

A game that mixes the real and virtual world to some degree is called a Mixed Reality Game [7]. If this is done using a mobile device such a game is called a Mobile Mixed Reality Game (MMRG). If the player has to change their location, the game is considered a truly mobile MMRG [13]. Mixed reality is interpreted rather freely, so that it does not only include Augmented Reality by means of a magic lens [1] but also by means of a map. Such games are also known as location based games.

As MMRGs are migrating from research projects [6, 12] to consumer products, they are also migrating from specialized hardware to common hardware like smartphones, i.e. phones with advanced sensors. As the input and output capabilities on these devices is limited compared to a static desktop computer, a high degree of usability is of even greater importance. In current literature little is known about how to measure usability of MMRGs. For both games in general [2, 9, 10] and MMRGs [4] heuristic evaluation is a tried and tested method. In [4] a definition for the usability of MMRGs is given, which will also be used here: “usability for MMRGs is the degree to which a player is able to learn, control, understand and safely play a game in the environment it was designed for.”

## 2 Method

After studying both the Android Marketplace [4] for games and literature on usability evaluation methods [4], it was surmised that different styles of games exist and postulated that each had usability evaluation methods that were better suited to their needs. Four methods were identified that merited testing for certain styles of games which are described using questions. Their relation is visualized in figure 1. As the focus of this study are MMRGs in a context sensitive manner, the evaluation of mock ups (**question 1**) is not further investigated.

**Question 2: game length** If a game is played over a longer period of time, the interface is most likely also complicated. In such a case it is advantageous if the player can also actually play it over a longer period of time. Therefore, using a diary is suggested. [11]

**Question 3: amount of players** If more players are needed to play the game than there are experts available to observe them, direct observation becomes impossible. In this case it is suggested to use interaction logs (IL) and a spoken diary. The latter means that the participants vocalize their likes and dislikes, but not every single thought like in a think aloud (TA) [3] session.

**Question 4: session length** When enough experts are available to observe players, but a play session takes a long time, e.g. over an hour, one may want to consider not having the experts present all the time as this is expensive. In order to make evaluating such a session easier, it is suggested to use IL and a spoken diary again.

**Question 5: physical activity** If you have enough experts and the game doesn't last too long, e.g. less than an hour, the last thing to take into account is the amount of physical activity required to play the game. If one needs to run a lot, performing a concurrent TA becomes very difficult both for the participant as for the expert. It was therefore suggested to use retrospective TA and IDA [5] when a lot physical activity is involved and just IDA when it isn't.

To validate the suggested methods they are applied to relevant games which are also evaluated using heuristics for MMRGs [4]. For each list all issues will be rated on severity using the five-point scale introduced in [8]. The median severity ratings were compared between methods, within games using a Mann-Whitney test and the amount of issues were compared within games, between methods using a  $\chi^2$ -test.

For the diary study Parallel Kingdom (<http://www.parallelkingdom.com/>) was used, for the slow paced game Tidy City [14] was chosen and for the other options Portal Hunt (<http://www.totem-games.org/?q=portalhunt>) was selected.

## 3 Results

In table 1 the amount of issues that have been identified using each method is shown per game, method and severity rating. For IDA ( $Mdn = 1$ ) the median severity was significantly different from the heuristic evaluation ( $Mdn = 2, U =$

Severity Heuristics Diary			Severity Heuristics RTA/IDA IL/AD				Severity Heuristics IDA		
0	0	0	0	0	1	0	0	0	10
1	6	7	1	0	5	3	1	4	17
2	11	15	2	19	13	11	2	17	12
3	17	16	3	18	11	32	3	9	4
4	2	3	4	13	9	10	4	1	0
Total:	36	41	Total:	50	39	56	Total:	31	43
Median:	3	2	Median:	3	3	3	Median:	2	1

(a) Parallel Kingdom

Table 1: Issues per severity rating, per game, per method

(b) Portal Hunt

(c) Tidy City

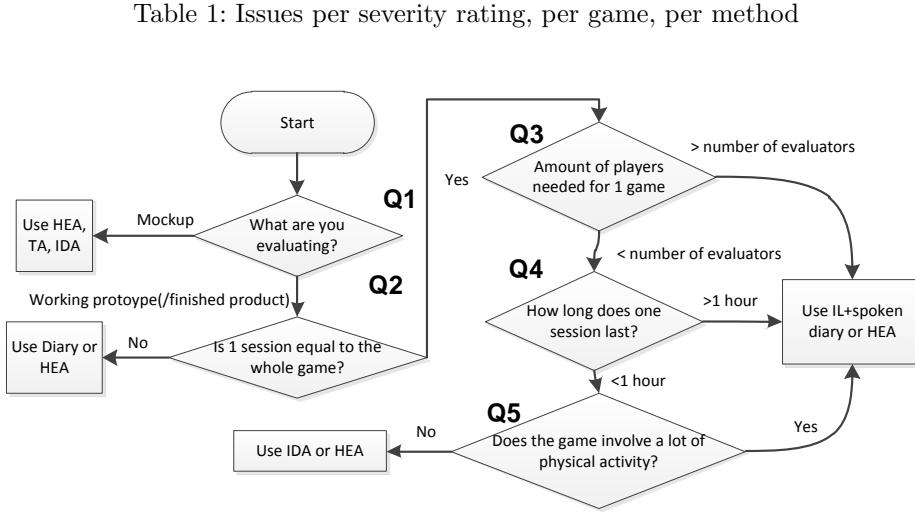


Fig. 1: ARGUEMENT – Augmented Reality Usability Evaluation Method Elec-tioN Tool

286.00,  $z = -4.372, p < .01$ ), but not for the other methods. The amount of issues did not differ significantly in any case.

## 4 Discussion

Notwithstanding the lack of significant differences, the extra effort that's required for RTA combined with IDA makes a strong case for replacing this method by interaction logs combined with audio diary as that methods resulted in more issues being identified with similar effort. The final version of the framework is visualized in figure 1.

Some limitations should be taken into account when reading this. E.g. the limited amount of participants, the use of the same usability expert in multiple evaluations and the fact that the results presented here are all based on a single observation. This means that further validation is required to make sure the results presented here are completely reliable.

**Acknowledgements** The authors would like to thank Herre van Oostendorp, Christof van Nimwegen, their colleagues of the Mixed and Augmented Reality Solutions department of Fraunhofer FIT and TOTEM, the German-French project on Theories and Tools for Distributed Authoring of Mobile Mixed Reality Games under the Programme Inter Carnot Fraunhofer from BMBF (Grant 01SF0804) and ANR (<http://www.totem-games.org>).

## References

1. Bier, E.A., Stone, M.C., Pier, K., Buxton, W., DeRose, T.D.: Toolglass and magic lenses. In: Proceedings of the 20th annual conference on Computer graphics and interactive techniques - SIGGRAPH '93, New York, New York, USA, ACM Press (September 1993) 73–80
2. Desurvire, H., Caplan, M., Toth, J.A.: Using heuristics to evaluate the playability of games. In: Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04, New York, New York, USA, ACM Press (April 2004) 1509–1512
3. Ericsson, K., Simon, H.: Protocol analysis : verbal reports as data. 1st paperb edn. The MIT Press, Cambridge Mass. (1985)
4. Gielkens, C.: A Framework for Usability Evaluation of Mobile Mixed Reality Games. Master, Utrecht University (2012) Available from [http://igitur-archive.library.uu.nl/student-theses/2012-0607-200545/thesis\\_charleygielkens\\_final.pdf](http://igitur-archive.library.uu.nl/student-theses/2012-0607-200545/thesis_charleygielkens_final.pdf).
5. Kjeldskov, J., Stage, J.: New techniques for usability evaluation of mobile systems. International Journal of Human-Computer Studies **60**(5-6) (May 2004) 599–620
6. McCall, R., Wetzel, R., Löschner, J., Braun, A.: Using presence to evaluate an augmented reality location aware game. Personal and Ubiquitous Computing **15**(1) (June 2010) 25–35
7. Milgram, P., Kishino, F.: A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems E series D **77**(12) (1994) 1321–1321
8. Nielsen, J.: Severity Ratings for Usability Problems (1995)
9. Pinelle, D., Wong, N., Stach, T.: Heuristic evaluation for games. In: Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08. CHI '08, New York, New York, USA, ACM Press (2008) 1453–1462
10. Pinelle, D., Wong, N., Stach, T., Gutwin, C.: Usability heuristics for networked multiplayer games. In: Proceedings of the ACM 2009 international conference on Supporting group work - GROUP '09, New York, New York, USA, ACM Press (May 2009) 169–178
11. Sharp, H., Rogers, Y., Preece, J.: Interaction Design: Beyond Human-Computer Interaction. Wiley (2007)
12. Thomas, B., Close, B., Donoghue, J., Squires, J., Bondi, P.D., PiekarSKI, W.: First Person Indoor/Outdoor Augmented Reality Application: ARQuake. Personal and Ubiquitous Computing **6**(1) (February 2002) 75–86
13. Wetzel, R., Blum, L., Broll, W., Oppermann, L.: Designing mobile augmented reality games. In Furht, B., ed.: Handbook of Augmented Reality. 1st edn. Springer (2011) 513–529
14. Wetzel, R., Blum, L., Feng, F., Oppermann, L., Straeubig, M.: Tidy City: A Location-based Game for City Exploration Based on User-created Content. In: Proceedings of Mensch & Computer 2011, Chemnitz, Germany (2011) 487–498