

VocabulARy replicated: comparing teenagers to young adults

Maheshya Weerasinghe*
University of St Andrews and
University of Primorska

Verena Biener†
Jens Grubert‡
Coburg University of
Applied Sciences

Jordan Aiko Deja§
Nuwan T. Attygalle¶
Karolina Trajkovska||
University of Primorska

Matjaž Kljun**
Klen Čopič Pucihar††
University of Primorska and
Faculty of Information Studies Novo mesto

ABSTRACT

A critical component of user studies is gaining access to a representative sample of the population researchers intend to investigate. Nevertheless, the vast majority of human-computer interaction (HCI) studies, including augmented reality (AR) studies, rely on convenience sampling. The outcomes of these studies are often based on results obtained from university students aged between 19 and 26 years. In order to investigate how the results from one of our studies are affected by convenience sampling, we replicated the AR-supported language learning study called VocabulARy with 24 teenagers, aged between 14 and 19 years. The results verified most of the outcomes from the original study. In addition, it also revealed that teenagers found learning significantly less mentally demanding compared to young adults, and completed the study in a significantly shorter time. All this at no cost to learning outcomes.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Treemaps; Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Reproducibility of user studies is a well known problem affecting different scientific fields. For example, an attempt to replicate 100 studies in psychology, published within a year in three high-ranking psychology journals, revealed that only 36% of results were replicated successfully [2]. Furthermore, a survey including scientists from various disciplines (chemistry, biology, physics/engineering, medicine, earth and environment sciences, etc.) found that the vast majority of them (64%-87%) reported having problems replicating results from other studies [3]. Another problem is the lack of replication studies. A paper in the field of human-computer interaction (HCI) reports that only 3% out of 891 studies attempted to replicate an earlier result [4]. Within augmented, mixed and virtual reality (AR, MR, VR) research such studies are even more scarce [7] with only a few examples available, such as [1, 6].

To conduct a robust user study, a careful consideration must be given to a representative sample of the population researchers intend to investigate. However, the vast majority of HCI studies rely on convenience sampling. As a result, the conclusions made in the literature are often based on the university students, aged 19 through 30. This is also one of the recognised limitations of one of our previous user studies called VocabulARy [8]. To investigate the effect of the convenience sampling on the results of the aforementioned study,

we decided to run a replication study that targeted a different age group, aged 14 through 19.

VocabulARy is an AR system for learning words in a foreign language. For our studies we selected Japanese as a language to be learnt because it is uncommonly spoken by speakers of Indo-European languages. The prototype displays visual and audio AR annotations for objects in users' surroundings. For each object users can see two words – an English word (first language) and Japanese translation (second language) – and the audio pronunciation of the latter. In addition, the prototype displays a keyword and its visualisation to enhance memory retention. In our studies we compare the AR system to the non-AR tablet computer and on each we compare keywords to keywords together with its visual representation.

2 USER STUDY

The study was conducted following the research method described in detail in the original paper [8]. The summary of the procedure and differences are explained hereafter.

2.1 Participants and procedure

The study was completed by 24 participants aged 14 through 19 ($\bar{x} = 15.8$, $SD = 1.5$). Half of them (12) participated in the AR (5 female) and half in the NON-AR condition (6 female). As we opted for a mixed design study, the between subject factor (i.e. AR and NON-AR) could be studied at two different locations. The NON-AR was studied at a scout camp, and the AR at the university as a part of a summer school. Since the participants came from different schools and different parts of the country, they represent a more varied sample compared to recruiting participants from one school or area only. All participants voluntarily took part in the study and the consent forms were acquired from their parents or legal guardians if they were younger than 18.

At each location, we randomly selected the instruction mode to be used first (KEYWORD vs KEYWORD+VISUALISATION). Finally, the learning scenario was randomly selected (kitchen or office environment). All randomisations were counterbalanced. After the training session the participants were asked to remember 10 Japanese words in the learning scenario given.

2.2 Differences between the replication and original study

Besides the age group (young adults 19-30 vs teenagers 14-19) and the sampling method (convenience sampling among computer science students from one university vs sampling from a more varied group of high-school students), there were two other differences. In the replication study we did not capture the delayed recall data, and the maximal available time for NON-AR condition was reduced from 15 to 5 minutes due to organisational limitations set by camp organisers.

3 RESULTS

All participants managed to successfully complete the study. The dependent variables (Immediate Recall (how many words could be successfully recalled after the study), Mental Effort (measured with

*e-mail: amw31@st-andrews.ac.uk

†e-mail: verena.biener@hs-coburg.de

‡e-mail: jens.grubert@hs-coburg.de

§e-mail: jordan.deja@famnit.upr.si

¶e-mail: nuwan.attygalle@famnit.upr.si

||e-mail: 89191037@student.upr.si

**e-mail: matjaz.kljun@upr.si

††e-mail: klen.copic@famnit.upr.si

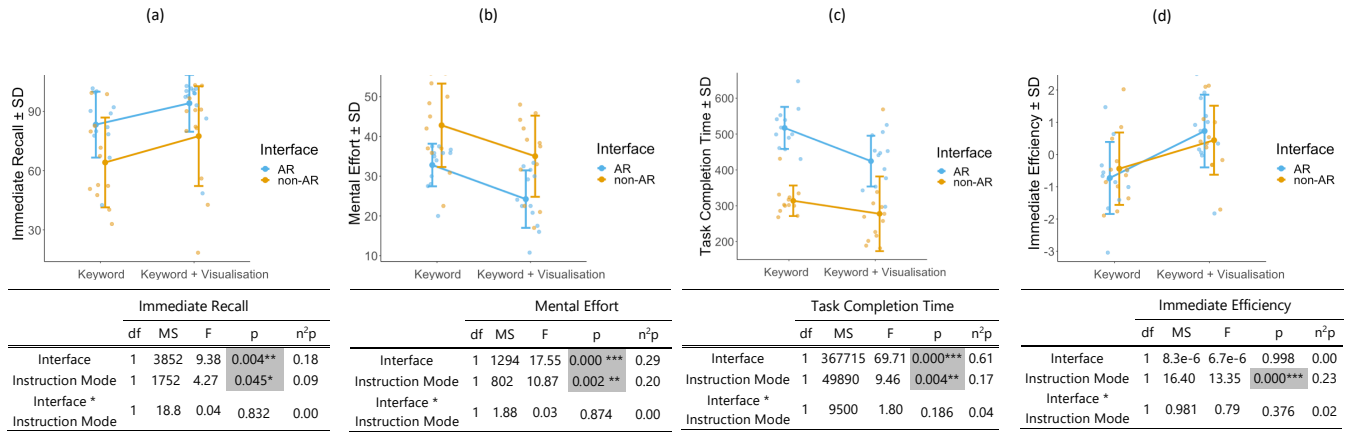


Figure 1: Means with standard deviation and ANOVA results for: (a) immediate recall performance in percentage of correctly remembered words; (b) mental effort invested during the study; (c) task-completion-time in seconds; (d) learning efficiency.

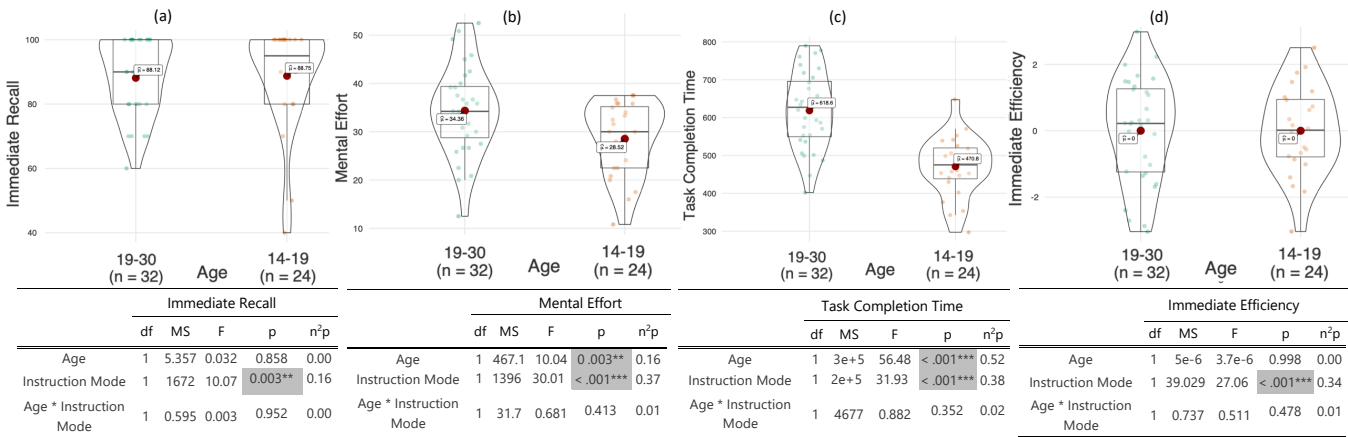


Figure 2: Means with standard deviation and ANOVA results for AR condition analysing the effect of age. This analysis includes data from the original and replication study for: (a) immediate recall performance in percentage of correctly remembered words; (b) mental effort invested during the study; (c) task-completion-time in seconds; (d) learning efficiency.

the NASA-TLX questionnaire), Task Completion Time (time needed to learn all the words), Learning Efficiency (the ratio of performance to the difficulty of the learning task)) as well as all statistical analysis carried out are described in detail in the original paper [8].

We conducted a mixed design analysis with 2 AGE (14-19 vs 19-30) x 2 INTERFACE (NON-AR vs AR) x 2 INSTRUCTION MODE (KEYWORD vs KEYWORD+VISUALISATION) conditions. AGE and INTERFACE conditions were analysed as between-subject factors, all others were analysed as within-subject factors. The results are organised according to dependent variables and divided into *Replication study* and *Replication and original study* subsections.

3.1 Immediate Recall

Replication study: The mean values of immediate recall and the ANOVA results for the INTERFACE (AR and NON-AR) and the INSTRUCTION MODE (KEYWORD and KEYWORD+VISUALISATION) conditions are shown in Figure 1a.

A significant main effect of the INTERFACE on immediate recall could be detected ($F(1,44) = 9.38$, $p < 0.05$, $\eta^2p = 0.18$). Immediate recall scores were significantly better in AR condition ($\bar{x} = 88.8\%$, $SD = 24.48$) compared to the NON-AR condition ($\bar{x} = 70.8\%$, $SD = 16.24$).

Also, a significant main effect of INSTRUCTION MODE on

immediate recall performance could be detected ($F(1,44) = 4.27$, $p < 0.05$, $\eta^2p = 0.09$). Immediate recall scores in KEYWORD+VISUALISATION condition ($\bar{x} = 85.8\%$, $SD = 21.85$) were significantly better than in KEYWORD condition ($\bar{x} = 73.8\%$, $SD = 21.83$). No significant effect could be found between the INTERFACE and INSTRUCTION MODE ($F(1,44) = 0.05$, $p > 0.05$, $\eta^2p < 0.001$).

Replication and original study for AR: The distribution of the data and the ANOVA results for immediate recall focusing on the effect of AGE are shown in Figure 2a. Statistical analysis showed no significant effect of the AGE on participants' immediate recall ($F(1,52) = 0.032$, $p > 0.05$, $\eta^2p < 0.001$). Also, no significant interaction effect could be found between the AGE and the INSTRUCTION MODE conditions ($F(1,52) = 0.003$, $p > 0.05$, $\eta^2p < 0.001$).

3.2 Mental Effort

Replication study: The mean values of the mental effort experienced during the task and the ANOVA results are shown in Figure 1b.

A significant effect of the INTERFACE on mental effort could be detected ($F(1,1) = 17.54$, $p < 0.001$, $\eta^2p = 0.26$). It was significantly lower for the AR ($\bar{x} = 28.5$, $SD = 10.87$) compared to the NON-AR condition ($\bar{x} = 38.9$, $SD = 7.61$). Also, a significant effect of the INSTRUCTION MODE on mental effort could be detected ($F(1,1) = 10.87$, $p < 0.01$, $\eta^2p = 0.29$). In KEY-

WORD+VISUALISATION ($\bar{x} = 29.6$, $SD = 10.25$) it was significantly lower than in KEYWORD condition ($\bar{x} = 37.8$, $SD = 9.61$). No significant effects were found between the INTERFACE and INSTRUCTION MODE ($F(1, 1) = 0.03$, $p > 0.05$, $n^2p < 0.001$).

Replication and original study for AR: The distribution of the data and the ANOVA results for mental effort focusing on the effect of AGE are shown in Figure 2b. A significant effect of the AGE on mental effort could be detected ($F(1, 52) = 10.04$, $p < 0.05$, $n^2p = 0.16$). Mental effort was significantly lower in the 14-19 age group ($\bar{x} = 28.5$, $SD = 7.61$) compared to the 19-30 age group ($\bar{x} = 34.4$, $SD = 9.17$). However, no significant interaction effect could be found between the AGE and the INSTRUCTION MODE ($F(1, 52) = 0.681$, $p > 0.05$, $n^2p = 0.01$).

3.3 Task Completion Time

Replication study: The mean values of task completion time for all study conditions are shown in Figure 1c. The data is analysed using a between-within subjects ANOVA on the 20% trimmed means [5].

A significant effect of the INTERFACE on task completion time could be detected ($F(1, 1) = 69.71$, $p < 0.001$, $n^2p = 0.61$). The completion time was significantly lower for the NON-AR condition ($\bar{x} = 296s$, $SD = 80s$) compared to the AR condition ($\bar{x} = 475s$, $SD = 79s$).

Also, a significant effect of INSTRUCTION MODE on task completion time could be detected ($F(1, 1) = 9.46$, $p < 0.01$, $n^2p = 0.17$). The KEYWORD+VISUALISATION ($\bar{x} = 351s$, $SD = 110s$) resulted in a significantly lower completion time than KEYWORD ($\bar{x} = 416s$, $SD = 115s$). There was no significant effect between the INTERFACE and INSTRUCTION MODE ($F(1, 1) = 1.80$, $p > 0.05$, $n^2p = 0.04$).

Replication and original study for AR: The distribution of the data and the ANOVA results for task completion time focusing on the effect of AGE are shown in Figure 2c. A significant effect of AGE on task completion time could be detected ($F(1, 52) = 56.48$, $p < 0.001$, $n^2p = 0.52$). The task completion time was significantly lower for 14-19 age group ($\bar{x} = 470.8$, $SD = 79.24$) than the 19-30 age group ($\bar{x} = 618.6$, $SD = 101.11$). No significant effect could be found between the AGE and INSTRUCTION MODE ($F(1, 52) = 0.882$, $p > 0.05$, $n^2p = 0.02$) conditions.

3.4 Learning Efficiency

Replication study: The average learning efficiency for immediate recall across study conditions are shown in Figure 1d. The data is analysed using a between-within subjects ANOVA on the 20% trimmed means [5].

Statistical analysis showed no significant effect of the INTERFACE on learning efficiency for immediate recall ($F(1, 1) = 6.8e - 6$, $p > 0.05$, $n^2p < 0.001$). A significant effect of the INSTRUCTION MODE condition on learning efficiency for immediate recall could be detected ($F(1, 1) = 13.35$, $p < 0.001$, $n^2p = 0.23$). The learning efficiency was significantly higher in KEYWORD+VISUALISATION ($\bar{x} = 0.585$, $SD = 1.08$) compared to the KEYWORD ($\bar{x} = -0.584$, $SD = 1.10$) condition. There was no significant effect between the INTERFACE and INSTRUCTION MODE for immediate recall ($F(1, 1) = 0.79$, $p > 0.05$, $n^2p < 0.05$).

Replication and original study for AR: The distribution of the data and the ANOVA results for learning efficiency focusing on the effect of AGE are shown in Figure 2d. Statistical analysis showed no significant effect of AGE for learning efficiency for immediate recall ($F(1, 52) < 0.001$, $p > 0.05$, $n^2p < 0.001$). Also, no significant interaction effects could be found between the AGE and INSTRUCTION MODE ($F(1, 52) = 0.511$, $p > 0.05$, $n^2p = 0.01$) conditions.

4 DISCUSSION AND CONCLUSION

A comparison of replication and original study results shows that there is little difference between the two. The results for statistical tests for the dependent variables such as immediate recall, mental

effort and immediate efficiency lead to the same conclusions. However, this was not the case for the time taken to complete the task. The results of both studies agree on the effect of INSTRUCTION MODE, whilst show the opposite in case of INTERFACE condition. For this condition the completion time is significantly lower for the NON-AR compared to the AR condition. This is probably due to different time constraint for the NON-AR and AR conditions.

Furthermore, there is an observable difference in the significance levels that were detected for all the variables. As the sample size in the replication study was substantially smaller (i.e. $n = 24$ vs. $n = 32$ in the original study) one would expect higher or similar p-values. This was indeed observed for all p-values except for the p-value of Immediate Recall for INTERFACE condition (i.e. $p = 0.004$ vs. $p = 0.01$ in original study). We hypothesise that the NON-AR condition in the replication study was tested outside the laboratory where the researchers did not have a complete control over the environment. Thus, various disruptions could occur, such as noise, people walking into the room, the presence of other observers. In addition, it is important to note that the replication study did not capture data for delayed recall, thus this part was not presented here.

Finally, the results analysing the effect of AGE condition showed that teenagers found the study significantly less mentally demanding and completed it in a significantly shorter time also in AR that had the same time constraints as the original study. However, despite overall better performance in immediate recall and learning efficiency, no significance was detected. Why this is the case remains to be answered.

ACKNOWLEDGMENTS

Authors would like to thank Nikola Kovačević, Nina Chiarelli and Ana Zalokar for their help with the study. This research was supported by European Commission through the InnoRenew CoE project (Grant Agreement 739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund). The work was also supported by the Slovenian research agency (program no. BI-DE/20-21-002, P1-0383, J1-9186, J1-1715, J5-1796, and J1-1692).

REFERENCES

- [1] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard, and J. E. Swan. Impact of ar display context switching and focal distance switching on human performance: Replication on an ar haploscope. In *In proc. of 3DUI 2020*, pp. 571–572, 2020. doi: 10.1109/VRW50115.2020.00137
- [2] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716
- [3] D. Henderson and K. Thomson. What makes science true? <http://www.pbs.org/wgbh/nova/body/reproduce-science.html>. Accessed: 2022-08-27.
- [4] K. Hornbæk, S. S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen. Is once enough? on the extent and content of replications in human-computer interaction. In *In proc. of CHI 2014*, CHI '14, p. 3523–3532. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2556288.2557004
- [5] P. Mair and R. Wilcox. Robust statistical methods in r using the wrs2 package. *Behavior research methods*, 52(2):464–488, 2020.
- [6] P. Mohr, M. Tatzgern, J. Grubert, D. Schmalstieg, and D. Kalkofen. Adaptive user perspective rendering for handheld augmented reality. In *In proc. of 3DUI 2017*, pp. 176–181. IEEE, 2017.
- [7] J. E. Swan. The replication crisis in empirical science: Implications for human subject research in mixed reality. In *In proc. of ISMAR-Adjunct 2018*, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00019
- [8] M. Weerasinghe, V. Biener, J. Grubert, A. J. Quigley, A. Toniolo, K. Č. Pucihar, and M. Kljun. Vocabulary: Learning vocabulary in ar supported by keyword visualisations. *arXiv:2207.00896*, 2022.