Raphael Zender et al. (Hrsg.): Die 18. Fachtagung Bildungstechnologien (DELFI),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 55

Comparing Virtual Reality and Screen-based Training Simulations in Terms of Learning and Recalling Declarative Knowledge

Leonard Meyer[1], Thies Pfeiffer[2]

Abstract: This paper discusses how much the more realistic user interaction in a life-sized fully immersive VR Training is a benefit for acquiring declarative knowledge compared to the same training via a screen-based first-person application. Two groups performed a nursing training scenario in immersive VR and on a tablet. A third group learned the necessary steps using a classic text-picture-manual (TP group). Afterwards all three groups had to perform a recall test with repeated measurement (one week). The results showed no significant differences between VR training and tablet training. In the first test shortly after completion of the training both training simulation conditions were worse than the TP group. In the long-term test, however, the knowledge loss of the TP group was significantly higher than that of the two simulation groups. Ultimately, VR training in this study design proved to be as efficient as training on a tablet for declarative knowledge acquisition. Nevertheless, it is possible that acquired procedural knowledge distinguishes VR training from the screen-based application.

Keywords: Virtual Reality Training; Screen Based Training; Declarative knowledge acquiring; Health Care Scenario

1 Introduction

In recent years, virtual reality has found its way into the global consumer market. Commercial VR headsets are widely used for entertainment purposes whose underlying applications are exploited by developers in all directions of new interaction possibilities. VR technologies allow the user to see and interact with virtual environments from a first person fully immersive view. They elevate virtual worlds, as they have been known for decades from simulations, video games and videos, to a new level of user experience. But this modern technology does not have to serve only for entertainment, there is a high potential of new concepts for digital learning media, of which various implementations have already been examined for effectiveness in education and teaching in the past years.

VR offers a wide range of features that could be useful for teaching. Any imaginable 3D environment can be displayed independent of the real available space. Users can interact freely in virtual space and can receive direct audio, visual and even haptic feedback. On the other hand, various possibilities of virtual reality are also available in other media such as videos or even closer in screen-based simulations. The sticking point that distinguishes

[1] Hochschule Emden/Leer, Abteilung E+I, Constantiaplatz 4, 26723 Emden, leonard.meyer@hs-emden-leer.de,
[2] Hochschule Emden/Leer, Abteilung E+I, Constantiaplatz 4, 26723 Emden, thies.pfeiffer@hs-emden-leer.de

the different media is the way of learning. While videos are passive learning objects and therefore only allow displaying information, simulations of all kinds allow an active interaction with the environment. What this interaction looks like can make a difference in learning success, which makes it part of a wide area in educational research considering different types of learning. In general, the concept of distinct learning types is very controversial, with several models, such as the visual-auditory-kinaesthetic learning styles model [BMS88]. Still there are differences in the way of learning and in the success of different teaching methods, but these differences are difficult to integrate into clear structures. Nevertheless, VR offers a wider range of interaction (and thus information) channels compared to screen-based simulations, especially on realistic training movements in life-sized environments. Therefore a decisive factor, which should positively distinguish the learning success in VR simulations from other methods, is the self-performing of a training or a procedure by the user, even if done virtually with the limitations of the respective system (lack of haptic feedback, controller-based input or limited walking space).

This advantage seems obvious for acquiring procedural knowledge (learning how) and has been proven in a couple of studies in the last years. For example, in a study from 2018, VR training was compared to non-VR training in the field of surgical education [Xi20]. 16 doctoral students of surgery were divided into two training groups. The task to learn was the insertion of pedicle screws into the spine, which was subsequently evaluated on human cadavers. The subjects of the VR group performed a simulation of this procedure during training, the non-VR group received a classical introduction by the chief surgeon. In terms of success rate and precision as well as speed, the subjects of the VR group were significantly better in the test.

Beside this it is also a question of whether the acquisition of declarative knowledge (factual knowledge; learning what) can also be improved by running a training simulation in VR. Declarative knowledge as such can be taught interactively in a screen-based simulation as well as in a VR simulation. Therefore, the question is: Does the fact of completing a training in VR lead to a greater (declarative) learning success than completing the same training on a screen? Since the interaction possibilities in VR are much closer to reality, this could have a positive influence on the learning success, which would not be present in screen-based training. Furthermore, it is important to measure the long-term success: How much have learners forgotten again after a certain time?

Thus a study was designed which disregards the advantages of a VR training simulation in knowledge retrieval but not in knowledge acquisition and makes a comparison to a training technique without these extended possibilities. More precisely, the study deliberately avoided measuring procedural knowledge, where various possible advantages are obvious in VR training, especially with regard to later execution in reality.

2 Materials and Methods

2.1 Participants

The experimental sample included 45 participants, 21 women and 24 men with an average age of 27 (min 19, max 41) recruited from Bielefeld University (Germany). Nearly all participants were students, most of whom came from computer science, psychology or philosophy. Attention was paid to an even distribution of participants in the test groups.

2.2 Experimental Design

A univariate design with two repeated measurements was applied. The considered factor was the level of knowledge about the necessary steps in the conducted training.

The overall sample was divided into 3 equally sized subgroups. One of them conducted the training in VR, a second one completed a screen-based application (SB group). Since the study was thematically concerned with the acquisition of declarative knowledge, a third comparison group was tested to compare the learning success to pure memorization.

In this group, the subjects were to learn the necessary steps by heart using simple text and picture instructions without any further training (TP group). Each group got training specific standardized instructions in terms of handling. The two simulation groups ran through the scenario exactly once. The average time required by the subjects was then set as learning time for the TP group.

Measurements were taken in a pre-test before completing the respective training, in a subsequent post-test and in a second post-test one week later. As described by Ebbinghaus' forgetting curve, the strongest loss of uniquely learned knowledge takes place in the first days after learning and then flattens out [Eb85]. Thus, it could be assumed that after seven days a sufficiently high discrepancy of knowledge levels emerges.

2.3 Educational Content

In the context of the study, participants had to deal with a healthcare scenario. The main focus lied on learning the procedure of a medical infusion preparation, a scenario nursing students are supposed to learn. This training was developed as a fully immersive VR application as part of the Virtual Skills Lab Project [Pf18] and used for the VR test group.

The training included 23 separate steps to perform. The screen-based group got the same application in classical first-person point-and-click adventure layout, a video game genre that has been outlined because of its suitability for educational applications [Am01].

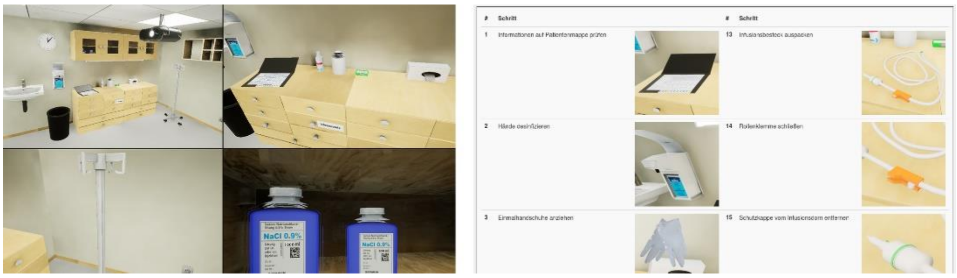

Fig. 1: Virtual lab in VR/screen-based app and instruction manual for third test group

Both virtual trainings were guided by the same instructions that were also used in the picture and text manual. The VR group received the instructions on a small widget on the left wrist, the SB group had the same widget in the upper left corner of the screen. Thus each group had the same training material at a different level of interaction. The TP group only had passive learning objects, the SB group had passive learning objects and an interactive environment and the VR group additionally had realistic behavior while interacting.

2.4 Knowledge Measurement

In general, a subjects knowledge level was defined as the set of known steps from the correct sequence both independently and dependent on the correct order. This way a participant who has forgotten the correct order but remembers the correct steps still knows more than someone who cannot remember anything.

To make the level of knowledge quantifiable, a digital flash card system had been developed and integrated in a web application. Objects used (bottle, disinfection cloth, ...) and activities performed (disinfecting, pricking, testing, taking, ...) were extracted from the individual steps of the infusion preparation and created as separate virtual "flash cards". The participants now had to build the tasks from the available objects and activities by combining them and putting them in an assumed correct order. In addition, distractor objects were added to the object pool, a rather obvious, one thematically roughly related but questionable and an absurd one: Bandages, scalpel and barbecue lighter. The latter served as a control item to eliminate unusable entities from the later data set. Since activities and objects were available in a data pool, they could be used as often as assumed for combining.

2.5 Calculating Distances Between Remembered and Original Procedures

If a lower case letter <x> is assigned to each object and an upper case letter <Y> to each activity, each created procedure can be described as a character string consisting of tuples of the type <xY> (e.g. *aDdEbCfN*; aD => "disinfect hands"). If the correct procedure is now defined as *iJaDdAcIbDdGaDfJeJeKjLkFlFgEfBmHjKhCnC*, a second procedure can be checked for similarity. The more similar the character string to be checked, the smaller the number of errors in the created sequence.

The Damerau-Levenshtein distance (DL distance) was used to calculate the minimum number of operations required to change one character string into the other (insertions, deletions or substitutions of a single character, or transposition of two adjacent characters)[MVM09]. By using the DL distance, however, a calculation problem can occur. A respondent who submits the character string *iJ* as a whole sequence has 36 errors according to the calculation, because 18 steps of the correct sequence are missing and must be inserted. In comparison, a second participant submitting a sequence that corresponds exactly to the correct sequence, but backwards, would result in the same distance. Both operations have various errors, but in relation to learning success the second subject knows considerably more, since at least all objects and activities are present and have been correctly combined. Omitting or forgetting correct objects/activities must therefore be given more weight in the calculation.

Two distinct sets can be derived from the correct procedure: Correct objects and correct activities. To positively weight the remembering of objects (and likewise activities) to be used, this set can be compared to the set of objects actually used in the submitted process. The similarity of the two sets 'CorrectObjects' and 'CorrectObjectsUsed' should influence the error rate. Therefore Jaccard-Similarity is used. Applied to activities and objects, two percentage values result, which correspond to the percentage similarity to the correct set of objects or activities.

In total, the following formula results:

$$Error = \frac{DLD}{0.5 * (object.similarity + activity.similarity)}$$

If all objects and activities are used, the denominator of the formula results in 1 and only the DL distance describes the error rate. If there are less objects/activities to be used, the denominator becomes smaller, which has a negative effect on the error rate.

The first subject from the previous example (submitted sequence = *iJ*) now has an error rate of 65.44, while the second subject still has 36 errors. An additional advantage of the separate consideration of correct objects/activities is to enable distinct analyses.

2.6 Sequence Blur

To calculate the learning success, a concrete solution for calculating deviations is needed. However, this is difficult to define. The free combination of objects and activities (albeit from a fixed pool) leaves room for interpretation.

In general, the verb "to take" is too versatile to be integrated into a concrete solution string. In order not to withhold existing steps, "take" was available as a verb for participants, but all created steps containing "take" were eliminated from the created character strings before the DL distance calculation.

Double interpretations were also handled. A step like "remove protection cap" has about the same meaning as "open protection cap".

2.7 User Experience Measurement

A difficult to use or distracting software can be a reason for a lower learning outcome. To exclude such cases, the User Experience Questionnaire (UEQ) was used to measure the usability of the training simulations [LHS08]. Participants had to fill out the questionnaire immediately after the first post-test (after the respective training).

2.8 Hardware

The VR group performed the training on a regular HTC Vive using the native Motion Controllers for interaction. The SB group used a Samsung Galaxy Tab S5e (10.5 inch). The TP group got the textual/picture instructions on a PC with a 24 inch monitor. All subjects performed the knowledge acquisition tests on a PC with a 24 inch screen.

3 Results

Informed written consent was obtained from each participant before the study. 5 out of 15 participants in the VR group stated to have already been using a VR application several times, 10 participants in the overall sample. The investigator observed that the handling of the interface for creating an infusion preparation sequence and the task definition was intuitively understandable for most of the subjects and for all after a brief explanation of the mechanics in the web application for the test runs. Training took place in a sound proofed room to ensure an undisturbed working atmosphere.

No participant was aware of the process of preparing an infusion before the start of the study and no one had integrated the control item "barbecue lighter" into the procedures created. Thus the entire data set was enabled for further analyses.

3.1 User Experience

Figure 2 highlights the mean values (columns) and variances (s^2; whisker) of the measurement results. The scales lie between values from -3 (extremely bad) to 3 (extremely good). Statistically, however, values above +/-2 are considered very unusual [LHS08].

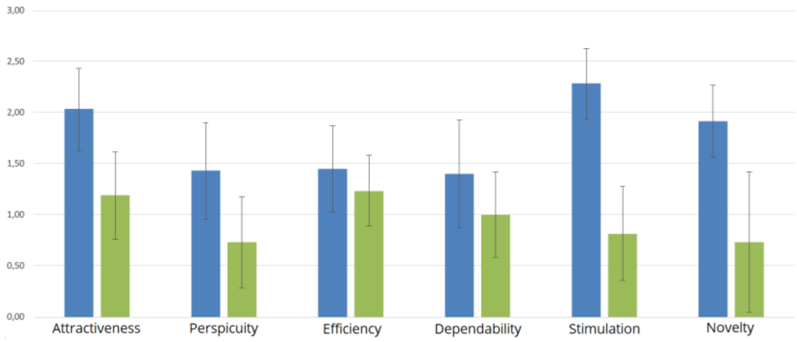

Fig. 2: UEQ Evaluation; blue = VR group, green = SB group

Overall the screen-based application received positive feedback. Attractiveness M = 1,189, s^2 = .71) and efficiency (M = 1,233, s^2 = .47) are the highest values. Dependability (M = 1.0, s^2 = .68) and stimulation (M = .817, s^2 = .83) are also still in the good range. The values for perspicuity (M = .733, s^2 = 0.79) and novelty (M = .733, s^2 = 1.68) are in the upper average range.

In the VR version transparency (M = 1.433, s^2 = .86), efficiency (M = 1.45, s^2 = .70) and controllability (M = 1.4, s^2 = 1.10) are in the good range with average values of 1.4. The originality follows higher (M = 1.917, s^2 = .49). Attractiveness (M = 2.033, s^2 = .62) and stimulation (M = 2.283, s^2 = .46) achieve very good values above 2.

A t-test for independent samples showed significant differences in attractiveness (p = .009), perspicuity (p = .044), stimulation ($p < .005$) and novelty (p = .007). In these 4 categories the VR app stands out from the screen-based app. There are no significant differences in efficiency (p = .444) and dependability (p = .256).

3.2 Knowledge Acquisition (analysis of error rates)

In all three groups an increase in knowledge (reduction of errors in the created sequence) was measured directly after the training and a decrease with the second measurement repetition one week later.

	pre	***difference***	**post**	***difference***	**scnd-post**
VR	32.518	57%	13.984	-26%	18.9
SB	29.663	55%	13.236	-30%	18.981
TP	30.984	88%	3.651	-75%	14.555

Tab.1: Mean values of error rates per sample, the lower the number, the better the learning

The strongest increase in knowledge with 88% after training (post) and the strongest decrease in knowledge (scnd post, one week later) with -75% was measured in the TP group.

The TP group has the lowest error rate in the last measurement with 14.6 errors, thus approximately 7-8 wrong steps.

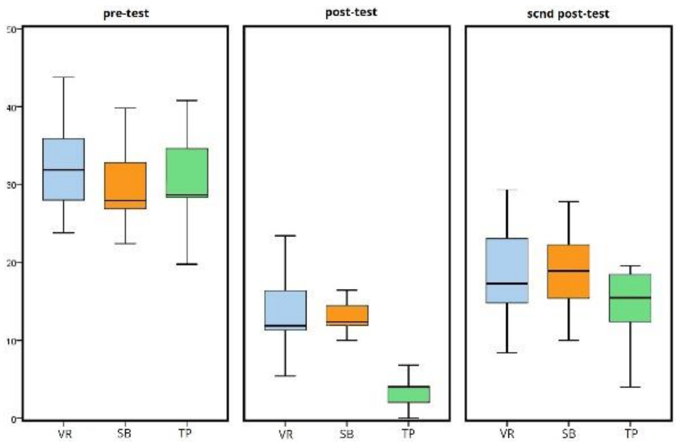

Fig. 3: Error rate, externally separated by measurement time, internally by condition

The error rates were analysed per measurement time using a univariate ANOVA with the betweensubject factor *condition* (VR, screen-based, text/picture). The ANOVA of the pre-test revealed no significant differences in the error rates of all samples ($F(2, 42)= 1,104$, $p = .341$). According to Welch-ANOVA of the post-test, there is a statistically significant difference between the groups (Welch test $F(2, 25,972)= 82,307$, $p < .001$).

Games-Howell post-hoc tests show in pair-wise comparison that the error rate of the TP group is significantly lower ($p < .001$) than in the VR and SB group (VR: -10.33, 95%-CI [-14.03, -6.64]; SB: -9.59, 95%-CI [-11.52, -7.65]). There is no significant difference between the two groups VR and SB with an average difference of 0.748 errors. The effect strength η (approximately: significance) according to Cohen [Co88] corresponds with $\eta = .66$ to a large effect (< =0.01 small effect, <=0.06 medium effect, <=0.14 large effect).

The ANOVA of the second post-test does not indicate a significant influence of the training condition ($F(2, 42)= 3.104$, $p = .318$). However, a slightly lower error rate (VR: $p = .081$; SB: $p = .074$) according to Bonferroni test is noticeable in the text group compared to VR/SB (VR: -4.36, 95%-CI [-9.08, 0.39]; SB: -4.43, 95%-CI [-9.16, 0.31]).

Another measured value was the percentage of correct objects and activities integrated into the built processes in the respective groups at the individual measurement points. Except for one extreme outlier, all participants of the TP group used 100% of the required objects in the post-test. With an average value of 87.18% (*SD* = .129) in VR group and 90.27% (*SD* = .114) in SB group, which corresponds approximately to one forgotten object, the results are just below the TP group.

The mean values in the scnd-post-test have a slightly decreasing tendency. The conditions VR and TP are also in the range of about one forgotten object and SB with one or two missing objects. (VR: M = 87.18%, SD = .122, n = 15; SB: M = 83.19%, SD = .13, n = 15; TP: M = 92.31%, SD = .122, n = 15).

Since no normal distribution can be assumed, Kruskal-Wallis tests are used instead of ANOVA to measure the central trends. The measurement of the post-test shows significant differences both in terms of forgotten objects ($\chi^2(2)$ = 8,445, p = .015) and activities ($\chi^2(2)$ = 10,682, p = .005). Paired comparisons with post-hoc tests (Dunn-Bonferroni) reveal that the amount of forgotten activities in both VR (z = -3,059, p = .007) and SB group (z = -2,526, p = .035) is significantly higher than in the TP group (effect strength according to Cohen [Co92]: medium effect (r = .46)). There is no significant difference between VR and SB (z = -.533, p = .594).

In addition, significantly more objects were forgotten in the VR group compared to the TP group when creating the process steps. The SB group has no significant differences to the other two groups.

In the scnd-post-test, the Kruskal-Wallis test shows no significant differences between forgotten objects ($\chi^2(2)$ = 3,987, p = .136) and activities ($\chi^2(2)$ = 1,793, p = .408).

3.3 Knowledge Acquisition (analysis of time)

The duration of an infusion preparation in the training simulations averaged 8.3 minutes. This value was rounded up to 9 minutes and set as learning time for the TP condition.

Table 2 shows the average test and training units duration in the respective groups. Although 9 minutes were allowed, the subjects in the TP group completed the learning phase after an average of 6.99 minutes (SD = 2.55).

	pre	***difference***	**post**	***difference***	**delay**	**scnd-post**	**training**
VR	7.24	11%	6.46	-27%	7.79	8.91	6.98
SB	7.35	13%	6.42	-7%	8.13	6.92	9.09
TP	8.16	7%	7.55	-8%	7.33	8.21	6.99

Tab. 2: Mean values of time measurements per sample (in minutes), duration post to scnd post test (delay) in days

The time participants required in both training simulations was analysed with a t-test. Accordingly, the infusion preparation was performed significantly faster in VR (M = 6.98, SD = 1.536, n = 15) than in the SB group (M = 9.09, SD = 1.935, n = 15) ($t(28)$ = -3.311, p = .003). The strength of the effect according to Cohen [Co92] is r = .53 and thus corresponds to a strong effect (r = .10 weak effect, r = .30 medium effect, r = .50 strong effect). As shown in column *delay* in table 2 the actual distance between the first and second post-test in each group exceeds 7 days on average. The greatest distance occurred in the SB group with a maximum of 10 and an average of 8.13 days (SD = .99).

The times of the test runs must be examined with regard to their influence on the error rate of the respective test. A Kruskal-Wallis test is used to identify significant differences within the conditions per knowledge test run. Due to the results there are no significant differences in the distribution of the consideration time across all three groups in pre-test ($\chi^2(1) = .269$, $p = .910$), post-test ($\chi^2(1) = .590$, $p = .622$) and scnd-post-test ($\chi^2(1) = 4.056$, $p = .186$). Spearman's rank correlation is used to test whether a low error rate correlates with a long consideration time. In pre- and scnd-post-test, there is a weak tendency for a slightly negative linear correlation according to the rank correlation coefficient (higher time for consideration leads to lower error rate). In pre-test ($r_s = -.230$, $p = .129$, n = 45) and scnd-post test ($r_s = -.212$, $p = .162$, n = 45) as well as in post-test ($r_s = -.109$, $p = .477$, n = 45) the values are not significant.

4 Discussion

4.1 Usability

The result of the UEQ evaluation corresponds to the expected result when comparing a well-known technology with an innovative and supposedly more exciting one. The point-and-click app has good results on its own, which highlights an adequate implementation of this genre. Nevertheless, it clearly loses in comparison to the VR app, which still has a kind of innovation bonus in terms of user satisfaction and offers a generally more exciting experience.

The dependability of both apps is positively assessed, which supports the results in the following learning effect evaluation. If greater problems had occurred here, higher cognitive load [SVMP98] could have had a significant negative impact on learning outcome. With a few weak exceptions, however, this was not evident in any condition neither through observation during the study nor through statistics.

4.2 Learning Effectiveness

The equality of the pre-tests in each group ensures a generally valid baseline in level of knowledge, therefore the measured improvement after the training is well comparable. Since no correlation between the duration of the knowledge test units and the error rate could be found, it can be ruled out that a different test duration leads to a biased result in favor of those who spent more time on the task. Unfortunately not all scnd post-tests were executed after exactly 7 days. Nonetheless an approximately equal distribution in the three groups relativizes this to a large extent. Just the TP group has a slightly earlier time of measurement repetition on average, which does not weaken the results. Assuming that most of the knowledge is forgotten in the first few days, as postulated by Ebbinghaus [Eb85], the difference after 7, 8 or 9 days is negligible.

Considered individually, each group has not only an average but also an absolute lower error rate after the training session. One week later, as expected, some content is forgotten. In comparison, the outstanding difference between the two virtual training simulations and the TP condition in the first post-test is striking. The increase of knowledge is clearly visible in each condition but is significantly lower in the simulation groups. Finally, it is not surprising that there is a clear difference, since the participants of the TP group had 9 minutes to memorize the 23 steps, whereas the participants of the VR and SB conditions could only see each step once. In addition, the TP group had the opportunity to learn all contents until shortly before the time expired, while in the simulation conditions the first steps (due to iterative performance) were already a few minutes ago. In the minimum time span between the end of the training and the post-test, learned contents are hardly forgotten, but time spans around 5-10 minutes can play a decisive role, especially if each step is performed only once [Eb85].

On the other hand, the participants of the simulation groups were able to perform the steps themselves within the scope of the interaction possibilities, thus using a further learning channel and dealing with the respective contents more detailed. Nevertheless, pure memorization is more advantageous in this study, as reflected in the statistics.

A comparison of the VR and the SB group shows no statistically significant difference in general. The small difference in slope (57% / 55%) is far from sufficient to certify a group as "more effective". Similarly, a look at the scnd-post-test shows no significant difference (drop 26% / 30%) between the two simulation groups. The data collected and experience gained during the study hardly allow any room for argumentation contrary to the statistics. Thus, it could not be shown that training in VR facilitates the acquisition of declarative knowledge compared to a related screen-based application. However, the sample size must be taken into account, which was rather small with 15 subjects per group.

Nevertheless, the data of the scnd-post-test shows another anomaly. After one week, the determined knowledge level of the TP group has clearly reduced. Compared to the VR and SB group, TP still has better values, but statistically significant differences are no longer present. This suggests that learning content in training simulations is more efficient in the long run. The comparatively strong knowledge loss of the TP Condition supports this thesis, since the VR and SB groups show just a small loss. Contrary to this assumption, the simulation conditions had less knowledge to lose from the beginning. The TP group had learned a larger number of steps in the training, which is more difficult to remember than the already faulty results of the other groups. In addition, the TP group with evaluation of the scnd-post-test is in the same range of values as VR and SB but has not turned out to be worse. This also allows the thesis that this level of knowledge represents a kind of limit to which test persons fall back after a few days, no matter what kind of training they have completed and how much knowledge they have acquired. The question would thus be whether the users of the training simulations, if they had learned more (e.g. by doing the training twice or three times), would also have forgotten more after one week.

4.3 Conclusion

The results leave room for further research, in which the groups "training simulation" and "text/picture" should have the same level of knowledge not only before but also directly after the training session. On the basis of the conducted study, it can be hypothesized that students who have learned all steps of a certain, previously unknown sequence in a training simulation retain the factual steps significantly longer in their memory than persons who have learned it by heart via textual instructions. Another approach could also consider a hybrid solution that takes advantage of the rapid acquisition of knowledge through text/picture instructions and then consolidates what has been learned through a training simulation. According to the results of this study, a differentiation of the simulation types between VR and Screen-Based (related to the acquisition of declarative knowledge) does not seem to be necessary anymore, but a repeated differentiation might, although not expected, provide a different result.

Bibliography

[Am01] Amory, Alan: Building an educational adventure game: Theory, design, and lessons.Journal of Interactive Learning Research, 12(2):249–263, 2001.

[BMS88] Barbe, Walter Burke; Milone, Michael N; Swassing, Raymond H: Teaching throughmodality strengths: Concepts and practices. Zaner-Bloser, 1988.

[Co88] Cohen, Jacob: Statistical power analysis for the behavioral sciences. Technical report,1988.

[Co92] Cohen, Jacob: A power primer. Psychological bulletin, 112(1):155, 1992.

[Eb85] Ebbinghaus, Hermann: Über das gedächtnis: untersuchungen zur experimentellen psychologie. Duncker & Humblot, 1885.

[LHS08] Laugwitz, Bettina; Held, Theo; Schrepp, Martin: Construction and evaluation of a userexperience questionnaire. In: Symposium of the Austrian HCI and Usability EngineeringGroup. Springer, pp. 63–76, 2008.

[MVM09] Miller, Frederic P.; Vandome, Agnes F.; McBrewster, John: Levenshtein Distance:Information Theory, Computer Science, String (Computer Science), String Metric,Damerau?Levenshtein Distance, Spell Checker, Hamming Distance. Alpha Press, 2009.

[Pf18] Pfeiffer, T. et al.: Virtual SkillsLab-Trainingsanwendung zur Infusionsvorbereitung (Wettbewerbssieger). In: DeLFI Workshops 2018. Proceedings der Pre-Conference-Workshops der 16. E-Learning Fachtagung Informatik (DeLFI 2018), Frankfurt, Germany, September 10, 2018.volume 2250, 2018.

[SVMP98] Sweller, John; Van Merrienboer, Jeroen JG; Paas, Fred GWC: Cognitive architecture andinstructional design. Educational psychology review, 10(3):251–296, 1998.

[Xi20] Xin, B. et al.: The efficacy of immersive virtual reality surgical simulator training for pedicle screw placement: a randomized double-blind controlled trial. International Orthopaedics, pp. 1–8, 2020.