Gamification intensity in web-based virtual training environments and its effect on learning

Thomas Bohné, Ina Heine, Felix Mueller, Paul-David Joshua Zuercher, Vera Maria Eger

Abstract—

Gamification approaches to learning use game-inspired design elements to improve learning. Given manifold design options to implement gamification in virtual environments, an important but underexplored research area is how the composition of gamification elements affects learning. To advance research in this area, we systematically identified key design elements that have shown promise in leading to positive learning results. We then conducted an experiment in which we varied gamification intensity in web-based virtual training environments for a procedural industrial task. 355 participants were divided into a baseline group without gamification, a basic, and an advanced gamification group. Analysis of participants' learning included learning outcomes (time-to-completion and number of mistakes), affective learning factors (motivation, self-efficacy, satisfaction), learning system usability and perceived cognitive load throughout the learning process. The results did not show any statistically significant difference favoring the higher levels of gamification intensity with respect to lower ones. Conversely, we found that participants' computer gaming habits and technical equipment (display size and computer pointing device) significantly influenced learning.

Index Terms— **Gamification, virtual reality, industrial training, virtual training, virtual assembly, game-based learning, non-immersive virtual reality**

I. Introduction

One of the advantages of virtual 2D or 3D environments is that they enable novel opportunities to implement gamification to increase the training motivation of learners and enhance learning outcomes [1], [2]. Gamification is defined as the use of game design elements in non-game contexts [3]. However, empirical understanding of gamification's potential to enhance virtual training remains limited, particularly in industrial learning contexts [4], [5]. Given manifold design options to implement gamification in virtual training environments, an important but underexplored research question is how different gamification elements affect learning. We focus on this research gap with an experiment comparing three different gamification intensities (defined by the number of game elements used) and their effect on the trainee's learning. Therefore, we investigated objectively measured learning outcomes as well as subjectively measured affective learning factors, learning system usability and perceived cognitive load throughout the learning process. To further investigate the learnings' sustainability, our research additionally measured knowledge retention over the time of 3-4 weeks after the initial training. Our research advances learning technologies research by contributing new empirical evidence on gamification intensity and how choices of game element configuration and learners' technological equipment influence learning. The insights of our study have important implications for educators and designers of virtual trainings contemplating gamification and aiming to advance future research on the optimal application of gamification in virtual learning contexts. In what follows below, we outline the theoretical foundation of our research, explain our research design, present results, and then discuss our findings and their limitations and opportunities for future research.

II. Theoretical Background

A. Virtual training and its evaluation

The learning context of our study is industrial training which can be defined as a systematic organizational process that equips employees with the required knowledge, attitudes, and skills to achieve an organization's mission and goals [6]. To evaluate the effectiveness of these training processes and frameworks several qualitative and quantitative schemes have been proposed [7]. Drawing on previous studies, we particularly considered a multilevel training evaluation framework because it includes subjective affective indicators and objective performance indicators [8]. Affective indicators typically include motivation, satisfaction, and self-efficacy. Performance indicators tend to vary depending on specific learning objectives. For learning industrial tasks, for example, common key performance indicators (KPI) are time-to-completion (TTC) and number of mistakes.

When investigating performance indicators, it is also important to assess their development over time because research has shown knowledge and skills decay over time [9], [10]. Scientific attempts to predict knowledge decay range from mathematical models to non-gradual knowledge collapse theories [11], [12]. Empirical studies indicate that retention intervals affect sustained knowledge [10], [13], [14]. With each additional training session taking up personnel and costs, organizations are generally interested in designing training to minimize the number of knowledge-refreshment needed.

Recent studies show that other metrics, such as perceived usability of the training platform are also important, especially for virtual training [15]–[19]. In general, high usability induces improved learning effects [17], [18]. Common attempts to quantify a system's usability include the System Usability Scale (SUS) which consists of 10 scale-based questions [20].

Cognitive load (CL) is induced throughout a training process. Cognitive load theory (CLT) proposes that human memory is composed of three memory systems: long-term memory

(LTM), working memory (WM) and sensory memory (SM) [21]. LTM is where the permanent or semi-permanent information is stored, WM is the structure devoted to the maintenance of information in short term during concurrent processing activities while SM is where sensory information is stored before being transferred to the WM. Thus, new information needs to pass by the WM to be stored permanently. CLT further claims that humans' WM storage and processing capabilities are limited, whereas the LTM theoretically does not face any constraint of that type [21]. Cognitive load theory is backed by recent studies revealing that without rehearsal, information in the WM is lost within 20 seconds [22]. Since comprehension and learning are determined by the capability of the WM, learning researchers conclude that the WM should solely be occupied by task-relevant information [23]. This results in the training objective to minimize the cognitive load by, for example, removing external or redundant information [24]. Based on the human brain's complexity, measuring the multidimensional concept of cognitive load proves to be a challenge, yet three measurement categories have emerged: (subjective) rating scales, psychophysiological methods, and secondary task techniques. Although the subjectivity of self-ratings may appear questionable, it has been demonstrated that people are quite capable of giving an accurate numerical indication of their perceived CL in past studies [25]. The most common NASA-TLX scale considers six different CL-dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration [26], [27].

B. Gamification

While cognitive load theory recommends removing all information not directly task-related to enhance learning outcomes, it neglects any positive effects such information or objects might have that outweigh their additional cognitive load [28]. These objects could, for example, increase other training evaluation metrics (e.g., motivation) while inducing only slight additional CL and resulting in an overall benefit [29], [30]. The concept of gamification is based on this strategy and can be defined and distinguished from other concepts as set out by [3]:

- *"the use* (rather than the extension) of
- *design* (rather than game-based technology or other game-related practices)
- *elements* (rather than full-fledged games)
- *characteristic for games* (rather than play or playfulness)
- *in non-game contexts* (regardless of specific usage intentions, contexts or media)".

Based on this definition, gamification is characterized by the presence of several game-like design elements (game elements). This results in two difficulties: firstly, gamification is manifold due to the potential composition of different game elements. Secondly, the lines between conventional design elements (e.g., instructional design elements) and game elements are sometimes blurry. We tackled the latter by grouping the game elements from the literature into basic game elements and advanced game elements. Basic game elements are design elements which some studies classify as instructional elements while others classify them as game elements. The advanced game elements on the other hand were solely classified as game elements across all previous studies.

Our literature review revealed 35 different game elements which are listed in the Appendix [31]–[35]. Mathematically and assuming – for the sake of simplicity – that each element can only be used in one way, this could possibly result in more than 30 billion different game element combinations, which makes research reproducibility all but impossible. Consequently, conducting empirical studies and selecting a set of game elements are important to progress knowledge [2]. A promising attempt to identify a suitable and consistent set of game elements is by considering their compatibility with underpinning gamification theories. The most frequently mentioned theories in this context are Flow Theory [36], Self-Efficacy Theory [37], Social Comparison Theory [38], Goalsetting Theory [39], Operant Conditioning Theory [40], and Self-Determination Theory (SDT) with its micro-theory Basic Needs Theory (BNT) [41]. Based on its wide range of applicability and acceptance [32], our research draws on SDT and BNT.

C. Self-Determination Theory (SDT) and Basic Needs Theory (BNT)

SDT is a macro theory comprised of several micro-theories that inform predictions made in self-controlled motor learning studies. The most relevant micro-theory is BNT which assumes that humans have three basic psychological needs that contribute additively to human thriving and work motivation [41], [42]. These three needs are autonomy, competence, and relatedness [41]. Autonomy involves feeling internal assent regarding one's behavior, instead of feeling controlled or pressured by outside forces. Competence involves feeling efficient, effective, and even masterful in one's behavior. Lastly, relatedness includes feeling meaningfully connected to others instead of feeling alienated or ostracized. Furthermore, research attributes universal relevance to these psychological needs independent of cultural background [41]. Due to its proven relevance for learning and working outcomes, we used BNT as a criterion to select suitable game elements for our gamification setting in the virtual learning environment [43].

D. Selection of Game Elements

Previous studies have reported mixed findings regarding gamification's influence on learning in virtual training environments, with some reporting positive effects with varying effect sizes and others reporting no effects [44]–[46]. One of the weaknesses of many previous studies is that the selection of game elements was essentially unsystematic and rarely justified on theoretical grounds. To advance research in this domain, we chose a structured approach considering 35 game elements and selected 14 elements based on three individual factors and two compositional factors.

The three individual factors of the five-factor methodology are:

- Sufficient effect size
- Compatibility with BNT
- Compatibility with Learning Objective and Learning Context

The first factor is the sufficient effect size of the game element on learning. Since most empirical gamification studies have investigated the impact of multiple game elements in

composition it is difficult to pinpoint how and to what extent each individual game element contributed to user motivation and behavior [47]–[49]. To tackle this problem, our research considers several studies per game element to identify individual impact. Sufficient effect size means that most of the previous studies in which the respective game element was investigated yielded positive learning results. The second factor considered if the game element's mechanisms align with BNT. Lastly, the game element must also suit the learning objective of procedural knowledge acquisition in an industrial context, which was the learning context of our study.

In addition to individual factors, we considered two compositional factors during the game element selection process: redundancy and counterworking. Both focus on game element interaction and whether game elements are redundant or counterworking. The category-based assessment of all game elements can be found in the Appendix.

Table II.1 summarizes a list of the 14 most promising game elements for our learning environment.

Table II.1: List of game elements applied in this study

Game Element	**Description**
Exploration	Gives the user the possibility to investigate and discover features of the system, through exploratory tasks [50]
Feedback	Returns relevant information about the (game) state to the user [51]
Virtual Instructor	Virtual (often human-like) persona accompanying the user [52]
Goalsetting	Clear goals are presented to the user [33]
Learning Examples	Exemplary demonstration of another user or the virtual system [53]
Signposting	Guiding signs by the system to support the user [54]
Prize	A reward that a user wins for his/her actions [38]
Badge	Visual representation of the user's achievements [55]
Choice	Enables the users to have the autonomy to determine their verdict among many possibilities [56]
Meaning	Allows the user to auto-identify with the virtual system via common purpose [57]
Narratives	Giving context to the user's tasks by implementing plots and stories [58]
Progress	Milestones and objects indicating the user's progress [34]
Reward Schedule	Schedule item that strengthens the user's behavior in anticipation of new rewards [58]
Roles	Role-playing elements of characters [35]

III. Research Design

To investigate gamification's potential to enhance the effect of web-based virtual training, we focused on five main hypotheses. Each hypothesis consists of the same independent variable (gamification intensity) and varying dependent variables covering the training evaluation metrics.

A. Hypotheses

The first set of hypotheses investigates the positive effect of gamification during the training process on the objective learning outcome of increased labor productivity after the training. Labor productivity in industry can be quantified by the number of mistakes (H1a) and the TTC (H1b) throughout a certain task. Based on previous findings, we hypothesize that higher levels of gamification intensity reduce both number of mistakes and TTC [59], [60].

The second group of hypotheses investigates the effect of implementing gamification during the training process on affective learning factors such as motivation (H2a), satisfaction (H2b) and self-efficacy (H2c). While previous studies revealed mixed results due to different game elements applied, we hypothesize that higher levels of gamification intensity significantly increase all three factors [2], [35], [58].

Furthermore, hypothesis H3 proposes a positive effect of gamification intensity on the system's usability perceived by the learner. Even though there are limited insights on whether gamification itself increases usability, the opposite effect of usability enhancing gamification was already identified [61]. Moreover, research identified usability to have a moderating effect on gamification's potential results [62]. Building on these findings, we hypothesize similar tendencies for the opposite effect. Since any additional visual element requires information processing by the sensory memory and the working memory, gamification intensity and its visual game elements are expected to induce an increased perceived cognitive load, which is expressed through hypothesis H4 [21], [24], [63].

Lastly, knowledge retention is considered to evaluate gamifications potential to enhance web-based virtual trainings. Based on previous studies investigating similar objectives, we hypothesize that higher levels of gamification intensity lead to increased knowledge retention with respect to lower levels (H5) [60], [64], [65]. All hypotheses are summarized in Table III.1.

Table III.1: Research hypotheses

	Hypotheses
H1	A higher level of gamification intensity during the training process increases labor productivity of assembly tasks as measured by: H1a: number of mistakes H1b: time-to-completion
H2	A higher level of gamification intensity during the training process improves affective learning factors as measured by: H2a: motivation H2b: satisfaction H2c: self-efficacy
H3	A higher level of gamification intensity during the training process increases the perceived system usability
H4	A higher level of gamification intensity during the training process induces higher perceived cognitive load
H5	A higher level of gamification intensity during the training process enhances knowledge retention

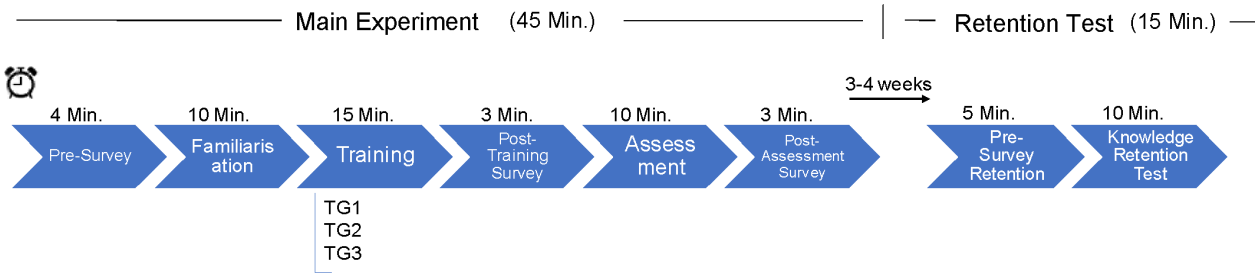

Figure III.2: Experimental study overview

B. *Experimental Study*

As all hypotheses posit a positive effect of an independent variable (gamification intensity) on several dependent variables (e.g., number of mistakes), we designed an experimental study to investigate the hypotheses.

1) Learning objective: manual assembly task

To investigate the hypotheses in a way that reflects real tasks, ideally a task was chosen that can be found across different applications. An adequate level of task complexity was needed to prevent ceiling effects. Based on these prerequisites, we chose the partial assembly of the low-voltage frequency converter. The product consists of 10 individual assembly steps, which are listed in Table III.2. Each step requires a sequence of pick and place tasks and the use of common tools (e.g., screwdriver), which mainly reflects procedural knowledge acquisition throughout the training as can be found in many industrial settings.

Table III.2: Assembly steps

Step	Sub-step	Description
1	1a	Place screening shield on the main part
	1b	Place two M4x12 screws
	1c	Fasten the two screws using the electric screwdriver
2	2a	Fit the main housing on the main part
3	3a	Place four M4x20 screws
	3b	Fasten the four screws (cross-wise) using the electric screwdriver
4	4a	Place the earth strap on the main part
	4b	Place two M4x9.5 screws
	4b	Fasten the two screws using the manual screwdriver
5	5a	Place the internal stirring fan on the main part

The assembly task has a complexity value of $C_{Product} = 6.34$, which ranks the task in the upper range of the medium complexity category, according to formula (2) [66].

$$C_{Product} = \left(\frac{n_P}{N_P} + CI_{Product}\right) * (N_P + 1) + \frac{n_S}{N_S} * (N_S + 1) \quad (2)$$

$$C_{Product} = product\ or\ task\ complexity$$
$$CI_{Product} = complexity\ index\ (parameter)$$
$$n_P = number\ of\ unique\ parts$$
$$N_P = total\ number\ of\ parts$$
$$n_S = number\ of\ unique\ fasteners$$
$$N_S = number\ of\ total\ fasteners$$

2) Experimental procedure

As can be seen in Figure III.2, our experimental study consists of two separate remote online experiments: the main experiment and the retention test. The retention test was conducted in the period 3-4 weeks after the main experiment. Apart from the training phase in the main experiment, each participant went through the exact same process.

The main experiment included six steps which take approximately 45 minutes to complete. After obtaining participants' consent, each participant completed a pre-survey covering items such as demographics, computer gaming habits, pre-motivation, and previous assembly and VR experience. As any kind of pre-experience or initial motivation might influence the assembly task performance, we checked them a-priori. Moreover, in the unlikely case participants had already assembled the converter they were excluded from proceeding to the next phase.

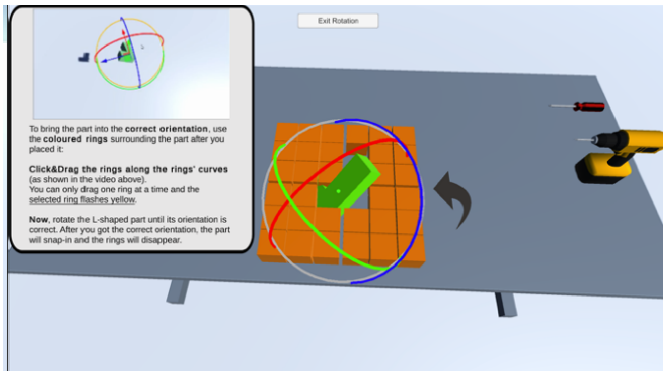

Figure III.1: Interface during the familiarization phase

During a familiarization phase, participants were shown the virtual environment and interaction controls to ensure their performance was not biased by insufficient interaction skills in the virtual environment. Implementing a familiarization phase is a common practice in virtual reality experiments [67], [68]. As can be seen in Figure III.1, the familiarization phase's design, including workstation and tools, was similar to the real training and assessment phases to simplify orientation for the participants. A virtual part rotation tool (gizmo) was introduced to each learner and an overview of all controls was presented

throughout the experiment if participants pressed the key "C" on their keyboard. The familiarization phase was completed once all essential controls were explained to and performed by the participant.

Upon successful completion of the familiarization phase, participants progressed to the training phase during which they were trained on the assembly of the converter. Each of the 10 assembly steps (see Table III.2) was learned by reading the instructions and virtually performing them. We designed three different virtual training environments (Treatment Groups, TG) which differed in their gamification intensity (i.e., number and type of game elements). All other factors were kept constant. Each participant was randomly allocated to one of the TGs with the help of a distributed algorithm and the participants did not know about the other TGs.

Figure III.3: Interface during the training phase

After completing the training phase by performing all assembly steps, participants were forwarded to the post-training survey which assessed the participants' post-training state in terms of affective learning factors and perceived system usability. Besides gathering data about the participants' post-training state, our survey served as a distraction break between learning (training phase) and being tested (assessment phase).

During the assessment phase, participants were asked to virtually assemble the converter, which was identical to the training phase. As can be seen in Figure III.4, the same objects are presented but no assembly instructions were shown nor given via audio. Participants were requested to perform the assembly to the best of their knowledge. Participants received feedback for incorrect placement of objects, as illustrated in the upper left corner of Figure III.3.

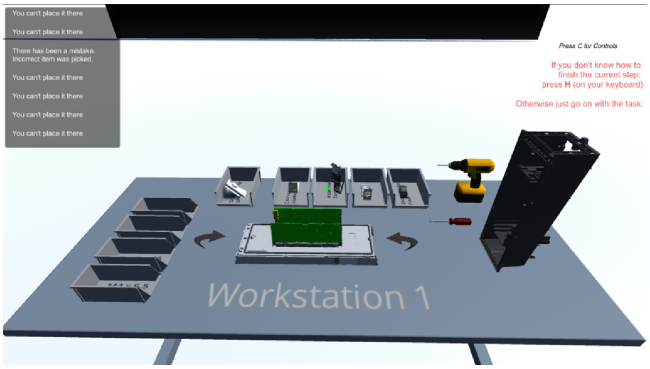

Figure III.4: Interface during the assessment phase

Since the assembly was procedural, participants were not able to proceed to the next step before completing the previous one. This could have resulted in some participants dropping out at an early step even though they might perform very well at the consecutive steps. To minimize this risk and to gather data based on the participant's performance throughout the entire assembly process a help function was implemented. After a pre-defined interval (60s) passed without progress, the participant received the option to ask help and the respective assembly step's information was shown on the information board, similarly to the training. This process also simulated real industrial procedures of checking, for example, the assembly manual again. In analogy to industrial assembly performance indicators, we measured the participant's number of mistakes and TTC of the assembly process during the assessment phase.

After completing all assembly steps, the participants were forwarded to the final step of the main experiment, the post-assessment survey. The post-assessment survey investigated the participants' setting and environment while conducting the experiment and measured perceived cognitive load. Since remote experiments may be prone to disruptions, it was essential to identify any disruptions causing data distortions.

After successfully finishing the main experiment, the final step of the experimental study was a retention test. Three weeks after the participant had finished the main experiment, they received an access link to the retention test. Retention tests are promising tools to assess long-term learning outcomes and several studies have already implemented them [16], [69], [70].

The retention web link sent to the participants was valid for one week. The precise time slots are important to allow comparison among the participants because knowledge deterioration increases over time. Moreover, previous studies indicate significant knowledge decay for retention intervals (RIs) of more than one week and provide promising results for retention intervals of 18 days [13], [14]. Thus, we chose 3-4 weeks as a suitable retention interval.

After the participant consent and information sheets were approved by the participant, the pre-motivation was assessed and general questions about the assembly process asked. In the next step, the assembly knowledge assessment took place. As can be seen in Figure III.4, the participant's knowledge about the assembly sub-steps was assessed by multiple choice questions, all following the same pattern. The participant was requested to define any assembly step by defining the part to be placed and select the target location. Moreover, screwdriver selection and crosswise screwing strategy were assessed using multiple choice methods. Thus, the retention test investigated the participant's retained knowledge about the process.

Figure III.5: Interface during the retention test

3) Experimental measurements

This section covers the variables measured and the respective measurement scales and methodologies applied during the experiment. All relevant scales are also listed in the Appendix.

During the pre-survey demographics, computer gaming habits, pre-motivation, and previous assembly and VR experience was recorded. Pre-motivation was investigated using the motivational scale developed by Noe and Wilk in 1993 and was compared with the level of motivation after the training to detect motivational changes [71]. The scales are listed in the Appendix section.

The post-training survey consists of 29 questions. The first block of questions investigated the participant's perceived usability of the training environment using the System Usability Scale which consists of 10 questions based on a 7-point Likert-scale (see Appendix) [20]. The next semantic block investigated the participants' perceived self-efficacy using one 7-point Likert scale question (see Appendix). To quantify the level of self-efficacy induced by the training session, the participants were asked to what extent they felt able to accomplish the task effectively [17]. After assessing self-efficacy, participants' motivation was examined. This consisted of four 7-point Likert-scale questions examining the intrinsic motivation, similar to previous virtual training motivation studies [72]. Lastly, post-training satisfaction was investigated with a 7-point Likert-scale question asking to what extent they felt happy about the training method, similarly to previous studies on virtual learning environments [73].

Since learners' motivation is quantified using subjective scales and is further highly dependent on their pre-motivation [74], it is essential to investigate the change in motivation induced by the training session. Thus, the variable *training motivation* is defined as the ratio of *post-training motivation* and *pre-training motivation*, as summarized by formula (4).

$$Training\ Motivation = \frac{Post - Training\ Motivation}{Pre - Training\ Motivation} \quad (4)$$

During the assessment phase of the main experiment, two absolute variables are measured: TTC and number of mistakes. Measuring TTC is self-explanatory as the time the participant needs to complete all assembly steps. A mistake was counted each time the participant did not complete a sub-step at the first try. Furthermore, the participants could not make more than one mistake during a single sub-step. Thus, the maximum number of mistakes is equal to the number of sub-steps listed in Table III.2. These mistakes could be either grabbing the wrong part or trying to place the correct part in an incorrect position.

The post-assessment survey consisted of 14 questions and investigated the participants' setting and environment while conducting the experiment and measured perceived cognitive load. Also, we investigated the participant's post-assessment satisfaction using a 7-point Likert-Scale to compute the overall variable *satisfaction* by taking the average value of two data points: *post-training satisfaction* and *post-assessment satisfaction*. We further used five 7-point Likert-scale questions to investigate the participants' settings and one question to investigate participants' computer system performance. Due to the remote nature of this experiment, the NASA Raw Task Load Index (NASA-RTLX) rating scale was used (see Appendix). The scale consists of six components, which are individually assessed: mental demand, physical demand, performance, effort and frustration. Each component is assessed on a scale from 0 - 100 and the average of all six components reflects the final NASA-RTLX value [26], [27], [75].

The level of knowledge retention was measured during the retention test by using a visual multiple-choice survey and tracking the participant's number of incorrect answers. In total, 14 knowledge retention questions were asked during this period which yielded the maximum number of possible incorrect answers (14). The decision to not conduct the exact same assessment as in the main experiment had several reasons. First, answering a short survey (15 min.) induces lower participation thresholds for participants and supports lower drop-out rates and thus higher data quality for a retention survey. Secondly, even though the two assessment scores cannot be compared due to the different assessment types, the knowledge retention of the different TGs can be compared.

4) Treatment groups (TG)

As explained in the previous sections, this study applied three different virtual training environments on the participant sample to compare their impact on learning.

The three different environments represent experimental treatment groups (TG) which differ in the level of gamification intensity, ranging from "No Gamification" (TG1), "Basic Gamification" (TG2), to "Advanced Gamification" (TG3). As can be seen from Table III.3, TG1 contains two essential elements and served as the study's control group. Despite the presence of *Exploration* and *Feedback*, TG1 is labelled as non-gamified because these elements are prerequisites for non-immersive virtual learning environments. TG2 includes four game elements. TG3 consists of TG2's elements plus 8 additional game elements.

Table III.3: Game element composition for the three treatment groups

Game Element	TG1 – No Gamification	TG2 – Basic Gamification	TG3 – Advanced Gamification
Exploration	X	X	X
Feedback	X	X	X
Virtual Instructor		X	X
Goalsetting		X	X
Learning Examples		X	X
Signposting		X	X
Prize			X
Badge			X
Choice			X
Meaning			X
Narratives			X
Progress			X
Reward Schedule			X
Roles			X

While the theoretical background of each of the game elements can be obtained from section II.D, the following paragraphs elaborate how we designed and implemented the respective game element in our web-based virtual environment.

Exploration was implemented by allowing the learner to navigate inside the virtual environment and to interact freely with objects. Feedback was implemented by giving the user visual and auditive feedback on each action, including task related and non-related interactions. As can be seen in Figure III.3, the game element Virtual Instructor (VI) was implemented with a human-like virtual instructor. Since facial expressions do not seem to influence learning outcomes the instructor was designed to embody a neutral-looking industrial worker [76]. While participants could not communicate with the VI, it accompanied and supported learners during the assembly training by directing them through the assembly steps. Goalsetting was implemented as a clear goal definition at the beginning of the training. Participants were told to be as fast as possible and conduct as few mistakes as possible. Learning examples were shown at the final part of the respective step. Signposting included visual attention cues marking the objects needed for the next assembly step. The required objects started color-flashing if they were required to be used in the current step. The game element Prize was implemented by golden coins emerging once an object was placed correctly. This aimed to trigger participants' extrinsic drive to collect monetary.

Badges involved three different levels of achievement. Three badges were used and named Beginner, Professional and Expert. For each correct interaction during the training process, the badge filled a little more. If one badge was filled it was rewarded to the participant and the progression to the next badge began.

The game element Choice was implemented by allowing the user to choose the level of difficulty for the training. However, the level of difficulty remained the same and the choice was just a perceptional one. Despite no non-linear gameplay being implemented, the mere perception of having a choice induces similar gamification effects on learners and can motivate the learner to assess their choice properly [77]–[79]. The game elements Meaning, Narrative and Roles were implemented in the same scene by informing the learner via text and audio about the context of the training and assigning them a role as an industrial worker assembling an important part for car tunnel ventilation. Prior to the actual training, participants received a motivational story about the importance of the converter for human safety in car tunnels and the participant's own special role in the assembly team. Moreover, the virtual environment surrounding the workstation was altered to an industrial setting, as can be seen in Figure III.3. While the game element Progress was implemented using a permanently visible, orange-colored progress bar, the game element Reward Schedule refers to the visibility of all three badges at any time. Thus, participants got a feeling of scheduled rewards in terms of badges. Both elements' designs can be seen in Figure III.3. The progress bar also included percentage information supporting the user's capability to track the individual progress during the training.

5) Recruiting strategy

Participants were able to take part at any time and no specific software was needed, apart from a computer. Access to the virtual training was granted via a web link to the application and the link was distributed across private networks and the crowdsourcing platform Amazon Mechanical Turk (MTurk) as the platform's experimental potential has been used in previous research and more than 15,000 papers have been published using MTurk as a source for data collection [80], [81]. Since the task required no previous knowledge, no required background knowledge was needed and any participant older than 18 years could participate. We tracked technological pre-experience prior to the experiment to identify potential moderator effects and variables. On MTurk, we only allowed people with an approval rating of more than 90% to take part in the experiment, which is a common threshold to ensure data quality [82]. Seven attention check questions (ACQ) were implemented in the experiment to filter potentially fraudulent or automated activity. To improve participation and data quality, all participants completing the experiment were given a £7 Amazon voucher for the main experiment and another £3 if they completed the retention test.

6) Data Analysis

Prior to the actual data analysis, incomplete data were excluded from the dataset. This included unfinished participation or participants who failed at least one of the ACQs, had technical difficulties, or faced distractions during the experiment, all of which are important for ensuring high data quality in remote experiments [83]. The latter was measured by metadata and participant self-assessment questionnaires. The remaining sample was then tested in terms of scale consistency by computing Cronbach's alpha (α_C) which must be larger than 0.7 [84]. Prior to the hypotheses testing, all control variables (e.g. demographics and pre-experience) were analyzed on their statistical correlation with the TGs to ensure an equal distribution across all three TGs. Both, the control variables, and the hypotheses were investigated by testing the null hypothesis stating that there are no significant differences between the groups. Depending on whether normal distribution

or homoscedasticity were given for the respective variable, either the Kruskal-Wallis Test (KW-Test) or an Analysis of Variance (ANOVA) were conducted. Statistically significant differences were then identified if the resulting error probability (p-value) is below 0.05. Non-parametric effect size *r* was computed using formula (3). While medium effect sizes range between 0.3 and 0.5, small effect sizes are below 0.3 and large effect sizes are above 0.5.

$$r = \left|\frac{Z}{\sqrt{n}}\right| \quad (3)$$

n = number of data points
Z = Z-value computed with Mann-Whitney U-Test

IV. RESULTS

A. Participant Sample

In total, 423 participants finished the experiment. 31 were filtered out because they failed at least one of the ACQs. 36 participants were excluded from the remaining 392 because they reported technical issues which impacted their performance. Lastly, 1 participant was excluded based on the environment check questions, resulting in a final sample size of 355 and an exclusion rate of 16.1%.

The final main experiment sample (n = 355) is distributed across the three treatment groups TG1 (n = 127), TG2 (n = 126) and TG3 (n = 102). The average completion time of the main experiment was 55 minutes. The mean age is 31.23 (SD = 9.32) and 27.3% of the participants are female, while 0.8% stated non-binary and 71.9% are male. Regarding the country of residence, the largest group was from Germany (43.1%) followed by the United States of America (36.6%), the United Kingdom (8.7%) and India (6.2%). 13 other countries were represented in the sample but with low ratios.

While 73.2% of the participants reported spending more than six hours per day using a computer, 80.3% reported playing computer games less than 3 hours per day. Most of the sample stated prior experience with VR (72.7%) and assembly tasks (74.6%). Only a small ratio (13.5%) had already faced other virtual assemblies. Office workers represented the largest ratio of occupations (36.6%).

The Cronbach's alpha value was larger than the 0.7 threshold for all scales, including pre-motivation (alpha=.924), post-motivation (alpha=.947), satisfaction (alpha=.778), usability (alpha=.876) and NASA-TLX (alpha=.746). Moreover, all control variables were tested on their correlation with the TGs using either ANOVA or KW-Test depending on the data characteristics. Since no statistically significant correlation was identified, we conclude that the TGs are balanced and independent regarding the following control variables: education, age, English language proficiency (listening and writing), occupation, computer use, computer gaming habits, VR experience, assembly experience, virtual assembly experience, display size, computer pointing device, recruiting channel and pre-motivation. Each control variable's mean value and standard deviation are listed in the Appendix.

B. Hypothesis H1 – Performance Indicators

The results regarding an effect of the level of gamification intensity on the learning performance indicators can be seen in Figure IV.1.

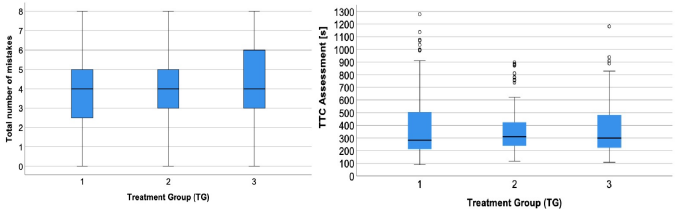

Figure IV.1: Boxplot diagrams for objective learning outcomes per treatment group

The total number of mistakes among the participant sample ranges from 0 to 8 mistakes and yields a median number of four mistakes which indicates sufficient variance. Regarding the three treatment groups, TG1 yielded the lowest mean number of mistakes (3.74, Sd=1.84), followed by TG2 (4.11, Sd=1.82) and TG3 (4.21, Sd=1.67). While this indicates a reverse tendency compared to H1a, the differences were not statistically significant based on the Kruskal-Wallis Test yielding an error probability of $p = 0.11 > 0.05$ for the null hypothesis $H1a_0$. H1a is rejected based on the results of our study.

Similar results can be obtained from the second performance indicator, namely time-to-completion. Prior to the data analysis, 10 extreme outliers allocated outside the range [1st quartile – 3* (interquartile range); 3rd quartile + 3* (interquartile range)] were excluded. This exclusion is an additional filter mechanism to eliminate participants that might have faced interruptions or software lagging during the assessment phase of the remote experiment without stating this in the post-survey. The remaining 345 data points ranged from 89s completion time to up to 1278s. The fastest (best) mean completion time occurred in TG2 (359s; Sd=188) followed by TG3 (373s; Sd=218) and TG1 (387s; Sd=248). Yet, the differences were not statistically significant based on the Kruskal-Wallis Test yielding an error probability of $p = 0.889 > 0.05$ for the null hypothesis $H1b_0$. Thus, H1b is rejected based on the results of our study. We did not find statistically significant evidence that implementing higher gamification intensity during the training process enhances learning outcomes as measured by number of mistakes or task completion time.

C. Hypothesis H2: Affective Indicators

The results regarding the effect of the level of gamification intensity on the training's affective factors can be obtained from Figure IV.2.

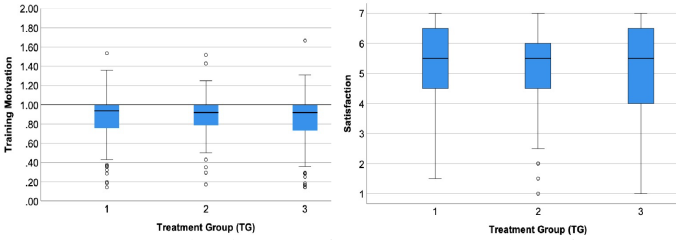

Figure IV.2: Boxplot diagrams for motivation and satisfaction per treatment group

This article has been accepted for publication in IEEE Transactions on Learning Technologies. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TLT.2022.3208936

The resulting training motivation ranges from decreased motivation (0.14) to increased motivation (1.67). Regarding the three treatment groups, TG2 yields the largest mean training motivation value (0.883; Sd=0.199), followed by TG1 (0.879; Sd=0.236) and TG3 (0.851; Sd=0.293). Yet, the differences in training motivation across the different levels of gamification intensity were not statistically significant, as determined by the Kruskal-Wallis Test ($p = 0.791 > 0.05$). Consequently, hypothesis H2a is rejected based on the results of this study.

Similar to motivation, learners' satisfaction also represents a subjective metric which usually changes throughout the training process [8]. As can be seen in Figure IV.2, TG1 yields the highest mean satisfaction value (5.44; Sd=1.29), followed by TG2 (5.27; Sd=1.25) and TG3 (5.07; Sd=1.59). Nevertheless, the marginal differences were not statistically significant based on the Kruskal-Wallis test yielding an error probability of $p = 0.316 > 0.05$ for the respective null hypothesis $H2b_0$. Consequently, the null hypothesis stands and H2b is rejected based on our findings.

The results of the third affective indicator, learners' perceived self-efficacy, are visualized in Figure IV.3.

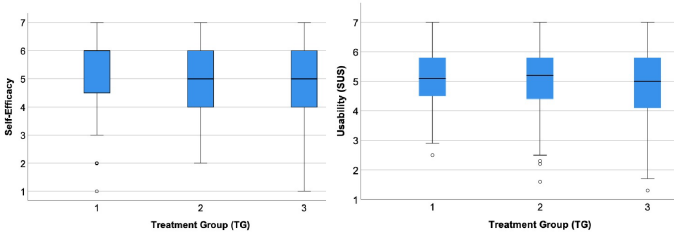

Figure IV.3: Boxplot diagrams for self-efficacy and usability per treatment group

As indicated in Figure IV.3, TG1 yielded the highest mean perceived self-efficacy (5.22; Sd=1.50), followed by TG2 (4.98; Sd=1.44), and TG3 (4.78; Sd=1.49). Conducting a pairwise Mann-Whitney U-Test yielded statistically significant differences between TG1 and TG3 ($p = 0.01$) but not between TG2 and TG3 ($p = 0.265$). Furthermore, the effect size is $r = 0.136$ which can be classified as a low effect size. Consequently, the difference in self-efficacy was statistically significant for learners in the non-gamified training environment compared to the advanced gamified. Thus, hypothesis H2c was rejected based on this study's results.

D. Hypothesis H3: Usability

The results regarding an effect of the level of gamification intensity and the training's perceived usability can be obtained from Figure IV.3. Based on the System Usability Scale (SUS) conducted after the training session, TG1 yielded the highest mean perceived usability (5.14; Sd=0.99). In contrast, the gamified learning environments TG2 (5.06; Sd=1.11) and TG3 (4.86; Sd=1.23) were described as less usable. Even though this indicates that gamification decreases the perceived usability, the differences between the gamification levels were not statistically significant, as determined by the Kruskal-Wallis Test ($p = 0.237 > 0.05$). As a result, hypothesis H3 is rejected based on our findings.

E. Hypothesis H4: Cognitive Load

The results regarding an effect of the level of gamification intensity on the training session's induced cognitive load on learners can be seen in Figure IV.4.

The cognitive load values were computed by using the NASA-RTLX which reflects the mean value across all sub-components. The lowest (best) cognitive load value was perceived in TG1 (44.4; Sd=20.4) followed by TG3 (45.9; Sd=19.5) and TG2 (46.0; Sd=20.9). Even though the results indicate alignment with the hypothesis that gamification increases the cognitive load, the differences were not statistically significant based on the ANOVA yielding an error probability of $p = 0.784 > 0.05$. Consequently, the respective null hypothesis stands and hypothesis H4 is rejected.

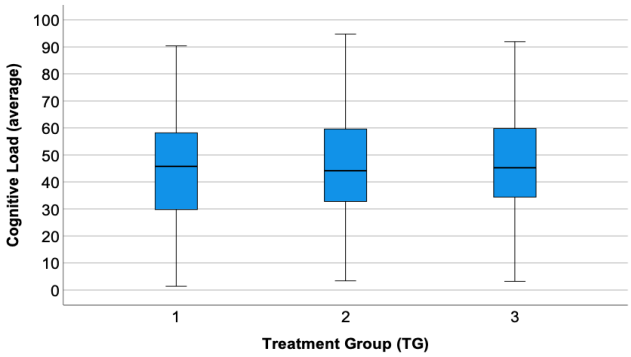

Figure IV.4: Boxplot diagrams for cognitive load per treatment group

F. Further factors influencing Learning

Additional factors influencing learning were identified during the data analysis. These factors include the computer pointing device, screen size, and the learner's computer gaming habits.

Participants who used a computer mouse ($n = 300$) for their computer interaction performed significantly better than those using a trackpad (n=50), as measured by the completion time during the post-training assessment. As indicated in Figure IV.5, computer mouse users yielded a mean TTC of 354s, while the trackpad users needed 444s to complete the task. These differences were proven to be statistically significant by the asymptotic Kruskal-Wallis Test ($p < 0.01$). Furthermore, using a computer mouse significantly increased the perceived usability of the training session, as computed by the KW-Test ($p = 0.046 < 0.05$). On average, computer mouse users stated a usability value of 5.08, while trackpad users only indicated 4.72. Both effects, TTC ($r = 0.179$) and usability ($r = 0.106$) can be categorized as small effect sizes.

Another equipment-related factor influencing the learning experience and outcome of the virtual training was the desktop's screen size. As can be seen in Figure IV.5, participants who used a large display (> 13inch) performed significantly better in the post-training assessment than those who used a small display (<= 13inch), as measured by TTC. The large display participants (n=301) yielded an average TTC of 355s, while the small display (n=49) participants needed 445s to complete the task. The statistical significance of these differences was computed using the KW-Test ($p = 0.023$) with a small effect size ($r = 0.122$).

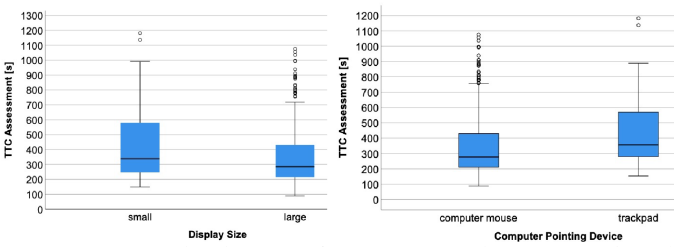

Figure IV.5: Boxplot diagrams for time-to-completion per technical equipment

Lastly, participants' computer gaming habits significantly influenced learning outcomes as measured by number of mistakes, TTC, satisfaction, self-efficacy, and usability. Computer gaming habits were investigated by assessing daily hours spent gaming during the pre-survey. As can be seen in Figure IV.6, trainees were asked to choose one out of six options ranging from less than one hour to more than eight hours per day.

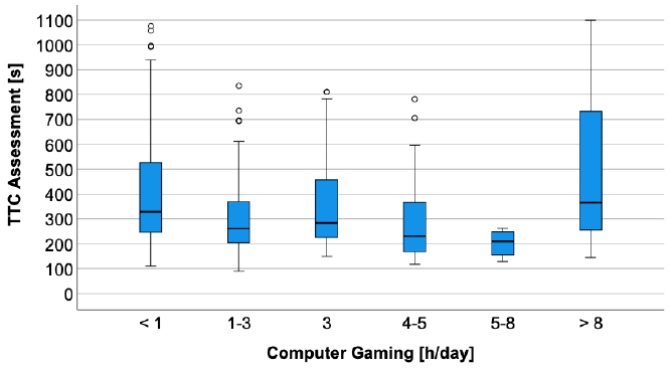

Figure IV.6: Boxplot diagrams for time-to-completion per computer gaming habit group

Since the groups playing the most [5-8h; >8] contained too few participants (n < 20) to yield valid results, they were excluded from the data analysis.

Table IV.1: Effect sizes of gaming habits' influence on time-to-completion

Group A	Group B	Effect size (r)	Tendency
Group 1 (< 1 h/d)	Group 2 (1-3 h/d)	r = .248 (medium)	A >B
Group 1 (< 1 h/d)	Group 4 (4-5 h/d)	r = .183 (small)	A > B
Group 3 (3 h/d)	Group 4 (4-5 h/d)	r = .120 (small)	A > B

As summarized in Table IV.1, the pairwise comparison using the U-Test revealed significant differences with small to medium effect sizes across the computer gaming habit groups regarding the completion time in the post-training learning assessment.

Similar findings were made for the number of assembly mistakes and affective learning factors. In summary, computer gaming habits influence satisfaction, self-efficacy, usability, and learning outcomes (total number of mistakes, TTC), as can be seen in the matrix in Figure IV.7.

Learning Out-come	Group A	Group B	Effect Size	Tendency
Satisfaction	Group 1 (< 1 h/d)	Group 2 (1-3 h/d)	r = .293 (medium)	A<B
	Group 1 (< 1 h/d)	Group 3 (3 h/d)	r = .181 (small)	A<B
	Group 1 (< 1 h/d)	Group 4 (4-5 h/d)	r = .120 (small)	A<B
Self-Efficacy	Group 1 (< 1 h/d)	Group 2 (1-3 h/d)	r = .294 (medium)	A<B
	Group 1 (< 1 h/d)	Group 3 (3 h/d)	r = .155 (small)	A<B
Usability	Group 1 (< 1 h/d)	Group 2 (1-3 h/d)	r = .260 (medium)	A<B
Mistakes	Group 1 (< 1 h/d)	Group 3 (3 h/d)	r = .160 (small)	A<B

Figure IV.7: Influence of computer gaming habits on learning outcomes

To summarize our findings, Figure IV.8 lists all variable correlations identified in this study in a matrix format. An "x" symbolizes that a statistically significant correlation was identified based on the common error probability of p = 0.05. Brown colored fields were not tested.

		Outcome (Dependent Variables)							
Independent Variables		Mis-takes	TTC	Motiv.	Sat-isf.	Self-Eff.	Usability	CL	Reten-tion
Design	TG					x			
Tech. Equip.	Pointing Device		x				x		
	Display Size		x						
Per-sona	Game Ex-perience	x	x		x	x	x		

Figure IV.8: Variable correlation

G. *Hypothesis H5: Knowledge Retention (Retention Test)*

In total, 110 out of the 355 participants completed the retention test within a pre-defined retention interval and thereby meeting the retention test's ACQs requirements. The retention sample was distributed across the treatment groups with TG1 (n = 37), TG (n = 38), and TG3 (n = 35). The average time-to-completion for the whole retention session was 31.52min. (SD = 226.56s). Participants aged between 18 and 54 years took part in the retention test, resulting in an average age of 29.76 years (SD = 8.16). The sample's gender distribution is 66.4% male, 31.8% female and 1.8% non-binary. Despite the international mix of the retention test's sample, consisting of participants from 10 different countries, nearly half of the sample were from Germany (43.6%), followed by the United States of America (34.5%), United Kingdom (11.8%), and India (4.5%). In terms of current occupation, the largest group of this sample were office workers (42.7%), followed by university students (29.1%). The retention sample's treatment groups were independent regarding all control variables.

As can be seen in Figure IV.8, the number of mistakes in the knowledge retention test ranges from 0 mistakes to the maximum amount of 14 with median values in the range of 5-8. TG1 yielded the best knowledge retention including the lowest average number of mistakes (M=5.68; Sd=3.54), followed by TG2 (M=6.45; Sd=3.24) and TG3 (M=7.09;

Sd=3.29). Yet, these differences were not statistically significant based on the Kruskal-Wallis Test ($p = 0.248 > 0.05$) and H5 is rejected. Implementing gamification during a non-immersive virtual training session therefore did not improve learning outcomes, as measured by knowledge decay over time based on our findings.

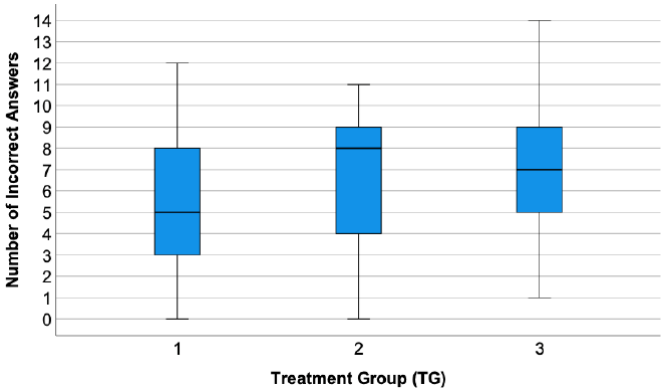

Figure IV.9: Boxplot diagrams for retention test performance (number of incorrect answers) per treatment group

V. DISCUSSION

The result that gaming intensity, which included systematically adding game elements proven to be effective in previous studies, did not improve learning for an industrial training is surprising and needs to be discussed. There have been previous studies finding mixed results for gamification [1], [2], [85], [86]. To identify potential causes for our finding, we conducted several additional analyses.

Hypotheses H1a (number of mistakes) and H1b (TTC) were rejected based on our study's experimental data. Our findings contrast with the literature which has identified a reduced number of technical assembly mistakes for gamified virtual reality training [4]. A possible explanation for our non-significant result could be an insufficient level of task complexity and the related ceiling effect [87]. Our study's assembly task can be categorized as medium-level complexity ($C_{Product} = 6.34$) [66]. Since previous virtual assembly training research indicates more significant results with rising task complexity ($C_{Product} > 7$), our assembly might not have met the required complexity threshold [88]. However, a counterargument to this line of reasoning is that the same task we used in the experiment yielded significant results in another experimental study [89], which in turn makes a ceiling effect explanation unlikely.

In contrast to other studies, increased gamification intensity did not increase affective learning factors measured by motivation (H2a), satisfaction (H2b), and self-efficacy (H2c) [49], [55], [90]. Interestingly, our research even identified a reverse relationship between gamification intensity and self-efficacy (H2b). The advanced gamified learning environment (TG3) induced significantly lower self-efficacy compared to the non-gamified environment (TG1). While the absence of a correlation between gamification intensity and self-efficacy is also common in other studies, the presence of a reverse effect is a surprising result [91]. A possible explanation for this result could be that even though our set of game elements aligns with BNT, gamification's competitiveness could have lowered learners' self-efficacy [32]. The effect size is very small ($r = .136$) which alleviates the implications of this finding. Lastly, the scales for satisfaction and self-efficacy only include two questions which also alleviates the implications.

Most research involving gamification and usability focuses on the influence of usability on gamification outcomes. Game usability was already identified as a moderating factor influencing learning [61]. With hypothesis H3 investigating the inverse causality we open an unexplored aspect. Since no significant causality was identified, we cannot confirm that gamification intensity influences perceived system usability. Similarly, perceived cognitive load, as measured by NASA-RTLX, did not vary significantly across the TGs (H4). This is contrary to previous studies detecting an increase in cognitive load due to implementing gamification [63]. This discrepancy could be explained by considering two limitations. First, the results of cognitive load tests vary across the different measurement methods [25]. Second, the investigation of the previous hypotheses indicates that our study's game elements may not have provided significantly different training experiences across the TGs. It seems reasonable that cognitive load also does not vary across the TGs.

No significant differences in knowledge retention were measured comparing the different levels of gamification intensity (H5), which contrasts with previous findings [60], [64], [65]. Yet, our finding aligns with the absence of learning enhancement investigated by H1. Similar to H1, insufficient task complexity could be a moderator blurring all gamification effects on learning in the first place and, subsequently, in the retention test [88] – although this seems unlikely in this case as mentioned above. The retention interval likely also influenced the level of knowledge decay [14], which leads to a second possible explanation for the absence of significant differences: the diversity of retention interval duration. Congruent with previous literature [13], our study identified a positive effect of retention interval duration and the level of knowledge decay. Therefore, varying retention periods ranging from min. 21 days to max. 26 days after the learning experience across the TGs may blur knowledge decay results.

The fact that no impact of gamification was detected across all dependent variables leads to a more general discussion. The absence of significant differences could have two potential reasons. First, the different treatment groups did yield different training experiences and learning outcomes, but the experimental setup failed to measure them. Second, the different treatment groups did not yield significantly different learning experiences. Regarding the first case, previous studies have already applied the exact same metrics and scales and yielded significant results [17], [20], [75]. Furthermore, the objective metrics of time-to-completion (H1b) and number of mistakes (H1a and H5) were tracked automatically, which forestalls any measurement inconsistencies. Increasing the maximum number of mistakes of 10 for the main experiment and 14 for the retention test could yield more differentiated results regarding the mistakes metric. Nevertheless, our research did identify a statistically significant correlation between the outcome metrics and the participants' technological equipment during the experiment, as well as their

computer gaming background. This supports the assumption that our study's metrics are intact and a suitable tool for measuring learning outcomes.

Assuming the three treatment groups' training designs did not influence learning outcomes, two other factors must be considered for remote experiments. The virtual training design (software) and the real environment in which the participant executed the training (hardware and environment). Both factors may contribute individually but also additively to the participant's training experience, which is directly related to the outcome metrics [83]. Both factors might balance and reinforce each other's influence on learning outcomes, potentially causing outcome metric limitations. Even though participants who faced severe software lagging or disruptions during the execution of the remote experiment were excluded from the data analysis, similar experimental settings across participants are difficult to guarantee in remote experiments.

Regarding the virtual design of the experimental study, the general structure of the experiment including the components of pre-survey, familiarization, training, assessment, post-survey and retention test have been shown to be a suitable experimental structure [16]. This leads to the consideration that the training's TG design could have provided insufficient differences to measure statistically significant effects, as TGs only differed based on game elements. However, the virtual experiment was tested through several user-studies and pilot experiments and all game elements were successfully used in previous studies already [2], [35], [53], [55], [77], [92]. This leads us to a discussion of whether the five-factor game element selection methodology applied in this research represented a suitable technique. Compared to essentially random game element selection in most other previous studies, it seems unlikely that a more logically structured approach caused worse results. Other factors, such as a maximum number of elements to prevent over-stimulation or applying other theories instead of SDT and BNT, could be considered in the future. It would be interesting to see if other selection approaches, such as tangible and intangible game elements will yield stronger results for virtual assembly training [93].

Assuming the virtual design itself provides different learning experience, participants' individual technical equipment and environment during the training could be limiting factors blurring the results. Since our experimental findings support this theory (see Figure IV.5), the assumption that the hardware and environment of the participants balanced the possibly positive effects induced by the game elements seems reasonable. This aligns with previous studies identifying, for example, display size as a factor influencing users' performance, especially in 3D virtual environments [94].

Another potential factor could be participants' computer gaming background. As our data analysis indicates (see Figure IV.6), the number of weekly hours spent on computer games influenced learning outcomes. Even though sorting participants into gamers and non-gamers did not yield significance (possibly because it decreased the sample size), the influence of their computer gaming background could have also balanced positive effects induced by the game elements. This proposition is supported by previous research with similar findings [95]. In addition to participants' computer gaming experience, other personal background factors or metrics related to computer gaming could be relevant, such as gender, age, or the attitude towards game-based learning and computer gaming, as investigated by previous studies [40], [96]. This idea aligns with theories proposing that gamification's effects depend highly on the individual user's personality and that personalized gamification yields the best results [31]. Attempts to investigate a person's individual gamification attitude in education were already made but further research in needed [97]. According to recent studies, additional factors impacting individual gamification responses include technology acceptance and task technology fit [98].

In summary, our research delivers three key findings: Firstly, increasing gamification intensity by selecting 14 game elements for gamified virtual training based on a five-factor method did not result in learning improvement. Secondly, users' technological equipment influenced the virtual learning experience and outcome. Thirdly, users' computer gaming habits and experiences influenced learning outcomes.

VI. Limitations and Outlook

A. Limitations

With a total of 355 participants during the main experiment (110 in the retention test) our study ranks in the field of large virtual training studies. However, we acknowledge it remains a small sample compared to all people possibly affected by desktop- or mobile-based procedural knowledge acquisition. Like most studies in more technical training, the sample was gender biased, including only 27.3% female and 0.8% non-binary participants. To the best of our knowledge no evidence exists how a user's gender might impact gamification's effect size. A strong representation of Germany and the USA accounting for more than 2/3 of participants limits cross-cultural comparisons.

Another limitation is the participants' background. Our participant sample is diverse in terms of background, assembly experience, and computer gaming habits. While this might be beneficial for analyzing the variables' impact, the results' transferability to a technical worker group is limited.

The platform MTurk that was used for our experiment is not uncontroversial regarding data quality [82]. Even though we filtered fraudulent MTurk participants using ACQs and metadata (e.g., unique user IDs), dishonest answers and unmotivated performance are impossible to prevent completely. While some external effects can be filtered, other unexpected ones might pass undetected [99]. Future studies should be conducted in fully controlled environments to assure all extraneous variables are accounted for.

The learning assessment took place remotely and virtually. Even though previous studies which have investigated knowledge transfer from virtual assembly to the physical assembly have concluded that knowledge in virtual assembly can be transferred, assessing knowledge only with a virtual assembly remains a limitation and a real-world physical assembly could yield different results.

A set of game elements was chosen to represent the concept of gamification. While this is the case for basically all

gamification studies, any conclusion regarding the overarching concept of gamification based on a specific set of game elements is limiting validity. Replacing or altering only one of the 14 elements applied in this study could result in different results, as shown in previous studies [1], [2]. Thus, insights regarding the impact of individual game elements and their interactions are limited.

B. Future Research Opportunities

More research is needed to better understand gamification mechanisms and design guidelines. The diversity in gamification's effect size on learning outcome is the result of the internal variety of gamification itself. Considering just the 35 different game elements identified in our literature review, mathematically more than 30 billion different combinations could be investigated. Such an investigation does not seem feasible with human participants, though theoretically-based subsets of game elements could be investigated. Such a fractional factorial design seems currently the only possible way to advance knowledge.

Our research applied SDT's micro-theory BNT as the framework to identify suitable game elements. Other theoretical frameworks such as Flow Theory [100], Social Comparison Theory [38] and the Lean Gamification Canvas [101] could be investigated in future research. The learner's personality and general attitude towards gaming could be further investigated to identify personality-based prerequisites or predictors for gamification effects. Such approaches could include broad concepts such as the Big Five Personality Test [102] or learning specific frameworks such as the Myers-Briggs Type Indicator personality test [103].

Future research could further examine the general feasibility of remote non-immersive virtual reality experiments and the influence of participants' technical equipment [104]. In addition, self-assessed variables could be matched with physiological data to prevent subjective bias in participants' responses. With the world's working and educational life shifting to more hybrid work settings, it is important to consider the potential of remote experiments and some promising remote lab concepts already exist [99].

The transferability of our findings to other tasks with different levels of complexity and types of knowledge could be further investigated. While other tasks refer to different assembly objects, different types of knowledge include moving from procedural knowledge to other levels such as conceptual knowledge defined in Krathwohl's taxonomy [105]. Another experimental parameter worth considering is the duration of the individual experiment (of the assembly task) and the retention period. A longitudinal study investigating learning outcomes through several repeating or type-like training sessions over a longer period could be valuable too.

REFERENCES

[1] S. Bai, K. F. Hew, and B. Huang, "Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts," *Educ. Res. Rev.*, vol. 30, no. December 2019, p. 100322, 2020, doi: 10.1016/j.edurev.2020.100322.

[2] F. F. H. Nah, Q. Zeng, V. R. Telaprolu, A. P. Ayyappa, and B. Eschenbrenner, "Gamification of education: A review of literature," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8527 LNCS, pp. 401–409, 2014, doi: 10.1007/978-3-319-07293-7_39.

[3] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining 'gamification,'" *Proc. 15th Int. Acad. MindTrek Conf. Envisioning Futur. Media Environ. MindTrek 2011*, no. September, pp. 9–15, 2011, doi: 10.1145/2181037.2181040.

[4] F. Palmas, D. Labode, D. A. Plecher, and G. Klinker, "Comparison of a gamified and non-gamified virtual reality training assembly task," *2019 11th Int. Conf. Virtual Worlds Games Serious Appl. VS-Games 2019 - Proc.*, no. March 2016, p. 1DUUMY, 2019, doi: 10.1109/VS-Games.2019.8864583.

[5] S. Oberdörfer, D. Heidrich, and M. E. Latoschik, "Usability of Gamified Knowledge Learning in VR and Desktop-3D," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13, doi: 10.1145/3290605.3300405.

[6] C. J. C. Jabbour, "Environmental training in organisations: From a literature review to a framework for future research," *Resour. Conserv. Recycl.*, vol. 74, pp. 144–155, 2013, doi: 10.1016/j.resconrec.2012.12.017.

[7] D. Kirkpatrick, *Evaluating Training Programs*, 3rd ed. Berrett-Koehler Publishers, 2006.

[8] T. Sitzmann and J. M. Weinhardt, "Approaching evaluation from a multilevel perspective: A comprehensive analysis of the indicators of training effectiveness," *Hum. Resour. Manag. Rev.*, vol. 29, no. 2, pp. 253–269, 2019, doi: 10.1016/j.hrmr.2017.04.001.

[9] S. Ginzburg and E. M. Dar-El, "Skill retention and relearning – a proposed cyclical model," *J. Work. Learn.*, vol. 12, no. 8, pp. 327–332, 2000, doi: 10.1108/13665620010378822.

[10] J. W. Kim, F. E. Ritter, and R. J. Koubek, "An integrated theory for improved skill acquisition and retention in the three stages of learning," *Theor. Issues Ergon. Sci.*, vol. 14, no. 1, pp. 22–37, 2013, doi: 10.1080/1464536X.2011.573008.

[11] A. N. Trani *et al.*, "Modeling and simulation of skill decay at the organizational team level," *Proc. Hum. Factors Ergon. Soc.*, vol. 2017-Octob, no. November, pp. 740–744, 2017, doi: 10.1177/1541931213601670.

[12] N. Cowan, "What are the differences between long-term, short-term, and working memory?," in *Progress in Brain Research*, vol. 169, no. 07, Elsevier, 2008, pp. 323–338.

[13] J. W. Kim, R. J. Koubek, and F. E. Ritter, "Investigation of procedural skills degradation from different modalities," *Proc. ICCM 2007 - 8th Int. Conf. Cogn. Model.*, pp. 255–260, 2007.

[14] C. E. Küpper-Tetzel, I. V. Kapler, and M. Wiseheart, "Contracting, equal, and expanding learning schedules: The optimal distribution of learning sessions depends on retention interval," *Mem. Cogn.*, vol. 42, no. 5, pp. 729–741, 2014, doi: 10.3758/s13421-014-0394-1.

[15] E. Ai-Lim Lee, K. W. Wong, and C. C. Fung, "How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach," *Comput. Educ.*, vol. 55, no. 4, pp. 1424–1442, 2010, doi: 10.1016/j.compedu.2010.06.006.

[16] G. Makransky and G. B. Petersen, "Investigating the process of learning with desktop virtual reality: A structural equation modeling approach," *Comput. Educ.*, vol. 134, no. February, pp. 15–30, 2019, doi: 10.1016/j.compedu.2019.02.002.

[17] D. Jia, A. Bhatti, and S. Nahavandi, "The impact of self-efficacy and perceived system efficacy on effectiveness of virtual training systems," *Behav. Inf. Technol.*, vol. 33, no. 1, pp. 16–35, 2014.

[18] A. Sutcliffe, B. Gault, and J. E. Shin, "Presence, memory and interaction in virtual environments," *Int. J. Hum. Comput. Stud.*, vol. 62, no. 3, pp. 307–327, 2005, doi: 10.1016/j.ijhcs.2004.11.010.

[19] P. C. Sun, R. J. Tsai, G. Finger, Y. Y. Chen, and D. Yeh, "What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction," *Comput. Educ.*, vol. 50, no. 4, pp. 1183–1202, 2008, doi: 10.1016/j.compedu.2006.11.007.

[20] T. S. Tullis and J. N. Stetson, "A Comparison of Questionnaires for Assessing Website Usability ABSTRACT : Introduction," *Usability Prof. Assoc. Conf.*, no. June 2006, pp. 1–12, 2004.

[21] J. Sweller, *Cognitive Load Theory*, vol. 55. Elsevier Inc., 2011.

[22] C. E. Zwilling, "Forgetting in short term memory : the effect of time FORGETTING IN SHORT-TERM MEMORY : THE EFFECT OF TIME presented to the Faculty of the Graduate School at the

University of Missouri-Columbia In Partial Fulfillment of the Requirements for the Degree Mast," no. January 2008, 2017.

[23] M. A. Just and P. A. Carpenter, "A capacity theory of comprehension: Individual differences in working memory," *Psychol. Rev.*, vol. 99, no. 1, pp. 122–149, 1992, doi: 10.1037/0033-295X.99.1.122.

[24] W. Schnotz and C. Kürschner, "A eeconsideration of cognitive load theory," *Educ. Psychol. Rev.*, vol. 19, no. 4, pp. 469–508, 2007, doi: 10.1007/s10648-007-9053-4.

[25] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educ. Psychol.*, vol. 38, no. 1, pp. 63–71, 2003, doi: 10.1207/S15326985EP3801_8.

[26] L. M. Naismith, J. J. H. Cheung, C. Ringsted, and R. B. Cavalcanti, "Limitations of subjective cognitive load measures in simulation-based procedural training," *Med. Educ.*, vol. 49, no. 8, pp. 805–814, 2015, doi: 10.1111/medu.12732.

[27] S. Rubio, E. Díaz, J. Martín, and J. M. Puente, "Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods," *Appl. Psychol.*, vol. 53, no. 1, pp. 61–86, 2004, doi: 10.1111/j.1464-0597.2004.00161.x.

[28] O. R. Runswick, A. Roca, A. Mark Williams, N. E. Bezodis, A. P. Mcrobert, and J. S. North, "The impact of contextual information and a secondary task on anticipation performance: An interpretation using cognitive load theory," *Appl. Cogn. Psychol.*, vol. 32, no. 2, pp. 141–149, 2018, doi: 10.1002/acp.3386.

[29] Z. Zainuddin, "Students' learning performance and perceived motivation in gamified flipped-class instruction," *Comput. Educ.*, vol. 126, pp. 75–88, 2018, doi: https://doi.org/10.1016/j.compedu.2018.07.003.

[30] M. Ninaus, K. Kiili, G. Wood, K. Moeller, and S. E. Kober, "To Add or Not to Add Game Elements? Exploring the Effects of Different Cognitive Task Designs Using Eye Tracking," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 847–860, 2020, doi: 10.1109/TLT.2020.3031644.

[31] A. C. T. Klock, I. Gasparini, M. S. Pimenta, and J. Hamari, "Tailored gamification: A review of literature," *Int. J. Hum. Comput. Stud.*, vol. 144, no. June, 2020, doi: 10.1016/j.ijhcs.2020.102495.

[32] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," *Comput. Human Behav.*, vol. 69, pp. 371–380, 2017, doi: 10.1016/j.chb.2016.12.033.

[33] R. Orji, L. E. Nacke, and C. Di Marco, "Towards personality-driven persuasive health games and gamified systems," *Conf. Hum. Factors Comput. Syst. - Proc.*, vol. 2017-May, pp. 1015–1027, 2017, doi: 10.1145/3025453.3025577.

[34] K. Berkling and C. Thomas, "Gamification of a software engineering course and a detailed analysis of the factors that lead to it's failure," *2013 Int. Conf. Interact. Collab. Learn. ICL 2013*, no. January, pp. 525–530, 2013, doi: 10.1109/ICL.2013.6644642.

[35] S. Nicholson, "A Receipt for Meaningful Gamification," in *Gamification in Education and Business*, T. Reiners and L. Wood, Eds. Springer, Cham., 2015.

[36] J. Nakamura and M. Csikszentmihalyi, "Flow theory and research.," in *Oxford handbook of positive psychology, 2nd ed.*, New York, NY, US: Oxford University Press, 2009, pp. 195–206.

[37] B. Gnauk, L. Dannecker, and M. Hahmann, "Leveraging gamification in demand dispatch systems," *ACM Int. Conf. Proceeding Ser.*, pp. 103–110, 2012, doi: 10.1145/2320765.2320799.

[38] E. Molleman, A. Nauta, and B. P. Buunk, "Social comparison-based thoughts in groups: Their associations with interpersonal trust and learning outcomes," *J. Appl. Soc. Psychol.*, vol. 37, no. 6, pp. 1163–1180, 2007, doi: 10.1111/j.1559-1816.2007.00207.x.

[39] E. A. Locke and G. P. Latham, "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey," *Am. Psychol.*, vol. 57, no. 9, pp. 705–717, 2002, doi: 10.1037/0003-066X.57.9.705.

[40] R. N. Landers, K. N. Bauer, R. C. Callan, and M. B. Armstrong, "Psychological Theory and the Gamification of Learning," in *Gamification in Education and Business*, T. Reiners and L. C. Wood, Eds. Cham: Springer International Publishing, 2015, pp. 165–186.

[41] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," *Psychol. Inq.*, vol. 11, no. 4, pp. 227–268, 2000, doi: 10.1207/S15327965PLI1104_01.

[42] R. Ryan and E. Deci, "Active Human Nature: Self-Determination Theory and the Promotion and Maintenance of Sport, Exercise and Health," in *Intrinsic Motivation and Self-Determination in Exercise and Sport*, Human Kinetics Europe Ltd., 2007, pp. 1–19.

[43] E. A. Sanli, J. T. Patterson, S. R. Bray, and T. D. Lee, "Understanding self-controlled motor learning protocols through the self-determination theory," *Front. Psychol.*, vol. 3, no. JAN, pp. 1–17, 2013, doi: 10.3389/fpsyg.2012.00611.

[44] B. Marin, J. Frez, J. Cruz-Lemus, and M. Genero, "An Empirical Investigation on the Benefits of Gamification in Programming Courses," *ACM Trans. Comput. Educ.*, vol. 19, no. 1, Nov. 2018, doi: 10.1145/3231709.

[45] Z. Zainuddin, S. K. W. Chu, M. Shujahat, and C. J. Perera, "The impact of gamification on learning and instruction: A systematic review of empirical evidence," *Educ. Res. Rev.*, vol. 30, no. March, 2020, doi: 10.1016/j.edurev.2020.100326.

[46] J. R. Rachels and A. J. Rockinson-Szapkiw, "The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy," *Comput. Assist. Lang. Learn.*, vol. 31, no. 1–2, pp. 72–89, 2018, doi: 10.1080/09588221.2017.1382536.

[47] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? - A literature review of empirical studies on gamification," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, pp. 3025–3034, 2014, doi: 10.1109/HICSS.2014.377.

[48] K. Seaborn and D. I. Fels, "Gamification in theory and action: A survey," *Int. J. Hum. Comput. Stud.*, vol. 74, pp. 14–31, 2015, doi: 10.1016/j.ijhcs.2014.09.006.

[49] E. D. Mekler, F. Brühlmann, A. N. Tuch, and K. Opwis, "Towards understanding the effects of individual gamification elements on intrinsic motivation and performance," *Comput. Human Behav.*, vol. 71, pp. 525–534, 2017, doi: 10.1016/j.chb.2015.08.048.

[50] G. F. Tondello, R. R. Wehbe, L. Diamond, M. Busch, A. Marczewski, and L. E. Nacke, "The gamification user types Hexad scale," *CHI Play 2016 - Proc. 2016 Annu. Symp. Comput. Interact. Play*, pp. 229–243, 2016, doi: 10.1145/2967934.2968082.

[51] M. A. Berg and S. A. Petersen, "Exploiting psychological needs to increase motivation for learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8101 LNCS, pp. 260–265, 2013, doi: 10.1007/978-3-642-40790-1_26.

[52] Z. Liang, K. Zhou, and K. Gao, "Development of Virtual Reality Serious Game for Underground Rock-Related Hazards Safety Training," *IEEE Access*, vol. 7, pp. 118639–118649, 2019, doi: 10.1109/ACCESS.2019.2934990.

[53] A. S. Bjerrum, O. Hilberg, T. van Gog, P. Charles, and B. Eika, "Effects of modelling examples in complex procedural skills training: A randomised study," *Med. Educ.*, vol. 47, no. 9, pp. 888–898, 2013, doi: 10.1111/medu.12199.

[54] É. Lavoué, B. Monterrat, M. Desmarais, and S. George, "Adaptive Gamification for Learning Environments," *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 16–28, 2019, doi: 10.1109/TLT.2018.2823710.

[55] C. Santos, S. Almeida, L. Pedro, M. Aresta, and T. Koch-Grunberg, "Students' perspectives on badges in educational social media platforms: the case of SAPO campus tutorial badges," *Proc. - 2013 IEEE 13th Int. Conf. Adv. Learn. Technol. ICALT 2013*, pp. 351–353, 2013, doi: 10.1109/ICALT.2013.108.

[56] G. Wulf and R. Lewthwaite, "Optimizing performance through intrinsic motivation and attention for learning: The OPTIMAL theory of motor learning," *Psychon. Bull. Rev.*, vol. 23, no. 5, pp. 1382–1414, 2016, doi: 10.3758/s13423-015-0999-9.

[57] D. Holmes, D. Charles, P. Morrow, S. McClean, and S. McDonough, "Rehabilitation Game Model for Personalised Exercise," *Proc. - 2015 Int. Conf. Interact. Technol. Games, ITAG 2015*, no. June 2016, pp. 41–48, 2016, doi: 10.1109/iTAG.2015.11.

[58] C. Butler, "A framework for evaluating the effectiveness of gamification techniques by personality type," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8527 LNCS, pp. 381–389, 2014, doi: 10.1007/978-3-319-07293-7_37.

[59] T. de A. G. Grangeia, B. de Jorge, D. Cecílio-Fernandes, R. A. Tio, and M. A. de Carvalho-Filho, "Learn+Fun! Social Media and Gamification sum up to Foster a Community of Practice during an

Emergency Medicine Rotation," *Heal. Prof. Educ.*, vol. 5, no. 4, pp. 321–335, Dec. 2019, doi: 10.1016/j.hpe.2018.11.001.

[60] L. M. Putz, F. Hofbauer, and H. Treiblmaier, "Can gamification help to improve education? Findings from a longitudinal study," *Comput. Human Behav.*, vol. 110, no. March, p. 106392, 2020, doi: 10.1016/j.chb.2020.106392.

[61] M. Rajanen and D. Rajanen, "Usability benefits in gamification," *CEUR Workshop Proc.*, vol. 1857, pp. 87–95, 2017.

[62] M. Rajanen and D. Rajanen, "Heuristic evaluation in game and gamification development," *CEUR Workshop Proc.*, vol. 2186, pp. 159–168, 2018.

[63] Z. Turan, Z. Avinc, K. Kara, and Y. Goktas, "Gamification and education: Achievements, cognitive loads, and views of students," *Int. J. Emerg. Technol. Learn.*, vol. 11, no. 7, pp. 64–69, 2016, doi: 10.3991/ijet.v11i07.5455.

[64] D. Dicheva, C. Dichev, G. Agre, and G. Angelova, "Gamification in education: A systematic mapping study," *Educ. Technol. Soc.*, vol. 18, no. 3, pp. 75–88, 2015.

[65] J. Majuri, J. Koivisto, and J. Hamari, "Gamification of education and learning: A review of empirical literature," *CEUR Workshop Proc.*, vol. 2186, pp. 11–19, 2018.

[66] S. N. Samy and H. Elmaraghy, "A model for measuring products assembly complexity," *Int. J. Comput. Integr. Manuf.*, vol. 23, no. 11, pp. 1015–1027, 2010, doi: 10.1080/0951192X.2010.511652.

[67] J. E. Brough, M. Schwartz, S. K. Gupta, D. K. Anand, R. Kavetsky, and R. Pettersen, "Towards the development of a virtual environment-based training system for mechanical assembly operations," *Virtual Real.*, vol. 11, no. 4, pp. 189–206, 2007, doi: 10.1007/s10055-007-0076-4.

[68] M. Murcia-López and A. Steed, "A Comparison of Virtual and Physical Training Transfer of Bimanual Assembly Tasks," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1574–1583, Apr. 2018, doi: 10.1109/TVCG.2018.2793638.

[69] L. Chittaro and F. Buttussi, "Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 4, pp. 529–538, 2015, doi: 10.1109/TVCG.2015.2391853.

[70] F. Buttussi, T. Pellis, A. Cabas Vidani, D. Pausler, E. Carchietti, and L. Chittaro, "Evaluation of a 3D serious game for advanced life support retraining," *Int. J. Med. Inform.*, vol. 82, no. 9, pp. 798–809, 2013, doi: https://doi.org/10.1016/j.ijmedinf.2013.05.007.

[71] R. A. Noe and S. L. Wilk, "Investigation of the factors that influence employees' participation in development activities.," *J. Appl. Psychol.*, vol. 78, no. 2, pp. 291–302, 1993, doi: 10.1037/0021-9010.78.2.291.

[72] Y. C. Huang, S. J. Backman, K. F. Backman, F. A. McGuire, and D. W. Moore, "An investigation of motivation and experience in virtual learning environments: a self-determination theory," *Educ. Inf. Technol.*, vol. 24, no. 1, pp. 591–611, 2019, doi: 10.1007/s10639-018-9784-5.

[73] S. W. Chou and C. H. Liu, "Learning effectiveness in a Web-based virtual learning environment: A learner control perspective," *J. Comput. Assist. Learn.*, vol. 21, no. 1, pp. 65–76, 2005, doi: 10.1111/j.1365-2729.2005.00114.x.

[74] S. Kasten, L. van Osch, M. Candel, and H. de Vries, "The influence of pre-motivational factors on behavior via motivational factors: a test of the I-Change model," *BMC Psychol.*, vol. 7, no. 1, p. 7, 2019, doi: 10.1186/s40359-019-0283-2.

[75] R. A. Grier, "How high is high? A meta-analysis of NASA-TLX global workload scores," *Proc. Hum. Factors Ergon. Soc.*, vol. 2015-Janua, pp. 1727–1731, 2015, doi: 10.1177/1541931215591373.

[76] D. Cui, D. Whittinghill, A. Fukada, C. Mousas, and N. Adamo, "Effects of Virtual Instructor's Facial Expressions in a 3D Game on Japanese Language Learning," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2021, pp. 401–405, doi: 10.1109/VRW52623.2021.00087.

[77] E. Lunts, "What does the Literature Say about the Effectiveness of Learner Control in Computer-Assisted Instruction ?," 2002.

[78] K. Scheiter, "21 The Learner Control Principle in Multimedia Learning," no. May, pp. 487–512, 2017.

[79] W. N. Dember, T. L. Galinsky, and J. S. Warm, "The role of choice in vigilance performance.," *Bull. Psychon. Soc.*, vol. 30, no. 3, pp. 201–204, 1992, doi: 10.3758/BF03330441.

[80] D. Difallah, E. Filatova, and P. Ipeirotis, "Demographics and dynamics of Mechanical Turk workers," *WSDM 2018 - Proc. 11th ACM Int. Conf. Web Search Data Min.*, vol. 2018-Febua, no. August 2017, pp. 135–143, 2018, doi: 10.1145/3159652.3159661.

[81] I. P. Kan and A. B. Drummey, "Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce," *Comput. Human Behav.*, vol. 83, pp. 243–253, 2018, doi: 10.1016/j.chb.2018.02.005.

[82] J. H. Cheung, D. K. Burns, R. R. Sinclair, and M. Sliter, "Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations," *J. Bus. Psychol.*, vol. 32, no. 4, pp. 347–361, 2017, doi: 10.1007/s10869-016-9458-5.

[83] T. Kozík and M. Šimon, "Preparing and managing the remote experiment in education," in *2012 15th International Conference on Interactive Collaborative Learning (ICL)*, 2012, pp. 1–4, doi: 10.1109/ICL.2012.6402077.

[84] L. B. Christensen, R. B. Johnson, and L. A. Turner, "Research Methods, Design, and Analysis," *Araştırma Yöntemleri Desen ve Anal.*, pp. 217–249, 2014.

[85] N. S. Uletika, B. Hartono, and T. Wijayanto, "Gamification in Assembly Training: A Systematic Review," in *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2020, pp. 1073–1077, doi: 10.1109/IEEM45057.2020.9309791.

[86] H. Cecotti, M. Callaghan, B. Foucher, and S. Joslain, "Serious Game for Medical Imaging in Fully Immersive Virtual Reality," in *2021 IEEE International Conference on Engineering, Technology Education (TALE)*, 2021, pp. 615–621, doi: 10.1109/TALE52509.2021.9678721.

[87] N. Gavish *et al.*, "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," *Interact. Learn. Environ.*, vol. 23, no. 6, pp. 778–798, 2015, doi: 10.1080/10494820.2013.815221.

[88] E. Gallegos-Nieto, H. I. Medellín-Castillo, G. González-Badillo, T. Lim, and J. Ritchie, "The analysis and evaluation of the influence of haptic-enabled virtual assembly training on real assembly performance," *Int. J. Adv. Manuf. Technol.*, vol. 89, no. 1–4, pp. 581–598, 2017, doi: 10.1007/s00170-016-9120-4.

[89] T. Bohné, I. Heine, Ö. Gürerk, C. Rieger, L. Kemmer, and L. Y. Cao, "Perception Engineering Learning With Virtual Reality," *IEEE Trans. Learn. Technol.*, vol. 14, no. 4, pp. 500–514, 2021, doi: 10.1109/TLT.2021.3107407.

[90] R. Brewer, L. Anthony, Q. Brown, G. Irwin, J. Nias, and B. Tate, "Using gamification to motivate children to complete empirical studies in lab environments," *ACM Int. Conf. Proceeding Ser.*, no. June 2014, pp. 388–391, 2013, doi: 10.1145/2485760.2485816.

[91] M. Ortiz-Rojas, K. Chiluiza, and M. Valcke, "Gamification in computer programming: Effects on learning, engagement, self-efficacy and intrinsic motivation," *Proc. 11th Eur. Conf. Games Based Learn. ECGBL 2017*, no. October, pp. 507–514, 2017.

[92] M. Roussou, "Learning by Doing and Learning through Play: An Exploration of Interactivity in Virtual Environments for Children," *Comput. Entertain.*, vol. 2, no. 1, p. 10, Jan. 2004, doi: 10.1145/973801.973818.

[93] M. M. Elaish, M. H. Hussein, L. Shuib, W. F. Wan Ahmad, and K. Becker, "A Proposed Gamification Elements of Educational Games," in *2021 International Conference on Computer Information Sciences (ICCOINS)*, 2021, pp. 14–17, doi: 10.1109/ICCOINS49721.2021.9497179.

[94] T. Ni, D. A. Bowman, and J. Chen, "Increased display size and resolution improve task performance in information-rich virtual environments," *Proc. - Graph. Interface*, vol. 2006, pp. 139–146, 2006.

[95] M. Denden, A. Tlili, F. Essalmi, and M. Jemni, "An investigation of the factors affecting the perception of gamification and game elements," *2017 6th Int. Conf. Inf. Commun. Technol. Accessbility, ICTA 2017*, vol. 2017-Decem, pp. 1–5, 2018, doi: 10.1109/ICTA.2017.8336019.

[96] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification," *Comput. Human Behav.*, vol. 35, pp. 179–188, 2014, doi: 10.1016/j.chb.2014.03.007.

[97] S. A. Andrade Freitas, A. R. T. Lacerda, P. M. R. O. Calado, T. S. Lima, and E. Dias Canedo, "Gamification in education: A methodology to identify student's profile," in *2017 IEEE Frontiers*

in Education Conference (FIE), 2017, pp. 1–8, doi: 10.1109/FIE.2017.8190499.

[98] V. Z. Vanduhe, M. Nat, and H. F. Hasan, "Continuance Intentions to Use Gamification for Training in Higher Education: Integrating the Technology Acceptance Model (TAM), Social Motivation, and Task Technology Fit (TTF)," *IEEE Access*, vol. 8, pp. 21473–21484, 2020, doi: 10.1109/ACCESS.2020.2966179.

[99] N. Aliane, R. V. Pastor, and G. V. Mariscal, "Limitations of Remote Laboratories in Control Engineering Education," *Int. J. Online Eng.*, vol. 6, no. 1, 2010, doi: 10.3991/ijoe.v6i1.1131.

[100] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Comput. Human Behav.*, vol. 54, pp. 170–179, Aug. 2016, doi: 10.1016/j.chb.2015.07.045.

[101] B. H. Yousefi and H. Mirkhezri, "Lean Gamification Canvas : A New Tool for Innovative Gamification Design Process," in *2020 International Serious Games Symposium (ISGS)*, 2020, pp. 1–9, doi: 10.1109/ISGS51981.2020.9375297.

[102] G. Saucier and F. Ostendorf, "Hierarchical subcomponents of the Big Five personality factors: A cross-language replication.," *Journal of Personality and Social Psychology*, vol. 76, no. 4. American Psychological Association, US, pp. 613–627, 1999, doi: 10.1037/0022-3514.76.4.613.

[103] I. B. Myers, *The Myers-Briggs Type Indicator: Manual (1962).* Palo Alto, CA, US: Consulting Psychologists Press, 1962.

[104] L. Chittaro and F. Buttussi, "Learning Safety Through Public Serious Games: A Study of 'Prepare for Impact' on a Very Large, International Sample of Players," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 3, pp. 1573–1584, 2022, doi: 10.1109/TVCG.2020.3022340.

[105] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview, Theory Into Practice," vol. 4, no. November, pp. 212–218, 2002, doi: 10.1207/s15430421tip4104.

APPENDIX

Pre-Motivation Scale (7-point Likert Scale):

1. I try to learn as much as I can from following the training material.
2. I am motivated to learn the skills which are taught within the following training material.
3. I would like to improve my skills.
4. I am willing to invest effort in the following training material to improve my skills.
5. The following training material has a high priority for me.

System Usability Scale (7-point Likert Scale):

1. I think that I would like to use this VR training frequently.
2. I found this VR training unnecessarily complex.
3. I thought this VR training was easy to use.
4. I think I would need assistance to be able to use this VR training.
5. I found the various functions in this VR training were well integrated.
6. I thought there was too much inconsistency in this VR training.
7. I would imagine that most people would learn to use this VR training very quickly
8. I found this VR training very cumbersome/awkward to use.
9. I felt very confident using this VR training.
10. I needed to learn a lot of things before I could get going with this VR training.

Self-Efficacy Scale (7-point Likert Scale):

1. I believe I have the ability to accomplish the assembly task effectively (in the time given and without errors).

Satisfaction Scale (7-point Likert Scale):

1. I feel very happy and satisfied about the training method.

Post-Motivation Scale (7-point Likert Scale):

2. I enjoyed experiencing the virtual world very much.
3. I thought experiencing the virtual world was quite enjoyable
4. I would describe the experience as very interesting
5. The experience in VR training was fun.
6. I feel very happy about the training method.

NASA-TLX (0-100 for each component):

1. How much mental and perceptual activity did you spend for this task? (e.g. thinking, deciding, remembering, looking, searching, etc.)
2. How much time pressure did you feel in order to complete the task?
3. How much effort was required (mentally) to accomplish your level of performance?
4. How insecure, discouraged, irritated, stressed, and annoyed were you during the task?
5. How successful do you think you were in accomplishing the goals of the task?

Control Variables per TG

Variable	TG1		TG2		TG3	
	Mean	Sd	Mean	Sd	Mean	Sd
Age	30.97	9.16	30.74	9.43	30.53	8.68
Gender	1.29	.487	1.31	.481	1.30	.481
English Reading	90.32	13.64	89.19	16.19	91.50	13.10
English Listening	89.38	14.46	88.64	16.42	90.91	13.12
Pre-motivation	6.11	.891	6.14	.905	6.08	1.086
VR Experience	1.15	.857	1.13	.940	1.24	.982
Assembly Experience	1.56	1.186	1.59	1.364	1.57	1.335
VR Assembly Experience	.144	.352	.110	.313	.156	.364
Computer Use	5.04	1.052	4.97	.999	5.04	1.066
Computer Gaming Habits	1.77	.937	1.82	1.097	1.90	1.106
Display Size	1.80	.395	1.85	.354	1.84	.364
Computer Pointing Device	1.16	.371	1.14	.354	1.15	.364

Game Element Assessment (for selection process)

○ = non-existing ◔ = poor ◑ = medium ◕ = rather strong ● = very strong

Game Element	Effect Size	BNT Theory Fit	Learning objective Fit	Redundant / Overlapping elements	Counterworking elements
Anarchy	◑	◕	○	Exploration	Goalsetting
Anonymity	◑	○	○		Avatar, Team
Avatar	◕	◕	◕		Anonymity
Badge (also: Achievements)	●	●	●	Prize	
Challenge (also: Quest)	◕	◕	◔		

Choice	●	●	◕	Customisation	
Collection (also: Virtual Economy)	◑	◑	◔		
Competition	◑	◔	◑		
Consequence	●	◕	●		Lottery, Easter Egg
Customisation	◕	◕	◔		
Easter Egg	◑	◔	○	Lottery	
Emotion	◑	◑	○	Meaning, Narrative, Roles	
Exploration	◑	●	◔		Time Pressure
Feedback	●	●	●		
Gifting	◑	●	◔	Collection	
Goalsetting	●	◕	●		Exploration
Honour System	●	●	◕	Reward Schedule, Prize, Badges	
Leaderboard (also: Ranking)	◑	◑	◑	Points	
Learning Examples	◕	◑	◕		
Level	◕	◕	◑	Badges	
Lottery	◑	◔	○	Easter Egg	
Meaning	◕	●	◕		
Narratives	◕	●	◕		
Points	◑	◑	◑	Leaderboard	
Prize	●	●	◕	Badge	
Progress	◕	◕	●		
Reward schedule	●	●	◕	Prize	
Roles	◕	●	◕	(Narrative)	
Signposting	●	◑	◕		
Social Network	◑	●	◔		
Strategy	●	●	○		Progress
Team (also: Guild)	◕	●	◔		
Time pressure	◑	◑	◑		Exploration,
Unlockable Content	◕	◕	◔		
Voting	◔	◕	○		

Thomas Bohné received a Ph.D. degree in engineering from the University of Cambridge, Cambridge, U.K. in 2010. He is the founder and head of the Cyber–Human Lab at the University of Cambridge since 2019. The lab focuses on how digital technologies can augment human abilities. By using primarily an experimental approach, his research aims to understand and optimize the performance of hybrid human–technology systems in industrial contexts. In the last 18 months, the lab's team has worked with over 100 organizations and run experiments with over 1500 participants. He received the Institute for Manufacturing's research excellence award 2019, best paper award of the 2020 IEEE International Conference on Human–Machine Systems (with Elisa Roth, Mirco Moencks, and Luisa Pumplun), and was nominated for the best paper award at the 2020 IEEE International Conference on Industrial Engineering and Engineering Management (with Mirco Moencks and Elisa Roth).

Ina Heine received a Ph.D. degree in psychology from RWTH Aachen University, Aachen, Germany, in 2016. Since 2017 she is head of the research area organizational development at the Laboratory for Machine Tools and Production Engineering WZL at RWTH Aachen University. Her team focuses on the sociotechnical design of human-machine systems with the aim of optimizing overall system performance. Current related third-party funded projects include the development of intelligent support systems for shop floor and service employees (e.g., AuQuA—"Augmented Intelligence-based Quality Assurance of Assembly Tasks in Global Value Networks" and AIXPERIMENTATIONlab—"Augmented Intelligence for supporting employee decision-making").

Felix Mueller received an M.Sc. degree in production engineering from RWTH Aachen University, Germany and an M.Sc. in management science and engineering from Tsinghua University, Beijing in 2021. From 2020 to 2021 he was a visiting graduate student with the Institute for Manufacturing at the University of Cambridge, United Kingdom. Currently, he is part-time working towards an MBA from Collège des Ingénieurs, Paris and works as an external consultant at Infineon Technologies in Regensburg, Germany. His research interests include human-computer interaction, mixed reality and industrial training.

Vera Eger received an M.Sc. degree in business, organizational, and social psychology from the Ludwig-Maximilians-Universität (LMU) München, Germany, in 2021. From 2020 to 2021, she was a Visiting Graduate student with the Institute of Manufacturing, Cambridge, Cambridge, United Kingdom. Currently, she is working as an Innovation Consultant and Business Creator at the UnternehmerTUM, München, Germany. Her research interests include socio-technical design systems, digital anxiety, and Human-Computer-Interaction.

Paul-David Zuercher received a Bachelor of Science in computer science from the Technical University of Darmstadt, Germany in 2021. Currently, he is the technical lead of the virtual learning environment optimization research project at the University of Cambridge's Cyber-Human Lab. His research focuses on the optimization of virtual training to improve human performance in the industry.