Virtual Reality (2024) 28:165
https://doi.org/10.1007/s10055-024-01058-0

ORIGINAL ARTICLE

# Training transfer validity of virtual reality simulator assessment

**Simen Hjellvik**[1] · **Steven Mallam**[1,2]

Received: 6 March 2024 / Accepted: 7 October 2024

**Abstract**

This study utilises computer-based simulations to explore the transfer effects of competency training in maritime education, addressing the current lack of research on their transferability to real-world scenarios.

The research explores the accuracy of procedural knowledge assessment using virtual reality (VR), positing that head-mounted display (HMD) VR offers stronger concurrent validity through training transfer measures than 3D desktop VR. This is evaluated by regression on a training transfer condition. It also investigates motivation's influence on training transfer and the regression model of this relationship.

Fifteen marine engineering students were divided into two experimental groups using 3D desktop VR and HMD VR systems, with eight experts in the control group. The students had previously received traditional lecture-based instruction and were given practical training using a 2D desktop simulator in the same scenario as in the VR treatment and in the training transfer condition.

The ANCOVA design experiment involved two levels of technical immersion before the operation of real-life equipment. Neither technical immersion nor expertise level as independent variables were found to have a significant effect in the relationship of the assessment predicting the training transfer. The direct relationship was significant ($R^2_{\text{adj}} = 0.436$) and further analysed with the influence of motivation, resulting in a moderation model with a decent effect size ($R^2 = 0.740$). Based on these findings, we can infer that both types of VR simulations used for assessment demonstrate concurrent validity in predicting real-life performance before we discuss and define the characteristics of the observed transfer according to theory.

**Keywords:** Computer-Based Simulation · Experimental Design · Marine Engineering · Maritime Education and Training · Simulator Training and Assessment · Training Transfer

## 1 Introduction

### 1.1 Problem

Virtual reality (VR) applications have become widely adopted in several domains of education and training (Radianti et al. 2020), offering immersive simulator experiences. VR refers to technology that delivers an interactive virtual environment to imitate a real-world experience, through hardware that provides a level of immersion of the user experience into this environment (Suh and Prophet 2018). Learners who use an immersive head-mounted display (HMD) in an educational setting are susceptible to be more engaged, spend more time on learning tasks, and improve their cognitive, psychomotor, and affective skills (Jensen and Konradsen 2018). Computer-based simulators are tools designed to create a secure atmosphere for learning by replicating professional environments and activities. Cutting-edge computer-based simulator technology necessitates investment and operational resources to achieve anticipated learning outcomes in education and training. The return on investment for such an allocation of resources can only be determined by the value it generates. By evaluating the training demand and the training outcome in relation to the residual impact of real-world changes resulting from the training intervention, one may make a direct

---

✉ Simen Hjellvik
Simen.Hjellvik@usn.no

Steven Mallam
Steven.Mallam@mi.mun.ca

[1] University of South-Eastern Norway, Horten, Norway

[2] Memorial University of Newfoundland, St. John's, Canada

comparison. In other words, a trainee undergoes a series of activities designed to stimulate changes in specific learning constructs. These changes are intended to be used and reinforced in the trainee's subsequent professional environment. The measurement taken during the training condition can be used to evaluate the generalization of learning constructs, but its effectiveness is limited to the training condition alone. However, if an evaluation of comparable criteria is also collected in a professional setting, it will provide a measure of the generalizability between the training outcome and its immediate impact in the actual world. The real-world effect is the professional competency. There is a scarcity of empirical research across the fields employing simulator training on the assessment of competency in immersive simulator conditions, that also includes measurements of the transfer of training outcomes (Doozandeh and Hedayati 2022; Moglia et al. 2016; Schmidt et al. 2021). Despite a century of research on training transfer, the existing models and theories still require further investigation due to advancements in technology and new applications and contexts created. This study aims to evaluate professional competency using advanced immersive simulator technology, ensuring the concurrent validity of training outcomes through training transfer measures. The technologies used in this study are referred to as (1) 2D desktop VR, (2) 3D desktop VR, and (3) HMD VR, and their characteristics are discussed further in Sect. 2.

## 1.2 Training transfer definition

The transfer of knowledge and skills has become a significant aspect of research in the training sector after a century of empirical and theoretical investigation. Over time, the perspectives and frameworks for transfer have been enhanced via this research (Barnett and Ceci 2002).The taxonomy proposed by Bloom (1956) remains applicable in contemporary times as it provides a means of assessing the extent to which knowledge can be applied, analysed, synthesised, and evaluated. From a broader viewpoint, it is reasonable to state that the proposed measures can represent evaluations of the transfer of information. The models introduced by Gagné (1965), which encompass eight types of learning, nine steps of planning instruction, nine events of instruction, and evaluation of instruction, have had a significant impact on the contemporary understanding of transfer as a process of generalising knowledge. Gagné (1965, pp. 231–236) provided a definition for the process of generalising information to new contexts, which occurs through two main mechanisms: *lateral transfer* and *vertical transfer*. *Lateral transfer* refers to the process of applying acquired knowledge or abilities to a wide range of contexts that have a similar level of complexity. Vertical transfer refers to the process in which previously learned knowledge or skills serve as a basis for addressing more advanced and demanding tasks. Vertical transfer can also extend downwards since the acquisition of a higher-level skill could facilitate the mastery of lower-level skills. The continued relevance of these concepts is evident in the study conducted by Blume et al. (2010). They characterised a measure as transfer if the targeted outcome was evaluated through either (1) tasks that were different in nature or difficulty or (2) in a different environment than the one used for training.

Another perspective on the conditions for transfer focuses on the degree of similarity between situations. When situations are comparable, it is referred to as a *near transfer*; however, when situations are divergent between training and on-the-job performance, it is referred to as a *far transfer* (Barnett and Ceci 2002). Barnett and Ceci (2002) employ two aspects in their taxonomy of transfer: (1) The substance being transferred and (2) the environment in which the transfer takes place are both subject to various levels of dissection within a range of proximity or distance. In order to determine the context dimension of a transfer event based on the taxonomy, it is necessary to ascertain the distinction between distant and proximity characteristics in relation to the knowledge domain, physicality, temporality, functionality, social aspects, and modality. If the training and transfer events take place in the same room as opposed to being in various environments, for instance, that will determine how close they are to one another in the physical context. If the two events are administered at, for example, a one-week interval, the temporal context can be considered neither proximate nor distant, in contrast to if they are administered on the same day or a year apart. Further, if training were provided in the form of lectures and the transfer of knowledge were assessed in terms of professional abilities, the modality context may be deemed significant. This taxonomy aids in identifying the characteristics of a training transfer before or after its occurrence, which will be revisited in the discussion to define the transfer's characteristics in our study.

In review of these definitions, transfer refers to the process of applying knowledge or abilities to a different context from the one in which they were first learned. The resemblance between the content being transferred and the context in which it is being expressed may vary. The intricacy of applying newly acquired knowledge and skill to a setting of greater or lesser complexity is a challenge to operationalization of training transfer research. However, with this challenge also comes a value to training research as resolving it can help better application of training resources and which training resources to enhance. After examining the definitions of transfer as reported in the field of training research, we now turn attention to how the field conceptualize these definitions.

## 1.3 Training transfer concept

Baldwin and Ford (1988) presented a model in their review that includes covariate characteristics that influence the link between training outcomes and on-the-job performance. The authors consider transfer in two dimensions: (1) the application of newly acquired knowledge and skills to the job setting, and (2) the ability to maintain performance over time. These aspects are depicted in Fig. 1. Remarkably, in order to delineate the impact of transfer and its state over time, they propose employing a pre-training condition as a metric to graph the decline in on-the-job performance and plan for the revision or repetition of training. In order to evaluate the application of knowledge and skills, it is necessary to establish explicit criteria for accurate measurement. Simply put, the specific knowledge and skills required in the job setting should determine the criteria for assessing training outcomes.

The Baldwin and Ford (1988) model is highly referenced used in training transfer research (Blume et al. 2010; Grossman & Salas 2011) and examines trainee characteristics, training design, and work environment as influential factors. Deriving at a concept for a specific training or research purpose is in itself a challenge, as Grossman and Salas (2011) state:

> “Organizations that seek guidance when developing training programs and promoting the transfer of training can rely on a vast database of literature that has resulted from decades of research. Such resources, however, contain numerous, sometimes inconsistent, findings that can make it difficult for organizations to pinpoint exactly which factors are most critical for training transfer.”

### 1.3.1 Motivation in training transfer

The trainee’s motivation is a characteristic that is positively linked to training outcome and training transfer (Baldwin and Ford 1988; Baldwin et al. 2017). It is rarely simplified to a single construct, but rather a multifaceted phenomenon that encompasses a variety of factors that help improve affective skills and behaviours (Dweck and Leggett 1988). It refers to conscious and unconscious judgements about how, when, and why we put effort into a task or activity (Parks and Guay 2009). Training research views self-efficacy as an independent motivational factor, alongside intrinsic and extrinsic motivation (Colquitt et al. 2000; Grossman and Salas 2011; Makransky and Petersen 2021), while others find it useful to view motivation as an umbrella of factors that can be individually analysed (Duncan et al. 2015; Holland et al. 2018). Motivation is interesting to training researchers as it can be predicted by cognitive abilities and yet explain variance when the effects of cognitive abilities are controlled (Bell et al. 2017).

Expecting to find motivation connected to the training transfer process we adapt motivation into a conceptual model that can be empirically tested. Figure 2 depicts the links between motivation, training outcome assessment, and the moderating effect on the direct relationship. Specifically, it shows how motivation, measured prior to the assessment of training outcomes, acts as a moderator in influencing the direct relationship. The training intervention is likely to have an impact on the measurement of motivation after training, which could also represent a mediation effect in the relationship. If we examine this within the context of an educational process that emphasises motivation related to learning, we can conceptualise a mediation and moderation model similar to Fig. 2. This reflect a mediation and moderation regression in Hayes (2022) which consist of both a moderation and a mediation that can also be analysed individually. Hayes (2022) is a methodology for regression-based path analysis that is useful for complex experimental designs. For other studies of this sort with VR, see Mulders

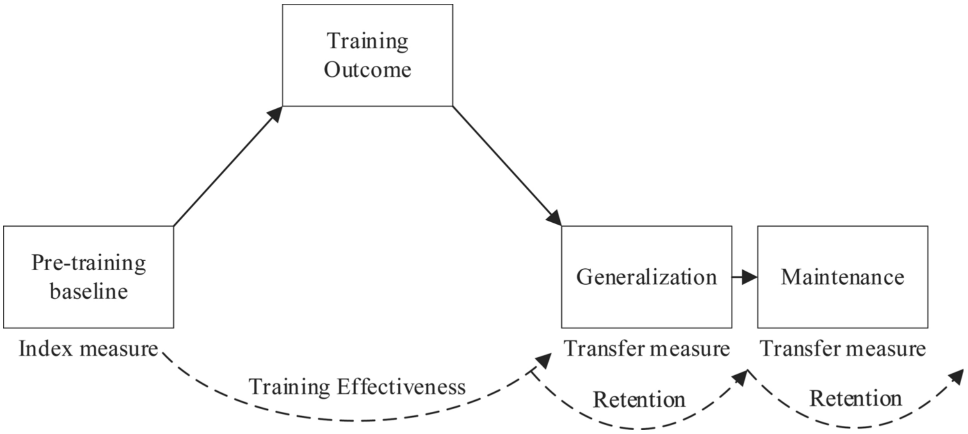

**Fig. 1** training transfer based on Baldwin and Ford (1988) and Grossman and Salas (2011)

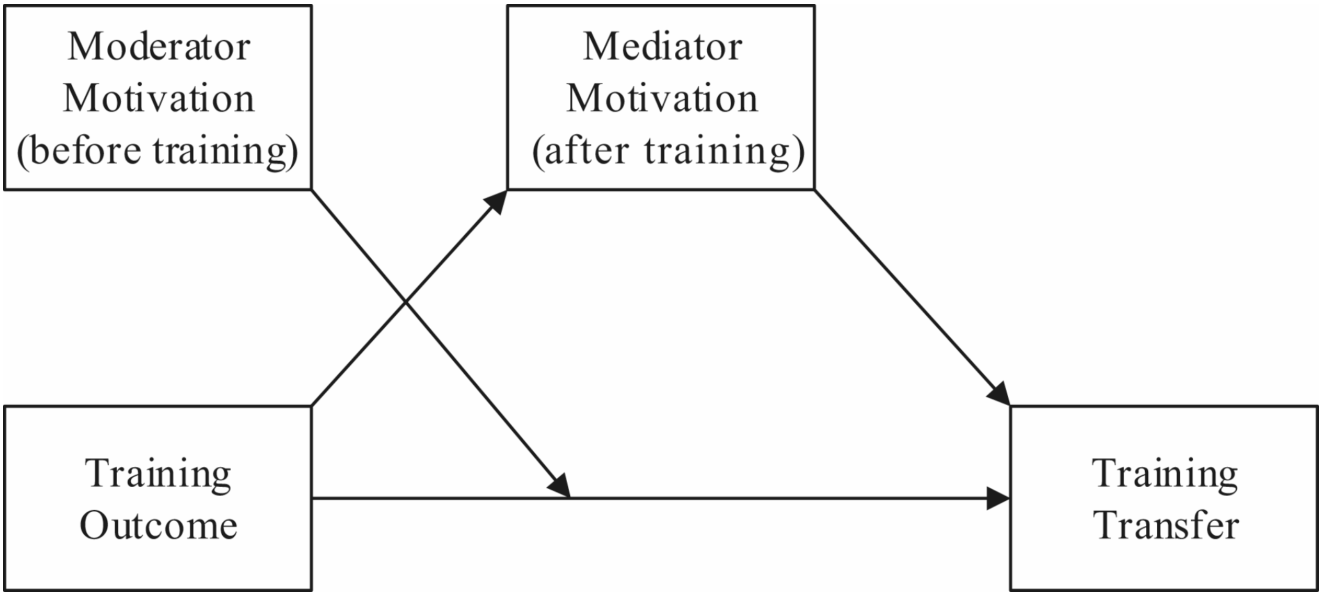

**Fig. 2** Training transfer conceptual model with motivation

(2023b). In our conceptual path model, the predictor variable would be the training outcome, which is related to a dependent measure of training transfer. Motivation, treated as a covariate, could be assessed before the start of the training, making it a possible moderator variable as it would not be affected by the training intervention itself. The training intervention itself could have an impact on a secondary measure of motivation, and if found to covary with the intervention this motivation measure could act as a mediator in this relationship.

The study conducted by Colquitt et al. (2000) integrates motivation with a path analysis model to examine its impact on transfer and job performance. Their comprehensive model reveals that the learning outcomes of skill acquisition and post-training self-efficacy are significantly correlated with transfer, although declarative knowledge and reactions to training are not significant. Notably, all four predictors of learning outcomes in transfer had a substantial effect size, with skill acquisition having a higher path coefficient compared to the motivational component of post-training self-efficacy. The endogenous variable of the model was job performance, with a large effect size from the transfer measure as the sole predictor. These results indicate that transfer should be considered a measurement condition separate from the on-the-job condition.

In view of this, the discrimination between a transfer setting and a job setting makes sense in a training situation. As a trainee is taken out of the workplace and provided training there is an opportunity to measure pre- and post-training performance in the workplace itself. However, in an educational context, the pre-training measurements may have to be implemented in a different setting than the post-training transfer measures. Hence, there is a difference between trainees and students as recipients of training; the students' professional competency after training can only be assumed without further effort to validate their training outcome.

### 1.4 Current training transfer research

Training transfer has been incorporated as a metric in previous and current models of training evaluation and training effectiveness. Training evaluation is a method used to quantify the results of learning that occur as a result of training. On the other hand, training effectiveness is the theoretical approach used to examine and comprehend the components that impact learning outcomes and transfer. Alvarez et al. (2016) developed an integrative model of trainee evaluation and training effectiveness in their meta-study. Regarding motivation, they note that prior models and research have included several elements of motivation and, in certain instances, combined scales. Specifically, the study indicated that self-efficacy prior to training was positively associated with transfer performance. Subsequently, they determine that both pre-training and post-training measures of motivation are necessary in research on training effectiveness in order to investigate the impact of motivational changes on training outcomes. This finding is supported by Ford et al. (2018), who found that motivation to learn and pre-training self-efficacy play a crucial role in facilitating the transfer of training for open skills. Open skills refer to training objectives that involve learning principles that are more challenging to acquire and require greater cognitive resources (Blume et al. 2010).

In recent years, empirical research on transfer has continued to advance the area by using proven theoretical models, including antecedents and variables in a more comprehensive and extensive manner. Ford et al. (1998) conducted an empirical study to examine the relationships that lead to training transfer. They evaluated the performance measures of 93 students who participated in a radar operation scenario. Through regression analysis of their transfer measure, the correlations between the training outcome measures of knowledge, end-training performance, and self-efficacy were found to be significant predictors. Various studies consistently emphasise the importance of motivation (Baldwin and Ford 1988; Colquitt et al. 2000; Grossman and Salas 2011), and in particular self-efficacy (Aguinis and Kraiger 2009; Alvarez et al. 2016; Blume et al. 2010; Ford et al. 2018), as powerful mediators for transfer. Seiberling and Kauffeld (2017) found that volition to transfer explained additional variance between motivation and the transfer of training. In simpler terms, individuals are not only driven by the need to acquire more knowledge, delve deeper into a subject, and excel in a certain evaluation, but also by the aspiration to amplify their influence in their everyday tasks.

In order for training transfer research to have practical use, empirical studies rely on a measurement criterion of transfer (Ford et al. 1998). Sullman et al. (2015) examined how professional drivers used their knowledge of economic bus driving in their study. To evaluate the transfer of skills, the study involved simulator training and on-the-job fuel consumption measurements. From the perspective of training effectiveness, this design utilised clear criterion measures to assess *near transfer* features. Specifically, the knowledge domain, functional aspects, social dynamics, and modality were similar across the training and transfer conditions. The sole instance of *far transfer* occurred in relation to the physical and temporal context. The performance metrics were gathered at four different time points: 1.5 months prior to training, during the simulator training, 1.5 months after training, and a retention collection 6 months after training. This design guaranteed a direct relationship between an initial measure, a training result, a transfer measure, and a maintenance measure. The second example is on the transfer of procedural skills. Taylor and Barnett (2013) conducted a study using three different types of training: desktop computer simulation, wearable VR simulation, and live training. There was a live test condition following these training conditions. The training focused on teaching military hostage rescue procedures to non-military individuals. The effectiveness of the training was evaluated based on the accurate execution of the procedural stages. Notably, this design offered three distinct variations in the proximity of the physical, functional, and modality context between the training and transfer conditions. However, there was no modification in terms of *lateral* versus *vertical transfer*, as the complexity and content remained unchanged. The study revealed superior performance in the live training condition as opposed to the wearable VR condition, which, in turn, exhibited greater performance than the desktop simulation. In the study conducted by Rauter et al. (2013) on complicated rowing skills, they also examined the transition from simulator- and live-training conditions to live-transfer conditions. Both training conditions were equipped with biomechanical sensors to evaluate performance quantitatively, in addition to expert observation of performance. Therefore, the objective and subjective instruments for measurement were identical in both the training and transfer situations.

Overall, the likelihood of successful training transfer is likely to be enhanced by the similarity between conditions. However, it is important to highlight the specific human and technical aspects that are relevant to this relationship. Immersion is a technical quality of the technology that should result in a perceived sense of presence, that is, a consciously feeling of "being there" in the virtual environment (Mayer et al. 2022). Cummings and Bailenson (2015) performed a meta-analysis on the link between technical immersion and perceived presence and find this to have a medium effect size. In simulator training G. Makransky et al. (2019a, b) anticipated that HMD VR would offer an increased sense of presence and facilitate cognitive processing, resulting in enhanced learning and transfer. The researchers discovered that students experienced a greater sense of being present in the VR science lab simulation when they used an HMD VR with high immersion. However, it was found that these students actually acquired a lower level of knowledge compared to those who used the low-immersion version of the simulation on a desktop computer. In another study Guido Makransky et al. (2019a, b) found HMD VR and 3D desktop VR to be no different in conveying knowledge than a paper text format, however they did find a positive motivational change with the immersive technologies.

In summary, research employing immersive technologies for training have shown different effect on training outcomes linked to the level of immersion. However, isolating the use of advanced VR simulations as a tool to assess training outcomes and then validating this result, seem relatively unexplored in the educational context for professional competencies. In remembering that training (acquiring) competency is not the same as assessing (exhibiting) competency, we cannot expect the current understanding of VR training to be applicable to VR assessment.

### 1.5 Research questions

Computer-based simulation has become firmly established as an instructional tool to enhance learning in education and training across various fields (Lajoie 2021). Advanced technology enables the use of immersive simulators for research and implementation in situations that target specific learning objectives. The objective of this study is to investigate the application of advanced VR simulation in maritime education. The aim is to design an experiment that validates the learning outcomes achieved through simulator training and assessment based on existing knowledge about training transfer. Training of *how to do* internalizes task procedures as procedural knowledge (Radianti et al. 2020). Training outcomes within a professional discipline typically involve more than just performing a task. To allow expertise to be the overarching goal of training towards a professional capacity, we consider professional competency to be the proper concept for a product of training that integrates different levels of learning constructs. Or, as stated by van der Vleuten and Schuwirth (2005, p. 313), *“a competency is the ability to handle a complex professional task by integrating the relevant cognitive, psychomotor, and affective skills."* For example, a nurse may have extensive training and education, but will the nurse be able to appease patients and hit the vein to draw blood every time as an intern?

The broad categories of learning constructs rarely occur as completely isolated factors in an education or training situation. It is common knowledge in training research and practice that skills can be developed from a foundation of knowledge and that motivation play a role in this process.

The primary research question pertains to the impact of VR simulator assessment on a transfer condition. Corresponding criterion measures for training performance and transfer performance will examine the effect in terms of procedural knowledge with regard to transfer between various conditions. The research question assumes the existence of a transfer effect that can be measured and that this effect varies depending on the level of immersion in the assessment environment.

*Research question 1* What is the accuracy of competency assessment of 3D and HMD VR?

In order to assess whether the transfer of procedural knowledge and skill varies based on levels of immersion, the experiment utilises two distinct experimental settings featuring VR simulation. We can view the scenario task as a specific professional competency, involving both knowledge and skill components. As the competency measured is developed through typical educational activities and desktop simulator training, we can assume that the level of that competency is possible to measure in different settings. The predictive features of the direct relationship between simulation assessment and training transfer would describe the concurrent validity of VR simulation as a tool to assess such competency. The training outcome will be evaluated using either a wearable HMD simulation or a 3D desktop VR simulation, and then equally measured in a real-life training transfer condition. Through the manipulation of technical immersion, hypothesis 1 states that as the HMD VR condition more closely approximates a real-life condition, it will provide a more accurate prediction of the criteria measure in the training transfer condition.

**Hypothesis 1** HMD VR group will have stronger concurrent validity of the assessment by training transfer than the 3D desktop VR group

The hypothesis makes three premonitions, namely, (1) that there will be a significant overall regression, (2) that the independent grouping variable gives a significant interaction, and (3) that the HMD VR group values prove a stronger regression than the 3D VR group. Moreover, the hypothesis suggests that we cannot adequately explain training transfer as a solitary occurrence. That is, by controlling suspected covariates and factors, we could improve any resulting regression model. We adopt a narrow perspective on motivation in this study, given its critical role in influencing training transfer. Therefore, we anticipate that the intervention will induce changes in motivation, possibly due to its manipulation. We expect motivation in relation to the transfer effect to lead to a more comprehensive understanding of causation.

*Research question 2* What is the influence of motivation on training transfer?

Hypothesis 2 posits that incorporating the motivational components into the study will enhance the relationship examined in the first hypothesis.

**Hypothesis 2** Motivation as a covariate will increase the effect and better explain the relationship between VR simulator assessment and training transfer

## 2 Methods

### 2.1 Procedure

The objective of this study was to investigate the assessment of professional competency using advanced immersive simulator technology. First, the participants were given the opportunity to train on a 2D model to acquire a baseline of competency which is then examined in an assessment scenario utilizing two types of VR simulation. Finally, the participants are examined again in a real-life scenario. An expert control group was included to provide validation and control for the measurements collected. With the training outcome being a specific professional competency in marine engineering, the experiment was designed to capture evidence for the concurrent validity through training transfer measures. The target population was students in marine engineering and experienced marine engineer officers. In order to avoid inconvenience of participation for the targeted student population, the study was incorporated into a mandatory marine machinery course that was directly related to the subject matter of the study. Through their seven-semester long education programme the students are trained and prepared for being responsible for operation, maintenance, and management of all technical systems onboard typical seagoing vessels. This course was given in the third semester as the second in a series of speciality courses that also includes compulsory simulator training as a component of the curriculum. The specific activities contributing to this study was undertaken as supplementary training after the students had finished their compulsory training exercises and had free time before their exams. The study consisted of three conditions, which were mainly conducted on separate days. Each participant in the study engaged for 4 to 5 h in total and involved one-on-one sessions with the first author serving as the instructor. The individuals from the expert population completed all three conditions in a single day. As a motivation for recruitment and engagement, all students who successfully completed scenarios were eligible for a chance to win a PlayStation 5 gaming console once all data had been gathered.

As shown in Fig. 3, the procedure included three conditions: training; assessment; transfer. First, all participants, including the expert population, conducted an initial knowledge test and 5 repetitions of the scenario on a 2D desktop simulator. By standardizing a 5-repetition requirement, it was assumed that the participants would reach a learning curve saturation. After each attempt, and automated performance score of the scenario would appear. Then, the students were assigned to one of two experimental groups (3D VR or HMD VR) and the experts to a control group (HMD VR). The assessment condition containing the experimental manipulation was only performed once. Finally, the participants performed a practical knowledge test and the same scenario in real life at a machinery laboratory. The participants were not made aware of their scores from the assessment condition nor the transfer condition in order to not manipulate any factors of the MSLQ. However, the practical knowledge test was administered as a familiarisation exercise and thus all incorrect items were corrected by the instructor before the scenario was performed.

A notification form (file number 683679) for ethical conduct in the treatment of personal data was granted from the Norwegian Agency for Shared Services in Education and Research, ensuring data management according to current academic and national standards.

## 2.2 Sample

The sample ($N = 23$) consisted of experts who were directly assigned to the control group and students who were assigned to either experimental group based on a matched-pairs assignment. The control group ($n = 8$) demonstrated seniority through their higher average age ($M = 47.8$, $SD = 8.2$) and longer duration of sea service ($M = 15.0$, $SD = 7.0$). Two experts were actively serving, while the rest were involved in educating marine engineering students. All experts had current and valid marine engineer officer class 1 certificates of competency, which allows them to serve as unlimited chief engineers. However, one expert held a class 2 certificate of competency. The experts reported a high level of familiarity ($M = 4.6$, $SD = 0.5$) with the machinery that was to be used in the experiment, as assessed on a 5-point scale ranging from no familiarity to extensive familiarity. The final population of novices ($n = 15$) in the sample was recruited from a third-semester cohort of marine engineering students, with only one participant withdrawing from the study. The average age of the student pool was 23.2 years, with a standard deviation of 2.0. Additionally, the average sea service experience from commercial or military service was 1.1 years, with a standard deviation of 1.5. Through such service none of the students

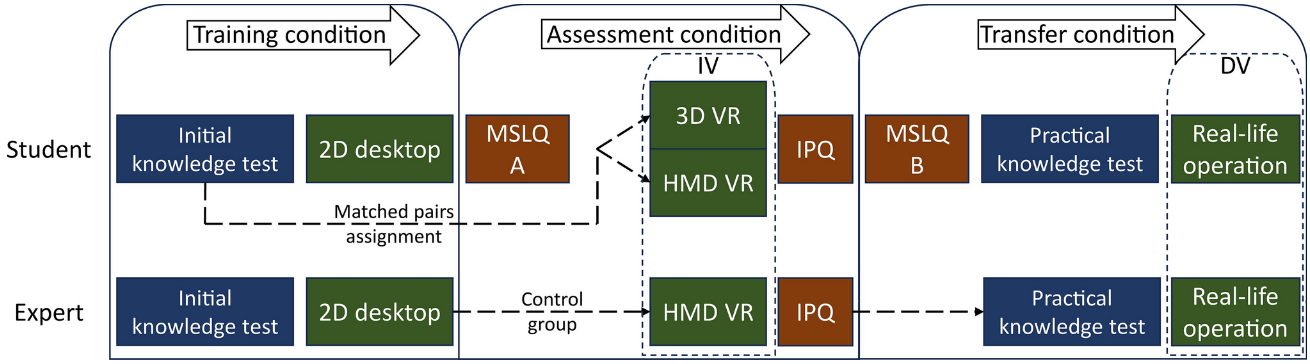

**Fig. 3** Research procedure chronologically from left to right

**Table 1** Condition characteristics.

| Condition | Setting | Scenario | Interaction | System complexity | Environment realism | Data collection |
|---|---|---|---|---|---|---|
| Training | 2D simulation | Fuel oil separator | Desktop | High | Low | Automated |
| Assessment (Fig. 4) | VR simulation | Fuel oil separator | Desktop or wearable | Moderate | High | Automated and observation |
| Transfer (Fig. 5) | Real-life | Lubrication oil separator | Physical | Moderate | Absolute | Observation |

had prior hands-on experience with the specific machinery model of this study, although a few had observed or helped in operation of a comparable model. The participants were assigned to either the 3D VR group or the HMD VR group based on their performance in the initial 20-item knowledge test. Coincidentally, each group was allocated one of the two female students from the sample. The matched-pairs assignment was deemed successful as there were no significant differences observed between the two experimental groups in terms of age ($t_{(1,13)} = 0.669$, $p = 0.515$), sea service ($t_{(1,13)} = 0.758$, $p = 0.462$), gaming console experience ($t_{(1,13)} = 0.650$, $p = 0.527$), HMD VR experience ($t_{(1,13)} = 0.854$, $p = 0.408$), and reported familiarity with the machinery equipment of the experiment ($t_{(1,13)} = 1.491$, $p = 0.160$). The experts were very different from the students when it came to age ($t_{(1,21)} = 8.385$, $p < 0.001$), time spent at sea ($t_{(1,21)} = 5.587$, $p < 0.001$), experience with video games ($t_{(1,21} - 4.965$, $p < 0.001$), how well they knew the experiment's machinery ($t_{(1,21)} = 6.000$, $p < 0.001$), and how they did on the knowledge test ($t_{(1,21)} = 3.025$, $p = 0.003$). There was no significant difference in experience with HMD VR between the professionals and the students ($t_{(1,21)} = -0.364$, $p = 0.360$). This indicates that the control group differs from the experimental groups in terms of demographic data that could be used to determine expertise in this regard.

To verify that five repetitions in the training condition were enough to reach a saturation point of learning, repeated measures ANOVAs were run with post hoc tests to compare all attempts with each participant's final attempt. We used this approach as it would be equivalent to running related sample *t*-tests between each attempt and the fifth for both demographic groups. In terms of the automated score from each attempt, neither the student population nor the expert population had any attempts with statistical significance compared to their last attempt ($F_{(1.976,41.966)} = 1.308$, $p = 0.274$). In terms of time consumption, however, there was a significant difference ($F_{(1,884,39.556)} = 26.066$, $p < 0.001$). The post-hoc show a difference between the last and first attempts of the experts (-657(95%*CI*, -1138 to -175) seconds, $p = 0.010$), and for the student between the last and both the first and the second (-241(95%*CI*, -403 to -80) seconds, $p = 0.005$). Thus, we could assume that the experts reached their learning plateau at their second attempt, and the students reached theirs at their third.

**Fig. 4** VR environment of the assessment condition accessible through either desktop or HMD

## 2.3 Condition and scenario

To ensure a fundamental system comprehension the student sample was engaged only after they had covered the relevant learning outcomes through lectures in their undergraduate programme in marine engineering. The experiment was designed as a series of different conditions on the same scenario. As shown in Table 1, there are different characteristics for the three conditions. Regarding the simulator software, the model used for the training condition was operated by interaction on a 2D line-diagram interface, while the model used for the assessment condition was an interactive 3D model (see Fig. 4). The task of each scenario was to perform a standard operating procedure of bringing a separator machinery from a shut-down to an operating state. The three conditions had some variation in system design of the separator machinery: however, the core features and functions were present in all. A separator is a machinery system that is used for purification of fuel oil and lubrication oil, see Fig. 5. The machinery itself operates on the principle of extracting water and solid contaminations from the oil by applying centrifugal force at 9500 rotations per minute to create a phase separation between different densities in oil and contaminants. The system as a whole involves tank arrangements, pipes, valves, instruments, electric power and automation panels.

**Fig. 5** Transfer condition with the MMPX oil separator in the machinery laboratory

**Table 2** Material characteristics.

| Condition | Environment | Technology | Model | Machinery |
|---|---|---|---|---|
| Training | Desktop simulator laboratory; Multiple student stations in a classroom layout | 2D desktop VR simulation | K-Sim Engine L11; Suezmax oil tanker model | Alfa Laval alcap type separator system with EPC50 control panel |
| Assessment (Fig. 4) | VR laboratory; one student station and one instructor station | 3D desktop VR simulation<br>HMD VR simulation | K-Sim Engine M11; Container feeder model | Alfa Laval alcap type separator system with EPC50 control panel |
| Transfer (Fig. 5) | Machinery laboratory, one student and one instructor | Real-life operation | | Alfa Laval MMPX purifier-type separator with EPC41 control panel |

## 2.4 Materials

All three conditions were prepared at the facilities of the authors' institution. Main differences are summed in Table 2. The training was conducted in a dedicated setting, specifically designed for this purpose. It included 15 student stations, each equipped with two desktop monitors, as well as keyboards and mice for user interaction. The assessment condition was set up in a capacious simulator laboratory specifically dedicated for research purposes. A high-performance laptop was necessary to operate the software for dual forms of interaction. The laptop utilised was an MSI GE66 equipped with an 11th Generation Intel® Core i7 CPU and NVIDIA GeForce RTX 3080 GPU, operating on the Windows 10 platform. A 86-inch 4K UHD i3TOUCH X-ONE interactive flat panel display was utilised, along with XBOX hand controllers, for 3D desktop VR engagement. The HMD VR interaction utilised HTC hand controllers and movement tracking, which were combined with a Varjo Aero headset. This headset offered a resolution of 2880 × 2720 pixels per eye and a refresh rate of 90 Hz. An area of 4 by 4 m was adjusted to allow for unrestricted mobility. The transfer condition was assigned to a machinery laboratory where students in marine engineering undergo practical training in their 7th semester of their degree to develop their professional skills. This laboratory houses machinery and equipment that replicate the engine room and engine control room features that are often found on an actual seagoing vessel. The equipment used for this occasion was an Alfa Laval MMPX oil separator, which belongs to the purifier category. Specifically for this study, we restored the machinery to a functional state after an 8-year period of inactivity.

## 2.5 Instrumentation

A scoring system was made for each condition based on the operating procedure of the specific separator machinery in each scenario. As the machinery systems differed in minor design details, such as pipeline layout and control panel model, the same approach was used to ensure that all procedures and scoring systems were as equivalent as possible. First, a hierarchical decomposition was made of the scenario goal, i.e., bringing the machinery to an operating state. Second, we designed an operating procedure based on the identified system interactions and their sequencing. Third, we programmed an automated data collection system for the two simulator models and prepared a scorecard for observation recording in the real-life scenario. See our previous work for details on simulator scenario programming (Hjellvik and Mallam 2023).The initial knowledge test given before the training condition was a 20-item multiple-choice test probing knowledge on system components and operation. The test items were developed based on the course curriculum and pre-piloted with a professional marine engineer officer to indicate that the experts would not all achieve a full score. The final knowledge test allocated to the transfer condition was a practical knowledge test in the machinery laboratory where the participant was tasked to assign 26 name tags to their respective system components. To design variables that represent procedural competency, knowledge and skill metrics were merged. For the assessment condition, the 20-item knowledge test was pooled with the achievement of the 29 possible correct operations of the scenario procedure. For the training transfer condition, the 26 possible correct items of the practical knowledge test were pooled with the 30 possible correct operations of the scenario procedure. Both of these quantifications were rescaled to a total range of 100 as variables to be analysed, as seen in Table 3 below. The programming and instrumentation were conducted by the first author, formerly a marine engineer officer. All instruments were piloted according to procedure with a researcher, who was also formerly a marine engineer officer. After the pilot the instruments for data collection were deemed ready and minor adjustments were made to the procedure to ensure efficiency of collection and safety of participants.

**Table 3** Descriptive statistics.

| Variables | Category | Sample | Central tendency | Application |
|---|---|---|---|---|
| Time delay between training and assessment | Continuous | HMD VR ($n=7$)<br>3D VR ($n=8$) | $M=2.29$ ($SD=2.43$)<br>$M=5.75$ ($SD=2.60$) | Descriptive |
| Time delay between assessment and transfer | Continuous | HMD VR ($n=7$)<br>3D VR ($n=8$) | $M=3.86$ ($SD=4.45$)<br>$M=1.13$ ($SD=1.80$) | Descriptive |
| Assessment Score | Continuous | $N=23$<br>HMD VR ($n=7$)<br>3D VR ($n=8$)<br>Control ($n=8$) | $M=74.05$ ($SD=12.50$)<br>$M=78.06$ ($SD=10.27$)<br>$M=86.22$ ($SD=5.42$) | Covariate/ predictor |
| Training Transfer Score | Continuous | $N=23$<br>HMD VR ($n=7$)<br>3D VR ($n=8$)<br>Control ($n=8$) | $M=71.93$ ($SD=16.02$)<br>$M=73.88$ ($SD=17.15$)<br>$M=91.71$ ($SD=4.26$) | Dependent variable |
| Immersion Level | Dichotomous | HMD VR ($n=15$)<br>3D VR ($n=8$) | | Independent variable |
| Expertise Level | Dichotomous | Student ($n=7$)<br>Expert ($n=8$) | | Control variable |
| IPQ | Continuous | $N=23$ ($\alpha=.851$)<br>HMD VR ($n=7$)<br>3D VR ($n=8$)<br>Control ($n=8$) | $M=4.91$ ($SD=0.61$)<br>$M=3.88$ ($SD=0.53$)<br>$M=5.01$ ($SD=0.72$) | Manipulation check |
| MSLQ A | Continuous | $N=15$ ($\alpha=.905$)<br>HMD VR ($n=7$)<br>3D VR ($n=8$) | $M=4.73$ ($SD=0.66$)<br>$M=4.91$ ($SD=0.61$)<br>$M=4.55$ ($SD=0.60$) | Covariate |
| MSLQ B | Continuous | $N=15$ ($\alpha=.935$)<br>HMD VR ($n=7$)<br>3D VR ($n=8$) | $M=4.81$ ($SD=0.72$)<br>$M=5.01$ ($SD=0.77$)<br>$M=4.64$ ($SD=0.68$) | Covariate |

## 2.6 Quality of measurements

Through the programming of scenarios for the two simulator models, the automated scoring systems were developed. In the Training condition, the score after each attempt was visible to the student. For the assessment condition, the automated scoring system was used as a supporting tool for the instructor, who also registered observations during the scenario. In the transfer condition, all metrics were recorded on paper by the instructor. The first author enacted the role of instructor for all conditions. We administered the initial knowledge test as an online survey under supervision. The I-group Presence Questionnaire (IPQ) and the motivation questionnaires were also given as online surveys under supervision. This allowed the instructor to clarify any uncertainties with questionnaire items and promote truthful answers.

### 2.6.1 Verification of manipulation

To determine if the experimental manipulation of technical immersion had a significant impact that was not influenced by random variations, a measurement of perceived presence was obtained immediately following the assessment condition. The IPQ is a 14-item scale ($\alpha=0.87$) specifically designed to measure the level of psychological presence experienced by individuals in virtual environments (Schubert et al. 2001). The 14 items were linguistically converted into Norwegian and thereafter presented alongside the original English text, utilising a 7-point rating. For the sake of analysis, the scale average was calculated after inverting the reverse items.

### 2.6.2 Covariate variables

Given that the study took place in an educational environment, the Motivated Strategies for Learning Questionnaire (Pintrich et al. 1991) was selected to provide a measurement of motivation. The MSLQ was used in a translated version (Kvinge and Engelsen 2016). The 31-item motivation part of the questionnaire consist of 6 subscales of motivation factors. This was administered two times to the student population. The first collection (MSLQ A) was given under supervision directly before the assessment condition and the second collection (MSLQ B) was given under supervision directly before the training transfer condition. See results section for details.

### 2.7 Data analysis

The data was analysed with SPSS 29. First, descriptive statistics were extracted before a manipulation check was run on the independent variable to test for a successful experimental manipulation. A control for demographic influence on the manipulation was also included to support the validity of the experiment design. Second, for hypothesis 1, an ANCOVA was run on the student's performance of the Training Transfer Score (outcome) with Immersion Level as the independent grouping variable (factor) and performance of the Assessment Score as the covariate (predictor). Then, an ANCOVA was run with the student HMD VR group and the expert control group to provide a control for the impact of Expertise Level (factor) on the regression. Third, for hypothesis 2, linear regressions were run to test for the interaction of motivation. This was explored further using the PROCESS macro (Hayes 2022) for SPSS.

## 3 Results

### 3.1 Descriptives and quality control

#### 3.1.1 Verification of manipulation

With the given sample size of 23, the IPQ scale demonstrated a high degree of internal consistency with a coefficient alpha of 0.851. As anticipated, the experimental group using the HMD VR reported a greater sense of presence ($n = 7$, $M = 4.9$) compared to the group using 3D desktop VR ($n = 8$, $M = 3.9$), and this difference was statistically significant ($t_{(1,13)} = 3.480$, $p = 0.002$). Furthermore, there was no significant difference in perception between the control group ($n = 8$, $M = 5.017$) and the student HMD VR group, which both got the identical experimental treatment ($t_{(1,13)} = 0.285$, $p = 0.780$). This confirms that the manipulation of immersion was successfully perceived differently and was not subject to confounding by demographic characteristics.

#### 3.1.2 Control of scoring system

As shown in Table 3, the expert population performed better than the student populations in both the assessment condition and the training transfer condition. The control group was given the HMD VR interaction of the experiment and performed significantly better than the student HMD VR group in both the assessment condition ($t_{(1,13)} = 2.506$, $p = 0.013$) and in the transfer condition ($t_{(1,13)} = 4.172$, $p = 0.008$). This verifies that the instruments made for scoring the scenarios and the use of these metrics as quantifications for the following analysis, was able to discriminate between expertise levels while all other experimental factors are equal. For the student population under investigation, there was correlation ($r = 0.690$) between the assessment condition and the transfer condition, while the mean difference was insignificant ($t_{(14)} = 1.070$, $p = 0.151$).

### 3.2 Statistical analysis

#### 3.2.1 Hypothesis 1: the HMD VR group will have stronger concurrent validity of the assessment by training transfer than the 3D desktop VR group

To test hypothesis 1 an ANCOVA was run on the student's performance of the Training Transfer Score with Immersion Level as the independent grouping variable and with performance of the Assessment Score as the covariate. Levene's test proved insignificant ($F_{(1,12)} = 0.199$, $p = 0.663$), supporting equal distribution of error terms. The White test also proved insignificant ($\chi^2 = 0.908$, $p = 0.923$), supporting heteroskedasticity of error terms. The ANCOVA model was significant ($F_{(1,12)} = 5.553$, $p = 0.020$, $\eta_p^2 = 0.481$), however, the main effect of Immersion Level ($F_{(1,12)} = 0.103$, $p = 0.735$, $\eta_p^2 = 0.009$) had no impact on the Training Transfer Score. Repeating with a simple regression ($F_{(14)} = 11.818$, $p = 0.004$) with only the Assessment Score predicting the Training Transfer Score increased the effect size from $R^2_{\text{adj}} = 0.394$ to $R^2_{\text{adj}} = 0.436$, suggesting that the assessment condition alone hold a strong prediction for the transfer condition. The hypothesis was rejected as there was no significant difference on the independent variable to be found. However, the direct effect of the research question was found with significant effect, and this is the largest empirical contribution of this study.

Further, to provide a control for the impact of Expertise Level, an ANCOVA was run with the student HMD VR group and the control group. With Expertise Level as the independent variable a significant model was produced ($F_{(1,12)} = 13.520$, $p < 0.001$, $\eta_p^2 = 0.693$). Interestingly, the main effect of Expertise Level on the Training Transfer Score was insignificant ($F_{(1,12)} = 3.293$, $p = 0.095$, $\eta_p^2 = 0.215$). This infers that the HMD VR assessment condition can be used to predict the transfer condition, regardless of Expertise Level. To be clear though, we cannot suggest collapsing the demographic population factor and analyse the HMD VR part of the sample as a whole due to between-group differences proven in the previous section.

### 3.2.2 Hypothesis 2: incorporating motivation as a covariate will increase the effect and better explain the relationship between VR simulator assessment and training transfer

Since there was no discernible change in the dependent variable as a result of the treatment, Immersion Level was eliminated for additional examination by collapsing the grouping variable for the student cases. Two linear regressions were conducted on the pooled student Training Transfer Score to investigate motivation. The predictor used in the regressions was the Assessment Score. MSLQ A was used as a control variable in the first regression, while MSLQ B was used in the second regression. Both models were determined to be statistically significant. The first model ($F_{(2,12)}=5.893$, $p=0.016$, $R^2_{\text{adj}}=0.411$) exhibited just a slight difference compared to the second model ($F_{(2,12)}=5.687$, $p=0.018$, $R^2_{\text{adj}}=0.401$). Motivation was measured before both the assessment scenario and the training transfer scenario; however, it remained constant during the experiment treatment ($t_{(14)}=1.167$, $p=0.263$). Therefore, it is important to note that the difference between MSLQ A and MSLQ B do not function as a mediator, that is, MSLQ B did not covary with the treatment. However, their existence can either improve or reduce the relationship as moderator variables. To test this, analyses were conducted using a simple moderation model (Model 1, Hayes 2022), in which both motivation measures were in turn included as moderators. Only the model in which MSLQ B was used as the moderator showed a significant interaction ($F_{(1,11)}=10.768$, $p=0.007$, $R^2_{\text{chng}}=0.254$), indicating the presence of a moderator effect. This analysis showed that the relationship between the assessment and the training transfer, controlled for motivation prior to the training transfer scenario, was statistically significant ($F_{(3,11)}=10.768$, $p=0.001$, $R^2=0.740$), see Fig. 6:A. Further, when individuals have a low level of motivation (1 standard deviation below the average), the assessment scenario has a significant positive effect on training transfer ($B=1.604$, $t_{(11)}=4.960$, $p<0.001$). At the higher level, this conditional effect of the predictor dampens as the moderator's levels increase by one standard deviation ($B=0.048$, $t_{(11)}=0.132$, $p=0.897$). In further detail, the model's output identifies a Johnson-Neyman point at a motivation level of 5.040 ($B_{\text{JN}}=0.589$, $t_{(11)=}2.201$, $p=0.050$), at which the moderator effect no longer remains statistically significant, see Fig. 6:B. Below this threshold of motivation, there is a significant degree of moderation, which accounts for 66% of the cases.

## 4 Discussion

### 4.1 What is the accuracy of competency assessment of 3D and HMD VR?

The statistical analysis yielded no evidence in favour of hypothesis 1, contradicting our initial assumption on immersion. Our hypothesis was that the group using HMD VR would demonstrate higher concurrent validity of the assessment through training transfer compared to the group using 3D desktop VR. To be true, the hypothesis required three elements: (1) a strong overall regression; (2) a significant interaction from the grouping variable; and (3) better results from the HMD VR group compared to the 3D VR group. The ANCOVA failed to demonstrate an interaction between the independent variable and the predictor, hence contradicting the second and third aspects of the hypothesis. Nevertheless, our research yielded a noteworthy prediction when the grouping variable was merged. The overall impact of the assessment scenario on the transfer scenario showed a significant result, indicating a substantial effect size.

Our approach deviates from typical training-focused studies, as discussed in Sect. 1, by incorporating immersive technologies into a singular VR experience in addition to training and educational activities. The differentiation

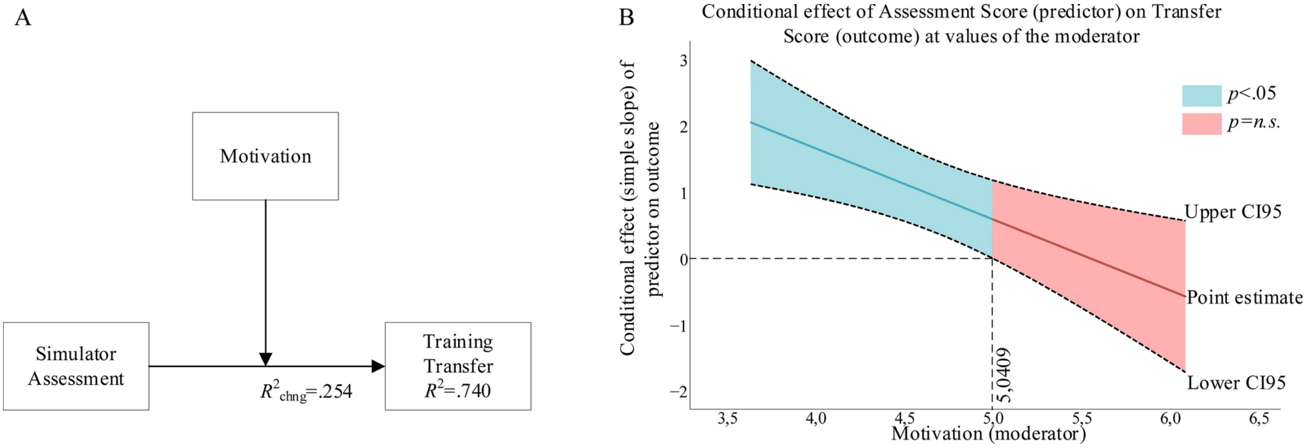

**Fig. 6** **A** Regression model. **B** Johnson–Neyman plot of conditional effects

between simulator training and simulator assessment, as this study has shown, was essential to isolate the specific effect of the research question. The primary distinction between a training-oriented approach and an assessment-oriented approach in experiments lies in the utilisation of a VR simulation to manipulate the magnitude of learning and performance effects. In our experiment, the VR simulation measures were employed to predict the conduciveness of the transfer condition. We expected that the results of the previous studies on training with different VR simulators would be applicable to our case: assessment with similar tools. We found some similarities. For example, the Taylor and Barnett (2013) study closely parallels our study in terms of content, as both studies focused on applying procedural knowledge in training to demonstrate procedural proficiency. Although their design was able to generate varying learning outcomes by manipulating the level of immersion in a series of training scenarios. The interaction with immersive technologies in our study occurred over a single instance, and no difference was found. Similar to Makransky et al. (2019a, b), we demonstrated that perceptions of presence varied based on different levels of immersion. However, their 2 × 2 study showed that declarative learning effects were better with desktop VR compared to HMD VR. Guido Makransky et al. (2019a, b) showed no change in declarative knowledge across text format training, desktop VR, and HMD VR after a short time of exposure.

The main contribution of this study is the empirical evidence that VR simulators hold concurrent validity in real-life terms in the magnitude of $R^2_{\text{adj}} = 0.436$, as performed through our experimental design. We found that 3D desktop VR versus HMD VR did not make any statistical difference in a single interaction. We also discovered that HMD VR has the potential to forecast the real-life performance of experts. That is to say, the experiment was able to predict training transfer regardless of expertise level, which further supports the concurrent validity of the assessment in predicting transfer.

**Table 4** Content characteristics discussed in Barnett and Ceci (2002).

| Content | Characteristics | | |
|---|---|---|---|
| Learned skill | Procedure | Representation | *Principle or heuristic* |
| Performance change | Speed | *Accuracy* | Approach |
| Memory demands | Execute only | Recognize and execute | *Recall, recognize and execute* |

Our identification is italicized

## 4.2 What is the influence of motivation on training transfer?

Hypothesis 2 posited that adding motivation as a covariate would make this effect stronger and give a better explanation for the link between VR simulator assessment and training transfer. The difference between the two measurements of motivation was deemed inconsequential, hence substantiating the inference that the experimental treatment did not induce a covariation in motivation. Identifying this aspect of a control variable is crucial, as it supports the use of moderation analyses. The simple moderation analysis revealed a statistically significant interaction effect for the MSLQ B variable, but no significant interaction effect was observed for the first measurement. The significant moderation index allows additional exploration of the analysis. The regression model (Fig. 6A), with MSLQ B as the moderator variable, demonstrated that motivation has a favourable impact on the relationship between the assessment scenario and the transfer scenario. However, this effect diminishes as the degree of motivation increases beyond the top third of the cases. To visualise this, Fig. 6B shows the effect (point estimate) of X on Y at any level of the moderator. The confidence interval bands show where a type 1 error becomes more likely than conventionally acceptable by illustrating the areas where both bands are either positive or negative. These findings confirm the existing knowledge of training transfer that motivation is a key factor in the conductivity of competency between conditions.

## 4.3 Characteristics of the effect

This experiment has shown the occurrence of transfer, first as a direct relationship and then as a relationship modelled with motivation. To clarify the characteristics of this occurrence, we revisit the taxonomy of Barnett and Ceci (2002) before making a generalisation argument. We recommend viewing this study through the taxonomy, which defines transfer content and context. That is, the occurrence of transfer is determined by what is transferred and then, when, and where this is transferred to and from. Table 4 delineates the specific characteristics of the transferred skills under the transfer condition. During the initial 2D desktop training, the experiment primarily focused on acquiring knowledge about a procedure while performing it. However, in the real-life transfer condition, participants were required to apply heuristic principles in order to execute a similar procedure. Specifically, none of the students had previously operated this particular machinery model in real-life scenarios, but a handful had seen or assisted in the operation of a similar model. The application of heuristic principles was observed in the transfer condition, particularly among students with commercial seagoing expertise. These students relied on their aural and tactile senses to obtain feedback from the machinery system while completing the scenario. As the performance indicators utilised were measured in terms of the proficient execution of procedures and retention of knowledge items, the dependent variable indicated accuracy in this regard. However, during the 2D desktop training, the duration of each repetition decreased, as anticipated in a repeated training scenario. On memory demands, all participants were well informed of the conditions they would be exposed to, so they could expect to confront the task of actualizing what the respective procedure specified. Despite the familiarity of the approach and strategy for solving the scenario in the first two conditions, participants in the transfer condition were required to utilise their declarative and procedural knowledge to identify unfamiliar components and functions that they had not encountered in real-life situations and to navigate themselves within this environment.

The transition between experiment conditions incorporated vertical transfer by enhancing the realism of the machinery system, as shown in Table 1, thereby making the scenarios more challenging. During the transfer phase, the participants encountered aspects that were not manipulated in the digital simulation, such as aural and tactile senses. Specifically, the presence of auditory and vibrational feedback was unprecedented with the operation of machinery in real life. The difficulty in solving scenarios was thus accretive to realism, as errors could potentially lead to risks and harm to both machinery and people. Although under the supervision of the instructor, the risk of actually making a critical error was mitigated. However, the instructor, also being a confederate agent in this regard, ensured to remind all students of the possible break-down events and consequences of mistakes.

The contextual factors relevant to our transfer scenario are categorised as both *far* and *near*, as shown in Table 5. This segment of the taxonomy aims to categorise features that exist on a continuous spectrum by imposing a subjective binary division. Based on the taxonomy, this results in the creation of 64 distinct profiles, making it difficult to provide support for generalisations based purely on this. Interestingly, Barnett and Ceci (2002) did not find any study that could be considered *far* in all respects. Nevertheless, Table 5 serves as a helpful instrument for reviewing our case of transfer and strategizing future research. The contextual factors found as *far transfer* are the ones of interest.

**Table 5** Dichotomous categorization according the taxonomy of Barnett and Ceci (2002).

| Context | Far versus near transfer | Continuum characteristics |
|---|---|---|
| Knowledge domain | Near | Procedural knowledge versus procedural skill |
| Physical context | Far | VR environment versus real life environment |
| Temporal context | Near | 0–11 days between conditions ($M = 2.40$, $SD = 3.48$) |
| Functional context | Far | Academic error-safe simulation versus real-life operation |
| Social context | Near | Individual training versus individual performance |
| Modality | Far | Digital interaction versus hands-on |

The transfer that takes place in the initial three contexts is of importance in a typical educational setting (Barnett and Ceci 2002). While acknowledging that the knowledge domain was initially acquired before being applied to a higher learning construct, the transfer of knowledge is considered *near*. In order for knowledge domain transfer to be considered *far*, it must extend beyond the boundaries of its original domain. For example, it is possible to accomplish this by having students who have only studied physics apply their knowledge to problems in marine engineering while solely relying on the learned principles. The transfer across the physical context we perceived as *far*. All three conditions were situated within the same university, yet, in terms of the micro-aspect, there were major differences between the physical rooms and the environmental conditions as sown in Table 2. To manipulate the temporal variable in a research project, a longitudinal study must be conducted. This factor has a crucial role in the maintenance of generalised knowledge (Baldwin and Ford 1988; Grossman and Salas 2011) and can be determined by repeating the transfer condition several months or a year after. The shift from an error-safe simulation to an error-exposed live operation requires the use of specific skills and a specific mindset, which represents a *far* functional transfer. Notably, the functional context also differed between the training and assessment conditions, as the stakes were higher with a single-attempt assessment compared to iterated training attempts. The social context remained very consistent between the assessment and the transfer conditions. Lastly, the modality was different as competency was transferred from a digital environment to a real-life laboratory.

Barnett and Ceci (2002) discuss the three contexts of transfer and remark that studies have demonstrated that the physical context in which a job is carried out can impact transfer, leading to variations in performance depending on the surroundings. The functional context, which pertains to the purpose or use of acquired skills, has attracted less research attention but is thought to impact transfer performance. The format in which learning is assessed, referred to as modality, has been proven to affect transfer according to their review. Different test formats may indeed result in varied levels of success to detect conductivity of knowledge and skill between conditions.

## 4.4 Implications

### 4.4.1 Limitation, causation and generalizability

Some fundamental criticism of media comparison studies stems from uncontrolled effects in the experimental design (Mulders 2023a). Although this study does not compare different types of VR for the learning process, we did compare them within the context of VR as an assessment tool. By designing a complex experimental design with standardised conditions in laboratory settings, we provided reasonable control over confounding variables.

There is an inherent risk of statistical limitation when designing experimental design research in discipline-specific fields. In our case, due to the current sample size of the expert group we chose not to report their isolated effect size and keep their purpose to the research design as validation for the student groups. In balance between quality and quantity, we spent our efforts on consolidating the former, as is evident in the control variables applied to the two populations of the sample. However, this often results in having to follow up emerging questions though new research efforts. For example, with the current sample we cannot state anything concrete about the 5 overly motivated cases, leaving an interesting explanation unaddressed. Nevertheless, quality over quantity is not necessarily a compromise, and a limitation might be translated into an opportunity.

When determining the impact of one variable on another, we usually seek (1) covariance, (2) chronological order, and (3) the exclusion of alternative explanations (Hayes 2022, pp. 15–19). When assessing the hypotheses, we discovered the mathematical support and limitation for statistical inference. This alone is insufficient to interpret the findings. The study's research procedure involved the design of a temporal order of events to meet the second assumption. For the third assumption, we tested and confirmed demographic parameters, group assignment, experiment manipulation, the instruments, covariate variables, and confounding variables. Within the frame of this research design, we can ascertain that the observed effects are indeed accurate.

This study focused on marine engineering students and professionals as the target population and may have uncontrolled characteristics that differentiate them from the general population. As such, the results of this study may have limited generalizability to similar populations based on our study alone. Further, given the limited sample size, it may even be difficult to apply our findings to the entire target group, that is, all marine engineering students or experts, anywhere at any time. However, given our trust in the research procedure, the significant effect sizes exceed the traditional acceptable limit for observed power. Therefore, it is probable that the results can be replicated based on the design, population, procedure, and phenomenon characteristics outlined in this study.

### 4.4.2 Future research

There are several directions that can be pursued with further research that could benefit both the field of training research and the practitioners of simulator training and assessment. And in the maritime domain particularly, as was our showcase. Considering the content and context of transfer, a broader range of training outcomes might be addressed to help define where training transfer is most useful. In terms of context, the characteristics the phenomenon might be relevant to isolate and explore individually. In our case, it will be useful and economic to repeat the transfer scenario with the same participants twelve months later. This would isolate the temporal context and expand this study including the maintenance or retention of transfer measurements. New efforts could be planned to focus deeper on the influence of motivation. With a larger sample, the subsuming factors of motivation could be explored individually and disclose how and which factors that can be manipulated in enhance learning effects in specific situations. As we incidentally found our control variable insignificant, expertise level could be interesting to expand from a dichotomous to a continuous variable. This would be relevant for the future application of state-of-the-art immersive technologies as competency assessment tools for both students and experienced professionals. By focusing on professionals, the training transfer model expands to another segment of the target population where real-life baseline measures are obtainable, and purpose of training is more specific.

## 5 Conclusion

The study aimed to create an experimental design to confirm the accuracy of learning outcomes from simulator training in relation to training transfer. We utilised cutting-edge commercial technology, including both software and hardware, in the field of maritime education. The dependent variable measured professional competency in a training transfer scenario, whereas the independent variable assessed the level of immersion in a simulator assessment scenario. We conducted an ANCOVA analysis to examine the association between the two conditions and the interaction of simulator immersion level. The idea that HMD VR would yield superior results compared to 3D desktop VR was not supported. The most important empirical finding of this study is the large direct effect, which shows that testing competency with maritime simulators transfers to real life, no matter how immersive the simulators are. We recruited an expert control group to test the design using a control variable and determined that the expertise level had no meaningful impact on HMD interaction. The control analysis found that simulators can be used to assess the competency of professional maritime engineering personnel. We examined how motivation affects the connection between competency in simulator assessment and the transfer of training. The moderation model confirmed with a large model fit the established knowledge regarding the significance of motivation as an essential factor in training transfer.

Our research findings support the use of immersive simulator technologies in assessing professional competency, benefiting the field of training research and practitioners of simulator training. This should be further investigated considering the commonalities among professional disciplines, relevance to other target populations, the content being transferred, and the characteristics of the transfer phenomena. We have demonstrated that state-of-the-art immersive technology is beneficial to predict real-world scenarios and invite additional contributions to further enhance competency development, whether in the education of students or in the training of professionals.

**Author contributions** SH led the work from start to finish, with the main contribution to theory, formulation of the concept, operationalization of the research design, data collection, data analysis, and discussion. SM supported the process from start to finish and contributed to the direction and finalisation of theory and discussion, language, and article structure.

**Funding** Open access funding provided by University Of South-Eastern Norway

**Data availability** Data is the property of the first author. The dataset is archived in accordance to FAIR principles on USN Research Archive. The data has been archived: Hjellvik, Simen (2024). Training transfer validity of virtual reality simulator assessment. University of South-Eastern Norway. Dataset. https://doi.org/https://doi.org/10.23642/usn.25442119.v1.

## Declarations

**Conflict of interest** The authors declare no competing interests.

## References

Aguinis H, Kraiger K (2009) Benefits of training and development for individuals and teams, organizations, and society [Review]. Annu Rev Psychol 60:451–474. https://doi.org/10.1146/annurev.psych.60.110707.163505

Alvarez K, Salas E, Garofano CM (2016) An integrated model of training evaluation and effectiveness [Article]. Hum Resour Dev Rev 3(4):385–416. https://doi.org/10.1177/1534484304270820

Baldwin TT, Ford JK (1988) Transfer of training: a review and directions for future research. Pers Psychol 41(1):63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

Baldwin TT, Kevin Ford J, Blume BD (2017) The state of transfer of training research: moving toward more consumer-centric inquiry. Hum Resour Dev Q 28(1):17–28. https://doi.org/10.1002/hrdq.21278

Barnett SM, Ceci SJ (2002) When and where do we apply what we learn? A taxonomy for far transfer. Psychol Bull 128(4):612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bell BS, Tannenbaum SI, Kevin Ford J, Noe RA, Kraiger K (2017) 100 years of training and development research: what we know and where we should go. J Appl Psychol 102(3):305–323. https://doi.org/10.1037/apl0000142

Bloom BS (1956) Taxonomy of educational objectives: the classification of educational goals: book 1 cognitive domain (Vol. 16). Longman

Blume BD, Ford JK, Baldwin TT, Huang JL (2010) Transfer of training: a meta-analytic review. J Manag 36(4):1065–1105. https://doi.org/10.1177/0149206309352880

Colquitt JA, LePine JA, Noe RA (2000) Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research. J Appl Psychol 85(5):678–707. https://doi.org/10.1037/0021-9010.85.5.678

Cummings JJ, Bailenson JN (2015) How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol 19(2):272–309. https://doi.org/10.1080/15213269.2015.1015740

Doozandeh P, Hedayati S (2022) The effect of simulation fidelity on transfer of training for troubleshooting professionals: a meta-analysis. IISE Trans Occup Ergon Human Factors 10(3):117–134. https://doi.org/10.1080/24725838.2022.2099483

Duncan T, Pintrich P, Smith D, Mckeachie W (2015) Motivated strategies for learning questionnaire (MSLQ) manual. Nat Center Res Improve Postsecond Teach Learn. https://doi.org/10.13140/RG.2.1.2547.6968

Dweck CS, Leggett EL (1988) A social-cognitive approach to motivation and personality. Psychol Rev 95(2):256–273. https://doi.org/10.1037/0033-295x.95.2.256

Ford JK, Baldwin TT, Prasad J (2018) Transfer of training: the known and the unknown [Review]. Annu Rev Organ Psych Organ Behav 5(1):201–225. https://doi.org/10.1146/annurev-orgpsych-032117-104443

Ford JK, Smith EM, Weissbein DA, Gully SM, Salas E (1998) Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. J Appl Psychol 83(2):218–233. https://doi.org/10.1037/0021-9010.83.2.218

Gagné RM (1965) The conditions of learning. Holt, Rinehart and Winston. https://books.google.no/books?id=4lpIAAAAMAAJ

Grossman R, Salas E (2011) The transfer of training: what really matters [Article]. Int J Train Dev 15(2):103–120. https://doi.org/10.1111/j.1468-2419.2011.00373.x

Hayes AF (2022) Introduction to mediation, moderation, and conditional process analysis: a regression-based approach, 3rd edn. Guilford Publications, New York

Hjellvik S, Mallam S (2023) Integrating motivated goal achievement in maritime simulator training. WMU J Marit Aff. https://doi.org/10.1007/s13437-023-00309-2

Holland DF, Kraha A, Zientek LR, Nimon K, Fulmore JA, Johnson UY, Ponce HF, Aguilar MG, Henson RK (2018) Reliability generalization of the motivated strategies for learning questionnaire: a meta-analytic view of reliability estimates. SAGE Open 8(3):215824401880233. https://doi.org/10.1177/2158244018802334

Jensen L, Konradsen F (2018) A review of the use of virtual reality head-mounted displays in education and training. Off J IFIP Tech Comm Educ 23(4):1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Kvinge LMR, Engelsen KS (2016) Omsetjing og tilpassing av MSLQ: IMTE-rapport nr. 1/16. http://hdl.handle.net/11250/2399353

Lajoie SP (2021) Multimedia learning with simulations. In: Mayer RE, Fiorella L (eds) The cambridge handbook of multimedia learning. Cambridge University Press, pp 461–471. https://doi.org/10.1017/9781108894333.048

Makransky G, Borre-Gude S, Mayer RE (2019a) Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. J Comput Assist Learn 35(6):691–707. https://doi.org/10.1111/jcal.12375

Makransky G, Petersen GB (2021) The cognitive affective model of immersive learning (CAMIL): a theoretical research-based model of learning in immersive virtual reality. Educ Psychol Rev 33(3):937–958. https://doi.org/10.1007/s10648-020-09586-2

Makransky G, Terkildsen TS, Mayer RE (2019b) Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learn Instr 60:225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mayer RE, Makransky G, Parong J (2022) The promise and pitfalls of learning in immersive virtual reality. Int J Human-Comput Interact 39(11):2229–2238. https://doi.org/10.1080/10447318.2022.2108563

Moglia A, Ferrari V, Morelli L, Ferrari M, Mosca F, Cuschieri A (2016) A systematic review of virtual reality simulators for robot-assisted surgery. Eur Urol 69(6):1065–1080. https://doi.org/10.1016/j.eururo.2015.09.021

Mulders M (2023a) Confounding in educational research: an overview of research approaches investigating virtual and augmented reality. Digital Psychol 4(1S):9–12. https://doi.org/10.24989/dp.v4i1S.2227

Mulders M (2023b) Learning about victims of holocaust in virtual reality: the main, mediating and moderating effects of technology, instructional method, flow, presence, and prior knowledge. Multimodal Technol Interact. https://doi.org/10.3390/mti7030028

Parks L, Guay RP (2009) Personality, values, and motivation. Personality Individ Differ 47(7):675–684. https://doi.org/10.1016/j.paid.2009.06.002

Pintrich PR, Smith DAF, Garcia T, McKeachie W (1991) A manual for the use of the motivated strategies for learning questionnaire (MSLQ)

Radianti J, Majchrzak TA, Fromm J, Wohlgenannt I (2020) A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput Educ. https://doi.org/10.1016/j.compedu.2019.103778

Rauter G, Sigrist R, Koch C, Crivelli F, van Raai M, Riener R, Wolf P (2013) Transfer of complex skill learning from virtual to real rowing. PLoS ONE 8(12):e82145. https://doi.org/10.1371/journal.pone.0082145

Schmidt MW, Köppinger KF, Fan C, Kowalewski KF, Schmidt LP, Vey J, Proctor T, Probst P, Bintintan VV, Müller-Stich BP, Nickel F (2021) Virtual reality simulation in robot-assisted surgery: meta-analysis of skill transfer and predictability of skill. BJS Open. https://doi.org/10.1093/bjsopen/zraa066

Schubert T, Friedmann F, Regenbrecht H (2001) The experience of presence: factor analytic insights. Presence Teleoperators Virtual Environ 10(3):266–281. https://doi.org/10.1162/105474601300343603

Suh A, Prophet J (2018) The state of immersive technology research: A literature analysis. Comput Hum Behav 86:77–90. https://doi.org/10.1016/j.chb.2018.04.019

Sullman MJM, Dorn L, Niemi P (2015) Eco-driving training of professional bus drivers—Does it work? Transp Res Part C Emerg Technol 58:749–759. https://doi.org/10.1016/j.trc.2015.04.010

Taylor GS, Barnett JS (2013) Evaluation of wearable simulation interface for military training [Article]. Hum Factors 55(3):672–690. https://doi.org/10.1177/0018720812466892

van der Vleuten CP, Schuwirth LW (2005) Assessing professional competence: from methods to programmes. Med Educ 39(3):309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x