CHAPTER 9

Evaluating Serious Game Trainings

P. Saskia Bayerl, Steffi Davey, Philipp Lohrmann, and Jonathan Saunders

9.1 Introduction

Developing and conducting serious game-based trainings can consume considerable resources and time. In order to ensure that such training methods provide optimal outcomes, it is therefore advisable to establish early on appropriate success criteria that quantify the acquisition or improvement of knowledge, skills, attitudes or behaviours. This chapter offers recommendations on how to plan and conduct evaluations of virtual reality (VR)-based serious games for training purposes. A concrete example of such an evaluation from the AUGGMED project (see Chap. 5) showcases the practical steps of running meaningful VR-based training evaluations as well as potential challenges that need to be taken into account.

P. S. Bayerl (✉) · S. Davey · J. Saunders
CENTRIC, Sheffield Hallam University, Sheffield, UK
e-mail: p.s.bayerl@shu.ac.uk

P. Lohrmann
ROQAS Solutions Limited, Hornchurch, UK

B. Akhgar (ed.), *Serious Games for Enhancing Law Enforcement Agencies*, Security Informatics and Law Enforcement,
https://doi.org/10.1007/978-3-030-29926-2_9

9.1.1 *Why Game and Training Evaluations?*

Virtual reality (VR)-based trainings using serious games are a relatively new development but are already changing the ways in which organisations such as police forces, military or first responders train their staff (see Chaps. 5, 9, 10 and 11). It is therefore imperative to ensure that such new training methods are (at least) as effective as traditional trainings. Systematic training evaluations are the primary tool for this purpose.

Generally, training evaluations aim to understand whether a training is effective in achieving the intended learning outcomes (for details on learning outcomes see Chap. 1). This is particularly important when developing new trainings to test whether the content, setup and presentation do deliver the expected results (*training validation*). Training evaluations can further establish which training format is better suited to a particular purpose, for instance, comparing learning outcomes using different materials, training durations or individual versus group settings (*comparative evaluation*). Lastly, evaluations can be used to test whether an existing training can be transferred to other settings (*training transferability*; e.g. establishing whether a training developed in one cultural context or industry can be used as effectively in another culture or industry) or whether an established training still delivers the expected outcomes or may need adaptation (*long-term viability*).

Thus, training evaluations can (and should) be conducted at various stages of the training process – from testing the original training concept, including the technological setup and scenario, to the continued and long-term performance of the training, once it has been deployed. To ensure valid results, evaluations need to be planned systematically, which means evaluation exercises are often quite resource-intensive. To help practitioners in this process, the following sections outline the main considerations in the planning and execution of training evaluations.

9.2 Aspects to Evaluate

What to evaluate hinges on the question what the purpose of the evaluation is: Is the intention to test the effectiveness of the complete training or only of specific aspects (such as the spacing of training sessions or specific VR materials or exercises)? In this chapter, we focus primarily on evaluating complete trainings, as our interest is in outlining a framework for evaluating the quality of VR-based serious games as a training method. (In the same regard, the steps and considerations presented in this chapter can

easily be adapted to a more targeted evaluation of sub-elements of a training.)

Our discussion focuses on three aspects that together can establish the effectiveness of VR-based trainings:

1. ***Employed technology including the training scenarios*** – aims to understand whether the technology works as expected and is usable (user-friendly) for the intended target group. For VR-based trainings, this test should also include a check of the appropriateness of the scenario(s).
2. ***Training satisfaction of participants*** – aims to understand whether participants are satisfied with the experience, either in its totality (general satisfaction) or in specific aspects (e.g. setting, trainer, perceived usefulness).
3. ***Impact*** – aims to understand whether the training achieved the intended learning outcomes.

9.2.1 Evaluation of the Technology Including Training Scenarios

Technology evaluations aim to understand whether the technology works as expected and is usable (functional as well as user-friendly) for the intended target group. This step should also include an explicit test of the appropriateness of the scenario(s).

Functionality Testing

Functionality testing assesses whether the technology (hardware and software) meets all functional requirements as set out by the end user. This stage will also include checking for any bugs in the system. It is vitally important to evaluate the performance statistics and acquire minimum requirements for the hardware to run the game, as any lag can cause cybersickness in VR (LaViola Jr, 2000; see also Chap. 2) and a generally uncomfortable experience.

Focusing on the VR scenarios, this step needs to check for and eliminate any inconsistencies or inaccuracies in the scenarios. This includes assessing the accuracy of scenarios against the initial design as well as testing all interactions that are possible in the virtual environment (e.g. for interaction fidelity; see Chap. 7). Serious games and virtual reality

trainings allow for considerable scalability, as simulations can be easily adapted to changing training needs. Whilst this is a key advantage of this training method, evaluators should consider that any changes to the simulations and scenarios may have an impact on the efficacy of the training. Thus, any scenario change or modification to the gameplay (including updates to hardware and game mechanics) should be re-evaluated.

Usability Testing

Virtual reality presents a new challenge for usability evaluations. Desktop applications tend to conform to a standard layout (Bevan, 2001), which utilises design techniques to ensure the application is as intuitive as possible – meaning that on first exposure to a desktop application, users will be able to make assumptions about interactions and where to find certain buttons and functionalities. Most users have not yet adapted to the immersive interfaces in VR. Therefore, it is important to evaluate the usability of an immersive virtual environment with the specific user group(s) for which the training is intended (also considering potential disabilities). This covers also simple issues such as the fact that users may be unsure where the intended interaction interface is located, which can cause confusion and aggravation. Already simple hints (e.g. arrows) can help to guide users' gaze to the correct location (cp. Chap. 2). Storytelling techniques[1] can also be employed to encourage users to face in the intended direction. As there are a multitude of interacting factors that may have an impact on the efficacy of such methods, a thorough comparative evaluating design choices is advisable.

Usability testing should further investigate duration scenarios, as users are advised not to spend a prolonged period of time in virtual reality. One recommendation is to take a break after 30 minutes (Fagan, 2018). Furthermore, as movement in VR can create cybersickness, it is important to test the controls of the movement, specifically their speed and acceleration (see Chap. 2). Important is also the required degree of immersiveness, i.e. the potential realism of situations. Immersiveness is one of the major benefits of VR-based trainings and a reason why they have found enthusiastic application from medical education (e.g. Alfalah et al., 2019) to psychotherapy (e.g. Morina, Ijntemaa, Meyerbröker, & Emmelkamp, 2015) to firearms trainings (Wei et al., 2019). In the same regard, highly realistic scenarios may be problematic when they lead to (unintended) stress or anxiety in participants (Slater, Khanna, Mortensen, & Yu, 2009).

Creating physical and mental stress reactions can be a necessary part of a training, for instance, if the training is meant to prepare first responders

[1] http://blog.leapmotion.com/art-storytelling-narrative-vr/

to physically and mentally handle the aftermath of bombings by highly realistic representations of the destruction, including wounds and people in distress (see Chaps. 5 and 9) or to train correct reactions in critical situations (see Chap. 10). In fact, indications are that stress as part of trainings may actually enhance learning (cp. Joël, Pu, Wiegert, Oitzl, & Krugers, 2006). However, stress may also inhibit learning when it is experienced at the wrong time (e.g. shortly before or a day after a learning task) or takes too long to abate (cp. Joël et al., 2006). In consequence, careful consideration should be given to the degree of realism in training scenarios in relation to the expected training effects (how much overall, for which elements, at which phase of the training process, etc.) and these elements tested accordingly with the intended user groups.

9.2.2 Degree of Training Satisfaction by Participants

Training satisfaction measures whether participants 'liked' or 'enjoyed' a training or whether they found it 'useful' for their work. Typical satisfaction questions are shown in Table 9.1. Such subjective evaluations are easy and quick to administer but will provide few insights into whether actual learning has taken place. Acceptance questions are therefore often considered the weakest indicator of training success.

Still, they can be helpful for gathering an overall impression by participants. Next to general satisfaction with or the perceived usefulness of the training, acceptance questions can also address a multitude of aspects from teacher(s) to class format, duration, size of the group, materials and assessment procedures to marketing and catering. They can thus identify potential irritants (e.g. materials that are hard to read or hear, examples that are too disturbing or seen as irrelevant, pleas for more or fewer coffee breaks) as well as strong points. Yet, while helpful, acceptance questions cannot replace proper impact assessments, and training evaluations should thus never rely solely on subjective assessments by participants alone.

Table 9.1 Examples of typical training satisfaction questions

1	Overall, the on-the-job training I receive is applicable to my job.
2	Overall, the training I receive on the job meets my needs.
3	Overall, I am satisfied with the amount of training I receive on the job.
4	I am generally able to use what I learn on the job.

Based on Schmidt (2007)

9.2.3 Impact Assessment

Impact assessments aim to establish whether the training has the expected (or any) influence on participants. To do this, impact measurements need to determine the type and degree of effects on participants in a systematic and repeatable way. A first consideration in the planning of impact assessments is the type of expected learning outcomes, i.e. the concrete aspect(s) or content(s) the training should teach or improve (see also Chap. 1). Secondly, evaluators must ask themselves whether they are primarily interested in testing the immediate effect of the training (i.e. training outcomes) or also in how well the leaning transfers to the workplace (i.e. its effect on job performance). Thirdly, a decision must be made whether effects are (only) assessed directly or shortly after the training (i.e. measuring short-term impacts) or also at later stages (i.e. measuring longer-term impacts). In the following, we provide more details on these aspects.

Different Levels of Learning Outcomes

The primary focus of an impact assessment depends on the purpose of the training (i.e. what are the aspects this training should teach or improve). This may be correct reactions in crisis situations (e.g. Haferkamp, Kraemer, Linehan, & Schembri, 2011; cp. Chap. 10), knowing the correct way to evacuate a building (Feng, Gonzales, Amor, Lovreglio, & Cabrera-Guerrero, 2018; cp. Chap. 9) or raising cyber security awareness (Jin, Tu, Kim, Heffron, & White, 2018; cp. Chap. 11). Hence, in planning impact assessments, a clear view on the outcome or outcomes is required: What is the purpose of the training?

A useful framework to support a systematic planning of impact evaluations is Kirkpatrick's model of training evaluation criteria (Kirkpatrick, 1967) and its extension by CAIPE (1999). This framework differentiates four different levels of outcomes (also Carpenter, 2011, p. 124):

1. ***Level 1: Reactions –*** participants' views of their learning experience and satisfaction with the training
2. ***Level 2: Learning***

    - Level 2a – modification in attitudes and perceptions
    - Level 2b – acquisition of knowledge and skills

3. ***Level 3: Behaviour***

   - Level 3a – changes in behaviour of the training participant
   - Level 3b – changes in organisational practice

4. ***Level 4: Results/benefits to training participants and others***

This framework systematically outlines the disparate levels of outcomes – from immediate reactions to the training (level 1) to internal states (emotions, attitudes, perceptions; level 2) and observable behaviours (level 3) to broader consequences to the individual, its organisation and related people or groups (level 4). Table 9.2 presents examples and possible methods for the assessment of each level.

As stated previously, which level(s) to focus on will depend on the training purpose. For a comprehensive understanding of training effects, it is often valuable, however, to consider multiple angles, i.e. include several levels in the evaluation. This is certainly recommendable for VR-based serious games trainings, as they allow for a high level of realism in their scenarios and are thus often employed for the training of complex and dynamic topics.

Training Outcomes Versus On-The-Job Performance

The real test for the usefulness and success of a training is whether new knowledge and skills are transferred into the work situation. This process is called *training transfer* and denotes the degree to which training participants manage to apply newly acquired knowledge, attitudes, behaviours, etc. in their work (Baldwin & Ford, 1988).

A simple differentiation between immediate training outcomes and training transfer may be stated as follows:

- ***Training outcomes*** **–** test whether participants can replicate the attitudes, knowledge, skills, etc. acquired during the training usually in the form of standardised tests (e.g. in the form of exams or during pre-designed VR scenarios)
- ***Training transfer*** **–** tests whether participants can apply the learned aspects as part of their job and thus in realistic, novel and usually non-standardised situations

Table 9.2 Methods to assess training impacts on different outcome levels

Evaluation level	*Examples*	*Method suggestions*
Level 1 – reactions	Satisfaction with the training, perceived usefulness for participants' work	Surveys/interviews with training participants
Level 2a – modification in attitudes and perceptions	Implicit biases in decision making, attitudes towards safety measures in the workplace	Surveys/interviews with training participants, assessment of reaction times in game (e.g. to test for implicit biases or stress reactions)
Level 2b – acquisition of knowledge and skills	Indicators for human trafficking, theoretical models for radicalisation	Written knowledge and skills, observation of skill/knowledge application in behaviours (in game or real life)
Level 3a – changes in behaviour	Compliance with safety procedures	Observation of the training participant (in game or real life), changes to indicators of job performance before and after the training
Level 3b – changes in organisational practice	Effectiveness of communication between agencies in crisis situations	Observations of people beyond the original training participant, objective measures of performance and practices at group and/or organisational level before and after the training
Level 4 – results/ benefits	Fewer incidents of stress-related disorders	For subjective benefits/internal states: surveys and interviewsFor objective/ quantifiable benefits: observations and objective measures

Hence, while training outcomes test the direct effectiveness of trainings, training transfer answers the much more relevant (as well as challenging) question whether the training actually improves job performance. Generally speaking, improved on-the-job performance can be considered as an even stronger measure of training effectiveness than successful training outcomes.

Training transfer is notoriously difficult. Not only do training effects often fade over time; learning from training sessions are often hard to apply in (work) situations that tend to be more complex, dynamic and unstructured than usually encountered in a training setting. According to some estimates, 40% of participants are unable to transfer learning directly after a training and 70% fail to do so after 1 year (cp. Saks, 2002).

Testing for training transfer must prove two things: 'the generalization *and* maintenance of newly acquired knowledge and skills on the job' (cp. Saks, 2002, p. 29; emphasis added). Saks gives further recommendations on how to assess transfer in a comprehensive way (p. 29):

- ***Establish rate of decay in learning*** – '[the assessment] should be specific as to the length of time following training'.
- ***Measure both outcome levels 3 and 4*** – 'estimates of the transfer of training should consider behaviour and results criteria'.
- ***Establish return on investment*** – 'a transfer estimate should … consider the percentage of training investments that result in transfer'.

Several things can support training transfer. While a comprehensive overview is outside the scope of this chapter, the following are examples of design decisions that facilitate training transfer (cp. Blume, Ford, Baldwin, & Huang, 2010; Burke & Hutchins, 2007 for a more extensive discussion):

- Trainees need to perceive a close relationship between training content and own work tasks (referred to as 'identical elements'; Thorndike & Woodworth, 1901 cited in Burke & Hutchins, 2007).
- A high number of practice elements and frequent feedback; especially for skills that are not used very often, a very high rate of practice repetitions beyond the point of first learning a skill (i.e. overlearning) helps retention.
- Active learning ('learning by doing').
- Systematic feedback not only when tasks are done correctly, but also detailed feedback about errors and mistakes.

Considering the above, VR-based serious game trainings already offer many of the elements that support training transfer (e.g. high level of realism and a strong focus on active learning). Although systematic evidence on this point is still missing, it suggests that VR-based trainings may be well-suited for trainings that require efficient transfer to real-world situations.

Short- Versus Long-Term Effects

The question of training transfer touches on the question of how long the acquired knowledge, skills or behaviours are retained. The successful retention of learned content is affected by many factors, not least by:

- ***Type of content*** – trainings for physical tasks and using natural tasks tend to result in better retention long-term compared to trainings for cognitive and artificial tasks (Arthur Jr, Bennett Jr, Stanush, & McNelly, 1998).
- ***Spacing of material*** – massed practice (i.e. trainings that present new knowledge in large chunks) leads to faster acquisition of learning and better immediate recall, while distributed practice (i.e. spreading smaller chunks of materials over more sessions) results in better retention long-term (Kim, Ritter, & Koubek, 2013; Schendel & Hagman, 1991).
- ***Type of training*** – game-based trainings seem to lead to better knowledge retention compared to paper-based trainings (Ricci, Salas, & Cannon-Bowers, 1996).

How long-lasting effects are can – and should – be tested. A post-test immediately after a training checks whether participants really understood the content and can replicate the new knowledge, skill, etc. directly after a learning session (short-term effects). Test(s) after a longer period – anything from days to years – indicates how much of the learning is still retained (long-term effects). Longitudinal approaches in the form of repeated measures can be used to assess the *degree of retention* (i.e. how much trainees still know after a given time) and the *rate of forgetting* (i.e. how much have trainees forgotten in the space of a given time). The next section presents longitudinal evaluation designs together with other common evaluation setups.

9.3 Common Evaluation Designs

Evaluations aim to establish whether a training serves its purpose, i.e. leads to the envisioned outcome(s) across participants (e.g. changes in behaviours, modifications in attitudes, improved knowledge, etc.). Several setups are feasible to accomplish this task: from the easiest – a simple test after the training – to more complicated ones such as longitudinal designs with systematically sequenced and mixed conditions to establish causal and differential effects among settings over time. Figure 9.1 presents an overview of common training evaluation designs. (Other options are available, of course, depending on the purpose of the evaluation; e.g. Hanley, Chambers, & Haslam, 2016; Howard & Jacobs, 2016.)

Below are short descriptions of the presented designs:

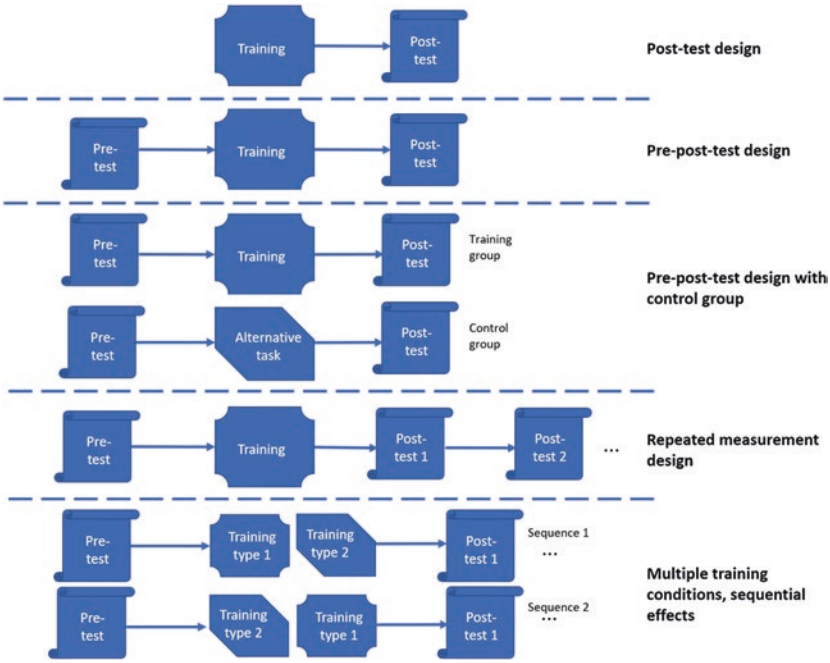

Fig. 9.1 Different evaluations designs

- ***Post-test design*** – All participants will complete the training and subsequently complete an assessment (post-test) in order to evaluate the extent to which the participants have achieved the learning outcomes. This evaluation design is the simplest to implement and requires the least resources and time commitment.
- ***Pre-post-test design*** – This design introduces an initial assessment (pre-test) to provide an indicator of the participants pre-existing knowledge on the subject area (baseline). Pre- and post-test should be identical – or at least similar enough – to allow a direct comparison between the two measurements and thus an indication of how much and which aspects the training has improved compared to what learners already knew before the training.
- ***Pre-post-test design with control group*** – This design adds a control group, i.e. a group that does not receive a training, but ideally an alternative task. The alternative task aims to ensure that it was the

training itself that improved participants' knowledge, skills, attitudes, etc., instead of other unrelated aspects such as the extra attention to trainees or the social interaction during the training. The alternative task should be unrelated to the training content but similar in nature to the training content (e.g. if the training is on a physical task, the alternative task also should be of a physical nature). This design can be easily expanded to compare training setups against each other by including additional groups that go through training variations.

- ***Repeated measurements design*** – Post-tests tend to be administered directly after the training and thus only measure immediate effects. Even if they are measured at a later stage, one test does not provide an indication of how the learned materials are retained or used over time. Repeated measures aim to do exactly that – assess how much and which knowledge is retained or forgotten. That is, using the same test multiple times can indicate whether the training leads to long-term changes to people's skills, knowledge or behaviours. These repeated tests can be either done in the same context (e.g. by using the same online knowledge tests) to facilitate direct comparability of results or conducted across different contexts (e.g. different work situations) to assess whether learning is transferred long-term and across situations.
- ***Multiple training conditions, sequential effects*** – This design is suitable when aiming to compare the effectiveness of combining multiple trainings to identify the best combination, sequence and/or spacing between trainings. This design can be expanded with sequences using the same type of training or adding intermediate tests between training blocks (not shown in Fig. 9.1).

The choice of the 'best design' is not always straightforward. Generally, the randomised controlled trial (RCT) design is considered a 'gold standard' for training evaluations. This design is a pre-post-test design with control group that also ensures that participants are assigned to training and control group conditions at random. Randomness in the assignment aims to reduce the possible impact of systematic differences among participants, i.e. random assignments should 'even out' potential effects of demographics, job function, educational background, etc. on training outcomes.

RCTs are resource intensive and can be problematic if participants in the control group may accrue negative effects from not being included in

the training group (e.g. if it means not receiving knowledge or advice that may safeguard people's work performance or personal well-being). Sackett and Mullen (1993) advocate a pragmatic approach. They acknowledge that in the organisational reality running a full RCT study may not always be feasible given the time and resources. They further suggest clarifying whether the evaluation needs to establish a *change or improvement* in knowledge, skills or behaviours, or whether it is sufficient to test that the participants demonstrate a certain *level or degree* of learned content. In the first case, pre-post designs are required; in the second case, a post-test suffices (although in the understanding that is then impossible to say whether it was the training that led to the level of knowledge, skills, etc.). As stated above, this pragmatic approach is not commonly shared.

While we advise to conduct comprehensive and systematic evaluations whenever possible especially for newly developed trainings, we agree with Sackett and Mullen that more pragmatic approaches are still better than no evaluation at all. If full evaluations are impossible, less formal designs may be chosen in the understanding that their results may be less easy to interpret and less generalisable. Any limitations in the interpretability of results – given a chosen evaluation design – should be clearly formulated and communicated when reporting on the evaluation results.

9.3.1 Are There 'Minimal' or 'Optimal' Setups?

The main concerns when aiming to validate the quality and fit of a (newly developed) serious game-based training must be to rule out that factors other than the training itself are responsible for the observed outcomes. Hence, a minimum standard for any training evaluation can be formulated as follows:

1. Use of a ***pre-test*** to exclude the possibility that the serious game training had no effect on expected outcomes compared to participants' status before the training
2. Use of a ***control group*** to exclude the possibility that factors other than the serious game training are responsible for the training results
3. Use of ***participants that are representative*** in all essential features for the target population of the serious game
4. Use of a ***sufficient number of participants*** to allow statistical testing, where relevant: per condition and across all time points (for

longitudinal designs also considering that drop-out rates can be as high as 70% if using survey methods for assessment; Overall, Tonidandel, & Schmitz, 2009)
5. Use of ***adequate assessment methods*** for each relevant outcome levels

An optimal design is much less easy to formulate, as it will depend on the specific goals of the serious game and of the evaluation. Yet, in our view, ideally a training evaluation should include all relevant aspects (from technology to acceptance to impact) and ensure that impacts are assessed longer term and in the best case directly on the job.

9.4 A Training Evaluation Example from the AUGGMED Project

The AUGGMED project[2] was an EU-sponsored research project conducted from 2015 to 2018. The project developed a number of serious games in virtual and augmented reality for the training of first responders, including police officers, coast guard personnel and paramedics (see Chap. 5 for details). As part of the project work, an evaluation study was undertaken to investigate the use of virtual reality in the training of police officers. Its intention was to establish whether the VR-based training was as effective as the more resource-intensive traditional hands-on training.

For this purpose, a pre–post-test design with control group and random assignment was employed. The traditional training acted as control conditions, since it constituted the current standard the VR-based training aimed to improve on. In total, 80 UK police officers were recruited as participants. Most of them were trainee officers with only a few months' experience; however, a few long-standing officers were included to make up numbers.

The training objective for the officers was to learn the correct handling of a suspicious parcel. All participants first watched a number of video lectures on the topic, introducing them to the recommended procedures and best practices. Their baseline level of understanding was then assessed with a multiple-choice knowledge check.

[2] http://auggmed-project.eu/AUGGMED has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653590.

Subsequently, the participants were randomly grouped into pairs, and each of the 40 pairs were randomly assigned to one of four training groups. Each group completed two consolidation exercises in which they had to apply the procedures they had learned to a simulated suspicious parcel scenario. These exercises were acted out either in VR or in the form of a live role play in a mock-up environment, which replicates realistic situations (see Fig. 9.2). To test for order effects of VR and live role play, the following four groups were tested against each other:

- ***Group 1*** **–** conducted two exercises in mock-up reality (representing the traditional training and thus the control condition)
- ***Group 2*** **–** conducted two exercises in VR
- ***Group 3*** **–** conducted the first exercise in mock-up reality and the second in VR
- ***Group 4*** **–** conducted the first exercise in VR and the second in mock-up reality

Each pair of officers was accompanied by a trainer throughout both exercises. The trainers actively took part in the scenarios by playing the

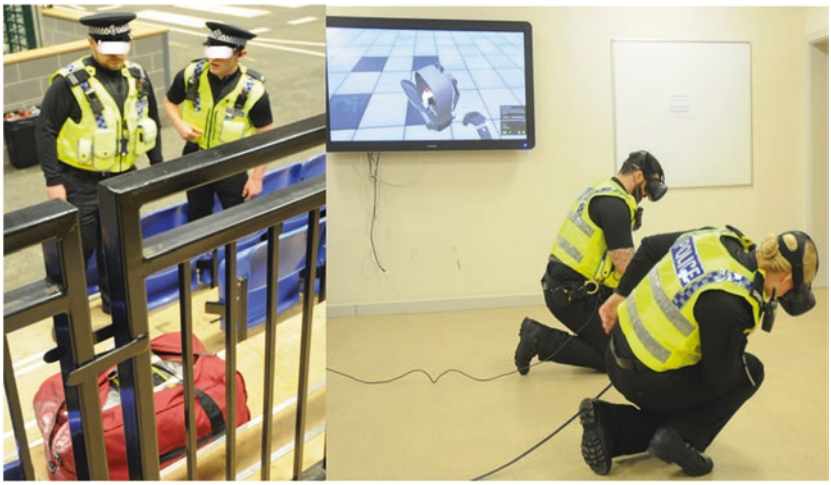

Fig. 9.2 Training participants investigate a suspicious parcel. Left: traditional training process using the mock-up of a re-created football stadium. Right: VR-based training

role of a local security guard who first meets the participants at the venue of each scenario. Throughout each exercise the trainer would stay in character and provide further background information when asked. In addition, the participants could use their radios to speak to the 'control room' in order to request the involvement of other services or to ask for further background information.

The participants using VR were given a brief technical introduction to the AUGGMED system before their first exercise in VR. This introduction was standardised and included familiarisation with the controls as well as a brief period of free play in a venue not used for any of the training scenarios.

All exercises were recorded with video cameras; in addition, the participants' body cams were used to capture further footage.

The impact of the training exercises was assessed in a number of ways. Immediately after completing their final exercise, participants were asked to retake the baseline knowledge check. In addition, the trainers used standardised marking sheets during the exercises to award points for following the correct procedures. These scores were later consolidated by an independent assessor using the video footage of the exercises. Finally, participants were asked to fill in another knowledge check 2 months after the training in order to assess the long-term learning outcomes.

The results of the knowledge check were used to compare the participants' level of competence before and after the consolidation exercises as well as to compare the four groups. Similarly, the trainers' scores were compared across the groups in order to assess the efficacy of the VR training in comparison to the traditional live role plays.

The evaluation showed that the training increased the overall knowledge of participants about correct procedures and their ability to correctly conduct checks. Further, the four groups did not differ significantly in their post-test outcomes (Saunders, Davey, Bayerl, & Lohrmann, 2019). This indicates that the VR-based training can be as robust and effective as the traditional, hands-on training. At the same time, it considerably reduced the resources needed to conduct the training sessions (i.e. no need to build large-scale mock-ups, to transport trainers and trainees to the location, etc.). In sum, the evaluation suggests that VR-based serious game trainings provide a viable alternative to hands-on trainings for the acquisition of complex procedures.

9.5 Additional Considerations

VR-based serious games have a number of attractive features for training in a law enforcement context. For instance, it is possible to put trainees in situations that are difficult or costly to replicate with sufficient realism in real life (e.g. the aftermath of natural disasters, humanitarian crises or war situations; cp. Chaps. 9 and 10). Furthermore, compared to mock-ups of locations such as in the AUGGMED evaluation, virtual scenarios are relatively easy to modify and adapt. VR-based trainings also possess characteristics that can support learning and training transfer such as an emphasis on active learning and the possibility to create a high number of practice elements that allow for immediate feedback (see Chap. 1).

One useful element of VR-based serious game trainings is that the medium used for the training can also serve as medium for the training evaluation itself. This reduces the disconnect that often occurs between what has been trained (e.g. the right way to search a suspicious bag) and the way it is tested (e.g. through online or paper-and-pencil tests). This similarity between training and testing increases the likelihood that trainees can correctly recall newly learned knowledge or behaviours. On the other hand, if the settings or scenarios used for training versus testing are too similar, testing can turn simply into 'rote repetition'. Test scenarios should thus at least vary in aspects directly related to the intended learning outcomes (e.g. if the intended learning outcome are cross-cultural negotiation skills, test scenarios could introduce variations in cultural context, the conflict that needs to be solved or the number and type of people engaged in the negotiation), with (well-reasoned) differences in the level of complexity between training and test.

It is not uncommon for users of virtual reality to experience symptoms similar to motion sickness (see Chap. 2). This can negatively affect learning as well as testing and thus skew results of an evaluation study. Factors that are said to affect the likelihood of cybersickness include gender (women tend to be more susceptible to cybersickness than men), age (susceptibility tends to decrease with age) and illness (individuals suffering from an illness may be more susceptible; LaViola Jr, 2000). These factors should be taken into consideration during training and testing to

ensure that any confounds caused by cybersickness are accounted for. Recent research further suggests that there may be a negative correlation between the sense of presence[3] and cybersickness (Weech, Kenny, & Barnett-Cowan, 2019). Creating scenarios with higher presence thus may be advisable especially for groups with a higher likelihood for cybersickness.

Mancuso, Chlup, and McWhorter (2010) conducted qualitative research to investigate barriers of learning in virtual environments from the perspective of adult users. One of the barriers identified were glitches in the technology, which participants found annoying and stifled learning during the VR training. Hence, during evaluation studies any technical glitches should be documented to prevent technical problems that are falsely attributed to problems with the training itself. Another barrier was the learning curve for interacting with the virtual environment. Attempting to learn how to interact with new technologies can lead to poor retention of training content. Individuals who are well-versed in computer games or virtual reality technologies may thus have an advantage over users will little experience. This could create an 'experience divide' in results, i.e. systematically better scores for people with previous exposure to VR technologies. In a similar vein, the current lack of familiarity with virtual reality can create an initial sense of awe. This can equally detract from the intended learning outcomes of the serious game. Hence, previous experience with VR should be assessed prior to each training and included as a control variable in the analyses.

Another issue arises due to the difficulty of accurately replicating the real world in virtual reality; for example, if students are subjected to harsh weather conditions during the real-world training, this would be difficult to replicate in a realistic manner in the virtual reality training. Therefore, it is important to establish which elements of a training are essential and whether they can be replicated in VR – or in a combination of VR and real-world settings (see Chap. 3). If crucial elements cannot be replicated in VR, then this training form may not be best suited and real-life trainings the better option instead.

[3] The sense of being present in the virtual place and time rather than in the actual real-world location (Sanchez-Vives & Slater, 2005).

9.6 Conclusion

A key benefit of virtual reality trainings is that they are often cheaper and less resource-intensive than traditional live training exercises. Still, it is important to understand the effects of moving to VR-based serious games. By evaluating virtual reality trainings in a comprehensive manner, it is possible to achieve an accurate idea of how effective the training is as a whole and how it can compare to traditional training methods. This is important to ensure that skills are not being lost with the modernisation of training methods. A comparison of training methods can also help to create an understanding of what elements are better trained using virtual (or augmented) reality and which may benefit from more traditional training methods. This chapter outlined important considerations and recommendations in planning and conducting evaluations of VR-based serious game trainings. Evaluations can be time- and resource-intensive. However, without systematic evaluations the effectiveness of trainings will always remain guesswork. This chapter hopes to provide helpful guidance on how to set up and conduct evaluations that encourages organisations and practitioners to obtain ongoing and systematic evidence about the effectiveness of their VR-based serious game-based trainings.

References

Alfalah, S., Falah, J., Alfalah, T., Elfalah, M., Muhaidat, N., & Falah, O. (2019). A comparative study between a virtual reality heart anatomy system and traditional medical teaching modalities. *Virtual Reality, 23*(3), 229–234. https://doi.org/10.1007/s10055-018-0359-y

Arthur Jr., W., Bennett Jr., W., Stanush, P., & McNelly, T. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance, 11*(1), 57–101.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63–105.

Bevan, N. (2001). International standards for HCI and usability. *International Journal of Human-Computer Studies, 55*(4), 533–552.

Blume, B., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105.

Burke, L. A. & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *Human Resource Development Review, 6*(3), 263–296.

CAIPE (1999). *Evaluating interprofessional education: A United Kingdom review for health and social care*. The British Educational Research Association and The United Kingdom Centre for the Advancement of Interprofessional Education, London.

Carpenter, J. (2011). Evaluating social work education: A review of outcomes, measures, research designs and practicalities. Social Work Education, 30 (2), 122–140.

Fagan, K. (2018). *Here's what happens to your body when you've been in virtual reality for too long*. [Online]. Available at: https://www.businessinsider.com/virtual-reality-vr-side-effects-2018-3. (Accessed 4 July 2019).

Feng, Z., Gonzales, V., Amor, R., Lovreglio, R., & Cabrera-Guerrero, G. (2018). Immersive virtual reality serious games for evacuation training and research: A systematic literature review. *Computers & Education, 127*, 252–266.

Haferkamp, N., Kraemer, N., Linehan, C., & Schembri, M. (2011). Training disaster communication by means of serious games in virtual environments. *Entertainment Computing, 2*, 81–88.

Hanley, P., Chambers, B., & Haslam, J. (2016). Reassessing RCTs as the 'gold standard': Synergy not separatism in evaluation designs. *International Journal of Research & Method in Education, 39*(3), 287–298.

Howard, M., & Jacobs, R. (2016). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): Two novel evaluation methods for developing optimal training programs. *Journal of Organizational Behavior, 37*, 1246–1270.

Jin, G., Tu, M., Kim, T.-H., Heffron, J., & White, J. (2018). Game based cybersecurity training for high school students. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18)* (pp. 68–73). New York: ACM.

Joël, M., Pu, Z., Wiegert, O., Oitzl, M., & Krugers, H. (2006). Learning under stress: How does it work? *Trends in Cognitive Sciences, 10*(4), 152–158.

Kim, J., Ritter, F., & Koubek, R. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science, 14*(1), 22–37.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig & L. R. Bittel (Eds.), *Training and development handbook* (pp. 87–112). New York: McGraw-Hill.

LaViola Jr., J. J. (2000). A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin, 32*(1), 47–56.

Mancuso, D. S., Chlup, D. T., & McWhorter, R. R. (2010). A study of adult learning in a virtual world. *Advances in Developing Human Resources, 12*(6), 681–699.

Morina, N., Ijntemaa, H., Meyerbröker, K., & Emmelkamp, P. (2015). Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of

studies applying behavioral assessments. *Behavior Research and Therapy, 17*, 18–24.
Overall, J. E., Tonidandel, S., & Schmitz, J. M. (2009). Testing the significance of difference in average rates of change in controlled longitudinal studies with high dropout rates. *Methodology, 5*(2), 46–54.
Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996). Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology, 8*, 295–307.
Sackett, P., & Mullen, E. (1993). Beyond formal experimental design: Towards an expanded view of the training evaluation process. *Personnel Psychology, 46*, 613–627.
Saks, A. M. (2002). So what is a good transfer of training estimate? A reply to Fitzpatrick. *The Industrial-Organizational Psychologist, 39*, 29–30.
Sanchez-Vives, M., & Slater, M. V. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience, 6*(4), 332–339.
Saunders, J., Davey, S., Bayerl, P. S., & Lohrmann, P. (2019). Validating virtual reality as an effective training medium in the security domain. In *Proceedings of the 2019 IEEEVR Conference*.
Schendel, J. D., & Hagman, J. D. (1991). Long-term retention of motor skills. In J. E. Morrison (Ed.), *Training for performance: Principles of applied human learning* (pp. 53–92). New York: Wiley and Sons.
Schmidt, S. W. (2007). The relationship between satisfaction with workplace training and overall job satisfaction. *Human Resource Development Quarterly, 18*(4), 481–498.
Slater, M., Khanna, P., Mortensen, J., & Yu, I. (2009). Visual realism enhances realistic response in an immersive virtual environment. *IEEE Computer Graphics and Applications, 29*(3), 76–84.
Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Frontiers in Psychology, 10*, 158.
Wei, L., Zhou, H., & Nahavandi, S. (2019). Haptically enabled simulation system for firearm shooting training. *Virtual Reality, 23*(3), 217–228. https://doi.org/10.1007/s10055-018-0349-0