Serious Games Evaluation: Processes, Models, and Concepts

Katharina Emmerich[1(✉)] and Mareike Bockholt[2(✉)]

[1] Entertainment Computing Group, University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany
katharina.emmerich@uni-due.de
[2] Graph Theory and Complex Network Analysis Group, University of Kaiserslautern, Gottlieb-Daimler-Straße 48, 67663 Kaiserslautern, Germany
mareike.bockholt@cs.uni-kl.de

Abstract. Serious games are developed with the goal of having a certain impact on players which goes beyond mere entertainment. This purpose-driven design is immanent to serious games and can be stated as the key characteristic that distinguishes serious games from other digital games. Hence, verifying that a serious game has the intended effect on the players needs to be an essential part in the development process. This and the following chapters are therefore dedicated to give a guidance how evaluation procedures can be planned and realized. The main focus is on aspects which are particularly distinctive to the evaluation of serious games, while methods and principles related to the evaluation of digital games in general will not be covered in detail. The structure of this chapter is as follows: After emphasizing the specific importance of evaluation for serious games, we describe a set of challenges which might occur in this context. In order to enable the reader to face these challenges, we present a framework of evaluation-driven design which offers guidance in the evaluation process. Other models which address different challenges are described before three examples of commendably evaluated serious games are discussed. These examples are intended to demonstrate how the presented abstract models can be applied in concrete evaluation procedures.

1 Introduction

This chapter gives comprehensive information about the evaluation of serious games. Each serious game is supposed to fulfil a certain purpose beyond mere entertainment. However, in order to ensure that this goal is achieved, it has to be tested in form of experiments and user tests. This process is called evaluation. Without evaluation, there is no evidence that the purpose of the game is achieved. Ideally, a serious game is evaluated with members from the target group in a comprehensive evaluation process. There are several criteria for a well-conducted and informative evaluation, as well as a range of challenges and problems that have to be met.

R. Dörner et al. (Eds.): Entertainment Computing and Serious Games, LNCS 9970, pp. 265–283, 2016.
DOI: 10.1007/978-3-319-46152-6_11

Serious games are often used in new contexts as a medium of intervention, and in many cases there are little to no existing successful examples for orientation. Thus, as there are few proven strategies or approaches to be guided by, the design and development of serious games is often based on literature, theories and to some part on intuition and personal experience (see also the chapter about game design in this book). Although this is a reasonable approach, it raises the need to test and confirm the underlying assumptions afterwards. However, the process of proving the effectiveness of a serious game is in general more involved than the evaluation of a commercial entertainment game, as a serious game needs to be especially evaluated with respect to its "serious" purpose. In the following, we will thus concentrate on the evaluation of the effectiveness of serious games in regard to their purposes, not on mere usability testing or the assessment of general player experience issues (e.g. fun), which is also a part of evaluation. Those aspects are interesting to study as well, but they are not serious games specific, thus techniques and methods from game design and evaluation in general can be applied easily and read about in other books (e.g. see [36,39,44] in the further readings sections). Here, evaluation is discussed in terms of its characteristics and specialities regarding the serious games context.

1.1 Overview

In order to approach the topic, this chapter is structured into four main parts: First, Sect. 2 declares the importance of comprehensive evaluation processes in general. Section 3 then emphasizes its complexity by pointing out characteristic challenges and problems related to the evaluation of serious games. Based on that, a framework for serious games evaluation is proposed in Sect. 4, which is supposed to provide guidance for the planning and realization of evaluations during and after the design process of a serious game. Finally, some examples of successfully conducted studies are presented in Sect. 5 and discussed with respect to the introduced framework of evaluation-driven design. The chapter concludes with a short summary and a selection of further reading regarding the topic of evaluation.

2 Importance of Evaluation

The idea of using digital games for purposes like learning, health promotion and persuasion is not new and has evolved into an extensive field of research. During the last decades, more and more serious games and applications have been developed and the endeavours to prove their effectiveness were high. Literature reviews show that there are several studies implying the benefit of serious games in general [3,6,25] or regarding different aspects, for instance as a tool to support learning processes or to induce health-related behaviour changes [4,7,9,11,26]. However, these reviews also indicate that there are even more studies and serious games which were not that successful due to sundry reasons. The question arises what makes a good evaluation and whether the results are worth the effort, i.e., why evaluation is important at all.

The overall goal of evaluation is to prove the game's effectiveness and suitability with respect to its designated purpose and application context. The purpose is thereby always in the center of investigation. Reliable results are supposed to lend credence to serious games, to convince diverse stakeholders and to inform future design approaches. Game researchers can learn more about the relationships between game design elements and the resulting player experience and thus gain insights into the impact of games in general. Additionally, researchers as well as developers are supposed to gain experience from both successful and failed game concepts and may thus improve in designing effective serious games [19]. Moreover, evidences of the effectiveness of serious games are necessary to convince the users of serious games themselves, as they have to believe and trust in their capability. Michael and Chen state that "[t]rainers and educators need to know whether or not the player has actually learned the content of the serious game" [22]. They thus underline the need of evaluation in order to gain a better acceptance of the games and to be competitive compared to established non-gaming interventions and programs. Those who have to support the use of serious games in their field of work, for instance teachers, physicians and other intermediaries, have to be convinced of their positive effect, because otherwise they will simply not recommend to use them [5,19]. Mayer et al. [21] mention two main reasons to conduct a structured and reproducible evaluation: accountability and responsibility. Accountability refers to the fact that users "have the right to know what they are actually buying, using or playing" [p. 234] and that they have to be convinced of the effectiveness of serious games. Responsibility, on the other hand, refers to developers and advocates of serious games and their duty to critically question the effects and consequences their games may have, especially in case of vulnerable target groups.

Furthermore, researchers and practitioners also emphasize the importance of evaluation for the commercial success and the growth of the serious games industry [2,10]. Successfully evaluated serious games can help to advance the dissemination and to optimize the image of serious games while at the same time adhering to the concept of responsibility. Without proper evaluation, the establishment of serious games as considered interventions is not possible, or as Kevin Corti, Managing Director from *PIXELearning*, puts it: Serious games "will not grow as an industry unless the learning experience is definable, quantifiable, and measurable. Assessment is the future of Serious Games." (quoted from Chen and Michael [2]). Summarizing these aspects, there are four different main groups of stakeholders that benefit from structured evaluations as shown in Fig. 1: serious games developers, researchers, intermediaries as well as the users. Overall, serious games evaluation is important for underlining the potential of serious games in various application fields and extending the effort in serious game development, research and application.

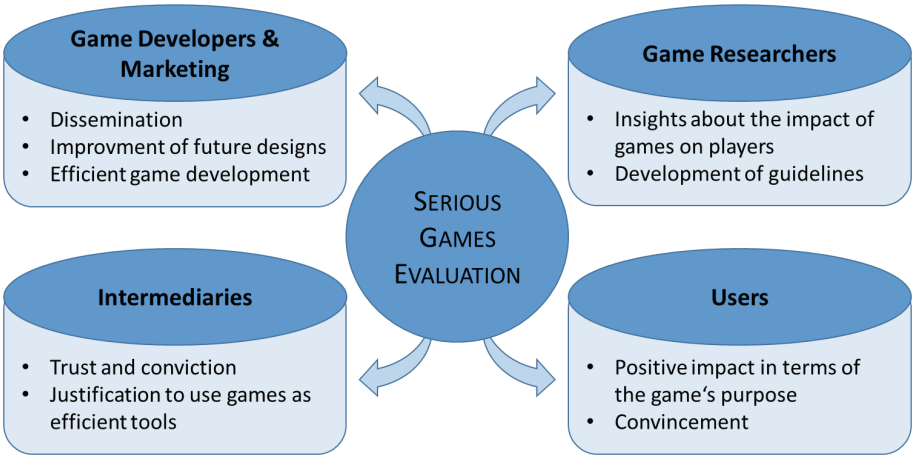

Fig. 1. Stakeholders and their advantages of serious games evaluation.

3 Challenges of Evaluation

As there are only few examples of comprehensively evaluated serious games, the question arises of what challenges and problems are there making evaluation an ambitious task. Most evaluation processes follow the same general structure: The study design is planned and set up, then participants are recruited and divided into different experimental groups. Normally, there are at least two groups, one that is provided with the serious game, called the *treatment* group, and one that does not use the game, called the *control* group. Of all groups, several data before, in the middle of and after the experiment needs to be logged and evaluated. Finally, this data is analyzed in terms of the purpose of the game and conclusions about its effectiveness are derived. However, Mayer et al. declare that "we lack an overarching methodology" [21]. According to them, the main problem is that there is a lack of comprehensive frameworks, theories, operationalized models, validated measurement methods, proper research designs, and generic tools for unobtrusive data gathering [21]. Hence, challenges appear all along the process of evaluation and can be further be classified into the following categories:

Recruitment of participants. Recruitment is often a time-consuming task, especially in the case of sensitive and vulnerable target groups like patients, children or disabled persons. If participants are already burdened with illness or complicated circumstances, which is true for many serious games for health, an experimental study is an additional strain. It is hard to find suitable volunteers and the risk of dropouts is high. However, a representative sample is important: Without a significant number of subjects a proper statistical analysis is not possible or will lead to results with just small effect sizes. Furthermore, it might occur that the recruited participants bring along different prerequisites, i.e., their previous knowledge or experiences with other

games are different. This can have an impact on the evaluation results and should be considered. If for example a serious game for learning is evaluated and learning through a serious game instead of conventional methods is a completely new experience for the participants, their attention for the game and also for the material to be conveyed might increase. In this case, the evaluation of the game will likely yield better results in the sense that the knowledge or skills of the participants have improved than if the participants are used to the application of games for learning. For the recruitment, it is therefore recommended to either choose a group of participants which is a representative sample for the intended target group of the game with regard to their prerequisites, or to determine their prerequisites beforehand and take them into account in the analysis and interpretation of the results.

Operationalization. Even if a representative sample and number of participants is achieved, the most sophisticated questions are what exactly has to be measured and how this can be done. Starting from the overall purpose of a serious game and the underlying theories, it has to be defined which concrete measurable aspects best reflect the game's purpose. This process is called operationalization. For instance, if the game's purpose is to support pupils in learning vocabulary, the number of recalled words after playing the game compared to their prior state of knowledge could be measured by a common vocabulary test. In other cases, the operationalization is much more ambiguous, for example if an increase of well-being is striven for. As operationalization and assessment are such complex processes, another full chapter goes into detail regarding this subject in this book.

Choice of measurement methods. Closely related to the issue of operationalization is the choice of measuring instruments. There are many different types of these instruments like questionnaires, physiological measures and observations, which all have different advantages and disadvantages regarding objectivity, validity and clarity. Hence, it is recommended to combine several kinds of methods to gain comprehensive insights into the effects of the game [1].

Design of control group conditions. Apart from methods of measurement, there are several more decisions to be made regarding the experimental design that can be challenging. One of those is the choice of an appropriate control group. Most serious games are supposed to be applied in contexts in which other (non-game) interventions with similar purposes already exist. In those cases, the evaluation of a serious game does not only have to aim at verifying a general intended effect of the game, but also has to take into account a comparison with existing solutions in order to answer the question of whether the game is better than established non-game alternatives. Girard et al. [9] propose to use at least two control groups instead of one: one group that receives no training at all, and one that receives a training with comparable contents as the game, but built on a different method, for example pencil and paper or classic teaching situations. Besides, it is also possible to create different versions of the game to identify the impact of single game elements. In any case, if the treatment of the control group is incomparable to the

serious games intervention, the results are incomparable and a badly chosen control group condition makes claims of effectiveness impossible.

Consideration of time-dependent effects. Another challenge related to the experimental design is timing. While it is already demanding to conduct a sound study with one point of measurement, most serious games would benefit from long-term assessment: Experiments in which participants are playing the game only once or for a short period of time, allow suggestions about short-term effects of the game, but do not take into account any wear-out effects. Due to its novelty, the game may attract more attention than established alternatives, but the resulting motivation to play it and to deal with the content might decrease after a while, impeding the impact of the game. Especially evaluation processes for games that are supposed to support long-term motivation of players should include this aspect as well.

Reach of effects. Besides considerations of time, it is also challenging to assess the reach of serious games effects in terms of the transfer to real-life contexts. Ideally, serious games evaluation includes both direct effects that playing the game has on players and subsequent effects that influence their future behavior in everyday life [9]. While this aspect is somehow interwoven with long-term effects, it does not describe effects over time, but rather defines on which level the effectiveness of a serious game is evaluated.

Processing of results. Finally, the evaluation of serious games bears the challenge to draw conclusions from results and meaningfully deploy them to improve the game. Evaluation does not terminate at the point that data is collected, but should lead to a process of revision to make the game more effective and appealing.

In sum, the result of an evaluation process should be data that is characterized by generalizability and validity [29]. To achieve this, we need a more structured way for evaluating serious games, hence in the following section we will present a framework of evaluation-driven design which is supposed to support the design of a proper and comprehensive evaluation process.

4 Evaluation Frameworks

The previous section contains a description of challenges that appear during an evaluation process and need to be handled. The following section is dedicated to the question of how to face these challenges. Because of the great variety of serious games regarding their background, purpose, or target audience, it is impossible to give concrete instructions for the evaluation procedure of every serious game, but it is possible to provide a set of guidelines and models which are general and abstract enough to be valid for the complete spectrum of serious games. For the specific evaluation procedure, guidelines and suggestions need to be derived carefully. They offer a useful standardization of evaluation procedures.

The section is structured into two parts: The first part contains the *framework of evaluation-driven design*, the second part contains a selection of further models

and frameworks proposed by other researchers which are of use for evaluation procedures.

The framework of evaluation-driven design embeds the evaluation and design process into the general structure of scientific working and can be considered as a "step towards a science of game-based learning" and serious games [27]. It offers guidance to when to evaluate, how to plan the evaluation process, and which questions need to be answered beforehand. Other existing models and existing frameworks either fall in one of the two categories of *when* and *what* can be evaluated. Models and frameworks of both categories are presented in the second part of the section in order to give additional guidance for an evaluation process. Models and frameworks which are not of specific use for the evaluation of serious games, but are valid in other areas and might be considered for the evaluation of serious games, are not explicitly mentioned. For those general methods, a list of references is given in the section for further reading at the end of this chapter.

4.1 Framework of Evaluation-Driven Design

The framework of evaluation-driven design offers guidance in the planning and realization of the evaluation of a serious game. It particularly highlights the role of evaluation during the design and development process which is why it is called framework of evaluation-driven design.

An illustration of the proposed framework is shown in Fig. 2, the next sections will describe the overall structure and the single components of the framework. The framework is intended to contextualize the well-known phases of a game development and evaluation process, set them in relation to each other and emphasize the similarity of the game development process with scientific processes. Therefore, the framework consists of mainly three phases: the preparation phase, the design phase and the evaluation phase, where the design and the evaluation phase are closely interlinked with each other and are iterated as often as necessary. It is clearly to see (and also intended) that neither of the elements is formulated in a detailed manner because each of the elements is a broad topic itself which needs elaboration and reflection. For the process of game design for example, there has been done a lot of research and practical studies which yielded a great variety of theories, models, guidelines, and best practices (see for example the chapter about game design in this book). This is intentionally not addressed in the present framework, but included implicitly on a high and abstract level because the framework is intended to model the whole process of designing and evaluating a serious game.

Preparation Phase. The preparation phase is in the beginning of every serious game project and starts with stating the problem which should be solved. The reason for investing effort and resources into the development of a serious game is that a problem exists which is to be remedied by the serious game. This might concern the society as a whole, or it might only concern a particular group of people, for example patients with a particular illness, students of a specific

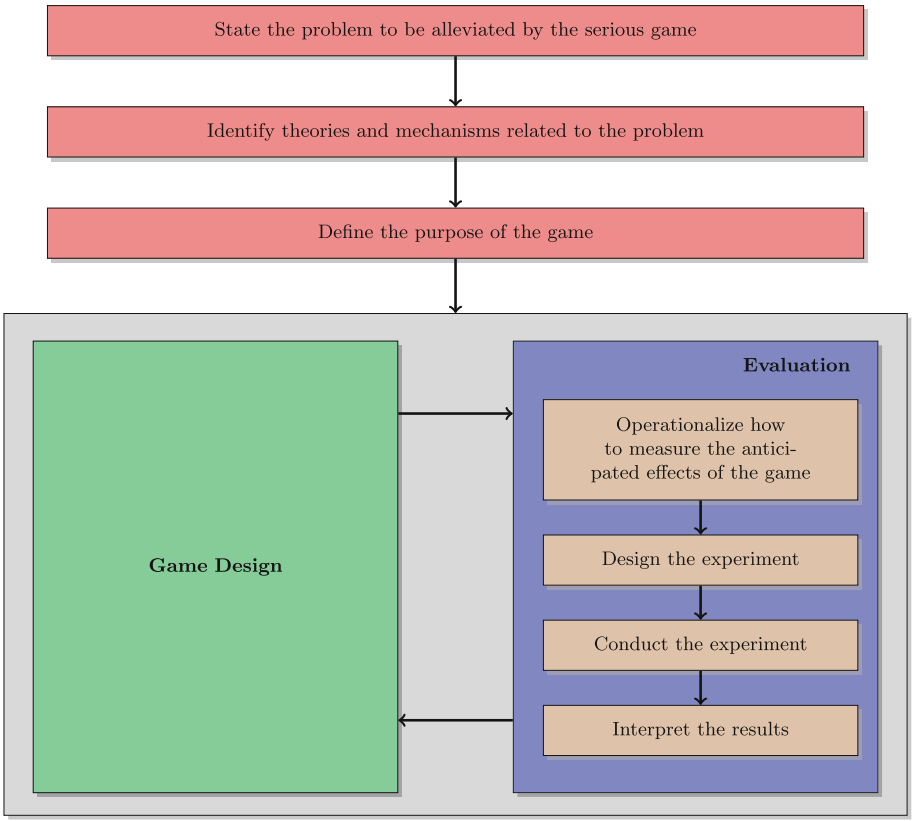

Fig. 2. The proposed serious game evaluation framework for evaluation-driven design.

grade, elderly people, etc. Such aspects of the current situation which should be improved by applying the serious game are for example:

- the percentage of the population which suffers from obesity is too high (and should be reduced)
- the awareness in society about human exploitation in third world countries is too low (and should be increased)
- the knowledge of cancer patients about their disease is too little (and should be increased)

In order to contribute to an improvement of the situation, the next step is to identify the theory behind the problem which is necessary to tackle the problem by a game. This means that we need to identify the reasons why the problem exists at all, the underlying processes or mechanisms, which factors contribute to the problem, etc. Furthermore, it needs to be analysed which of these factors can be changed by a game at all. For the example of obesity, possible reasons might be

an insufficient amount of physical activity or poor nutrition habits. But also the influence of lacking motivation, the social environment, or genetic predisposition might be considered. There is no doubt that in most cases it is impossible to identify all contributing factors and their relations to each other, but having identified at least one which is then addressed in the serious game is essential for the development and also for the evaluation of the game. Furthermore, the context of the game in which it will be applied, needs to be considered. It might be helpful to answer the following questions: (i) Why do other methods which have been used before did not work sufficiently to solve the problem? (ii) What is significantly different in a game than in previously applied methods? (iii) Is there an aspect that can be supported by a serious game in coordination with a traditional method? (iv) Is there an aspect which can even be treated better by a serious game than by other means?

Otherwise, a serious game will not be more successful than traditional methods. Having analysed the conventional methods and their effects has another advantage: the effects of the conventional methods set a lower bound on the effects of the serious game in order to rate the quality (or appropriateness) of the serious game: The measured effects of the serious game should be significantly larger than the traditional methods.

The next step is the definition of the purpose of the serious game. Based on the factors which contribute to the stated problem, the purpose of the serious game can be defined. The purpose of a serious game is to influence the identified factor which again has an impact on the situation and might solve the stated problem. The defined purpose is later the criterion by which the serious game is evaluated. Therefore, it is important that the purpose of the game can be operationalized, otherwise a proper evaluation is hardly possible. If it cannot be measured whether the serious game fulfils its designated purpose because the purpose can not be operationalized and measured, it is questionable whether this game should be called a serious game in the stricter sense.

Iterative Process of Games Design and Evaluation. The two other components, game design and game evaluation, build on the preparation phase and are tightly linked to each other. This means that the game is designed in an incremental way such that already at early stages of development, a prototype can be used for testing. Results from such an early evaluation will be taken into account in the next design step which results in another prototype which can be evaluated, etc. Hence, the results of an evaluation phase need to be usable in the game design in the sense that the feedback from the evaluation allows to draw conclusions which can be used to improve the game. The evaluation phase itself contains three main steps and is highly dependent on the results of the preparation phase. Depending on the defined purpose of the game, it needs to be analysed how the desired effect of the game can be measured in order to verify that the game fulfils its designated purpose. This is a great challenge in the evaluation procedure. This is the reason why a full chapter of this book is dedicated to the question of how to operationalize anticipated effects and

elaborates on the existing methods of measuring concepts as motivation, fun, learning effects, behavioral changes, etc. However, the step of operationalization is facilitated by a carefully and thoroughly performed preparation phase. The next step is then the design of an experiment which tests whether the game has the intended impact on the players. The design of an experiment should follow general scientific standards and contains a great amount of challenges itself for example the treatment of the control group or the recruitment of participants. Therefore, experimental design is discussed in another chapter in this book. The interpretation of the experiment results should be done carefully and with the appropriate methods. The results from the evaluation cycle then serve as input for the design process in order to adapt the chosen design. In the ideal case, this incremental, iterative and integrated design and evaluation process allows to develop a serious game which can be shown to have a significant impact on the players. However, this framework only focuses on the impact of a game related to the defined purpose of the game because the designated purpose is the quality which distinguishes serious games from other games. It should nevertheless taken into account that the serious game is also a good game which is fun to play, engages the player, etc.

Summarizing, the proposal of the present framework is mainly intended to highlight the following recommendations for evaluation: (i) evaluation should be an integral element of the development of serious games and be present in all stages of development and should be even considered before the game design starts, (ii) evaluation and design should be two processes in the game development which benefit and are dependent of each other, (iii) the development of serious games should follow the general process of scientific working, and (iv) the evaluation and development of serious games should be centered on the intended purpose of the game—therefore, a clearly defined purpose is a necessary prerequisite for evaluation.

4.2 Further Evaluation Frameworks and Models

The previously described framework of evaluation-driven design integrates evaluation and design into a process model, other existing frameworks or models of evaluation address other aspects of evaluation. In general, all existing models can be categorized along two dimensions: the time point of evaluation (when to evaluate), and the content of evaluation (what to evaluate).

When to evaluate. As pointed out in Sect. 3, timing is one of the challenges in the process of evaluation: when in the development and deployment should the evaluation phase(s) take place? In literature, there is usually the distinction between summative and formative evaluation. Formative evaluation takes place during the development of the game and is supposed to yield results which can be incorporated in the further development, while summative evaluation is carried out after the development phase and assesses the quality of the end product and its best use. Still, time-dependent effects—measuring long-term effects or short-term effects for example— need to be considered. Therefore, there are several

models which are concerned with the time point of evaluation and integrate the phase of evaluation in the design process. These models and frameworks can be found in the chapter about game design in this book, of particular interest are the classic *ADDIE* model [24] and the *Simulation-Games Instructional Systems Design Model* [17]. Both models contain at least one cycle meaning that several development stages are (intended to be) iterated which emphasizes that game development is not a sequential process, but rather each development stage is repeated several times such that the game improves in every iteration.

What to evaluate There exist many models that are concerned with the content of evaluation from which two are presented in the following: the model of Kirkpatrick [18], and the framework of Mitgutsch and Alvarado [23]. These are not redundant, but both valuable additional tools for the evaluation of serious games and address different challenges of evaluation.

Kirkpatrick's Model of Four Levels of Evaluation. Particularly for the evaluation of serious games, it is a difference whether only the quality of the game is evaluated or whether its effectiveness regarding its purpose is evaluated. This dimension might be summarized by the question of *what* exactly is intended to be measured, and can be quantified by Kirkpatrick's model of *four levels of evaluation* [18]. He developed his model for the evaluation of training programs in companies, but it is also applicable for the evaluation of serious games. In his model, evaluation can take place on the level of reaction, of learning, of behaviour, or of results.

The lowest level, the level of reactions, assesses whether a participant liked the training and therefore, applied to the area of serious games, measures the player's satisfaction with the game experience and evaluates the quality of the game mechanics, graphics, etc. However, a positive result in the evaluation on the reaction level does not necessarily imply that the desired outcome of the game has occurred. Therefore, the next level—the level of learning—tests to which extent participants "change attitudes, improve knowledge, and/or increase skill as a result of attending the program" [18]. Hence, evaluation on the second level assesses the short-time effectiveness of the game. Yet, having understood the pure contents of the game and being able to apply them in the context of the game does not mean that the players are able to transfer the acquired attitudes, knowledge, or skills in their normal environment. This is assessed on the level of behavior which evaluates "the degree to which learners have changed their behaviour outside of the learning environment because of their participation in the learning activities." [18]. Also for serious games, evaluation of the third level measures to which extent the players can apply the acquired skills outside of the game environment or to which extent the achieved change of attitude or condition are preserved in the long run. The last level—the level of results—measures the actual effect the participation in the training/game has in the larger context, i.e. in the institution or organization where it is deployed. For the field of serious games, evaluation on this level measures the long-term impact of the game on

the whole target group. It is clear that the evaluation becomes harder to realize with increasing level, but the results become more interesting.

Kirkpatrick's model of *four levels of evaluation* is a valuable tool for planning and designing an evaluation process. In order to choose appropriate methods for the evaluation, it is necessary to know on which level the evaluation is supposed to take place. For each level, other tools and methods are appropriate. This also holds for the challenges of time-dependent effects and the reach of effects described in Sect. 3.

Serious Game Design Assessment Framework. A totally different approach to an evaluation model is presented by Mitgutsch and Alvarado with their *Serious Game Design Assessment Framework* [23]. They emphasize that evaluation of a serious game should always be with respect to a clearly articulated purpose which is also pointed out in Sects. 2 and 3.

They notice that serious games are often assessed in terms of quality of their content, and not in terms of their intention-based design. In order to allow a structured discussion about the different elements of game design and to assess their cohesiveness as well as their coherence in relation to the purpose of the game, Mitgutsch and Alvarado propose a *Serious Game Design Assessment Framework* (see Fig. 3). It identifies six essential components of serious game design among which the purpose of the game should be the driving force for the design of the remaining elements. Besides the purpose of the game, the elements of the framework are *(i)* content & information, *(ii)* game mechanics, *(iii)* fiction & narrative, *(iv)* aesthetics & graphics, and *(v)* framing, of which all should reflect the purpose of the game. How these elements are related to each other has

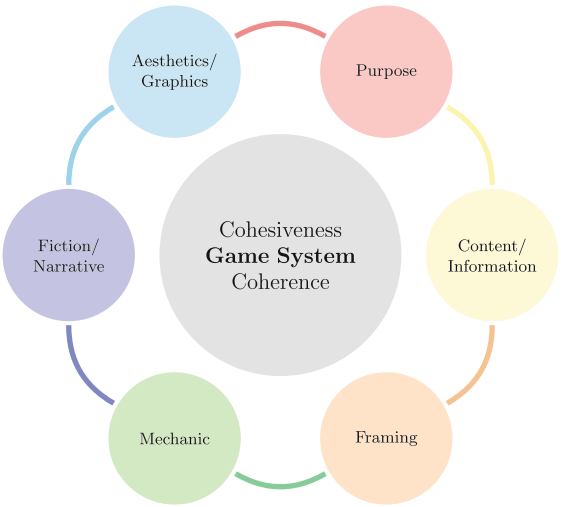

Fig. 3. Serious game design assessment framework, proposed by Mitgutsch and Alvarado [23].

an impact on the coherence and the cohesiveness of the game which is important for the game play experience since the game is perceived as one system by the player and not as single components. Their main contribution is the development of a questionnaire by which a given serious game can be analysed with respect to the identified components, their relation to each other, and their connection to the purpose of the game.

With this questionnaire at hand, a systematic and purpose-driven evaluation of a given serious game is possible. The authors deliver a well-thought decomposition of the elements of game design with a particular focus on the intended impact of the game which is essential for the design of serious games. However, this method of evaluation is not bale to measure whether the serious game is effective with regard to its designated purpose, but it might serve as a base for a discussion—as the authors also state it—and is raising attention to the fact that the intended impact should be in the focus of the design process of a serious game.

5 Examples of Commendable Serious Games Evaluation

After having introduced a general framework for serious games evaluation, this section presents practical examples of commendably evaluated serious games and discusses them in terms of the aforementioned framework. Though evaluation is a substantial part in the serious games development process, there are to date only few reported studies about full-fledged serious games that have successfully been evaluated in structured and comprehensive evaluation processes. Among these are the games *Re-Mission*, *SnowWorld* and *Frequency 1550*.

5.1 Re-Mission

One well-known example is the game *Re-Mission* by *HopeLab* for children and adolescents suffering from cancer [30]. In this third-person shooter-like game, the player controls a little nanobot inside the human body and uses different weapons related to cancer treatment, e.g. chemotherapy and radiotherapy, to destroy cancer cells and manage treatment side effects. The overall purpose of *Re-Mission* is to increase patients' well-being by inducing health-related behavioral change: The basic idea is to achieve a better treatment compliance due to an enhancement in knowledge and understanding regarding the disease and its treatment. Furthermore, the act of fighting against cancer in a digital environment was supposed to increase the patients' beliefs that they are able to influence their recovery and thus their feeling of cancer-specific self-efficacy. Hence, the purpose of the game was explicitly defined based on the problem of non-compliance and suffering of patients, and with respect to psychological determinants of behavior as well as related theories [30]. This is in accord with the preparation phase of the proposed serious games evaluation process.

Accordingly, the effectiveness of *Re-Mission* was then tested in a broadly conceived study [16] with 375 young cancer patients in the United States, Canada,

and Australia, focusing on health-related behavior changes, but additionally also evaluating fun and success as a digital game in general. Researchers faced the challenge of assessing a very vulnerable target group by extending the duration and draw area of the experiment in order to obtain a reasonable sample size. The long-term formal experiment was conducted for over two years and it included two evaluation conditions: One group of patients played *Re-Mission* over a period of three months, while the control group was given a comparable but not cancer-related digital game (*Indiana Jones and the Emperor's Tomb*), which is based on similar game mechanics and perspective. While this control group condition was a good choice to prove whether the gaming experience as such has an effect on compliance, another control group would have been beneficial in order to compare the game against alternative intervention methods such as a knowledge-providing text.

The focus of outcome measurements was on medication adherence, self-efficacy, cancer-related knowledge, feelings of control and stress level. Those variables were operationalized in several ways and assessed at three time points (baseline, after one month, after three months) by various measures ranging from subjective self-reports (questionnaires and scales) to objective control mechanics regarding adherence (blood tests and electronic pill-monitoring devices) and time spent playing the game (data logging). This multidimensional measurement approach combined with the long-term character of the experiment allowed for a comprehensive data analysis leading to informed results regarding the impact and appeal of *Re-Mission*. Consequently, besides just confirming the intended positive effect on medical adherence and knowledge, researchers were also able to gain valuable insights into the processes by which the game influences players [16,30]: Contrary to prior assumptions, the observed behavioral changes are mainly associated with an increase of self-efficacy, and can hardly be explained by knowledge acquisition and experiences made in the game world. Hence, motivational and emotional components were more influential than expected and should be considered intensely in future work on serious games for behavioral change.

5.2 SnowWorld

Another interesting example of a comprehensively evaluated serious game is *SnowWorld* developed by Hunter Hoffman and Dave Patterson at the University of Washington in cooperation with *Firsthand Technology*, a company focused on serious games and virtual environments[1]. In *SnowWorld*, the player becomes immersed in a virtual reality (VR) setting of an icy canyon using a head-mounted display and earphones. By using VR technology, the distraction from everything happening in the real world is high. The users are enveloped by the cold and chilly atmosphere inside the gameworld and are able to interact with it by navigating and throwing snowballs. This effect is the fundamental serious' mechanic

[1] http://www.firsthand.com/services/pain.html.

of the game, as it's purpose is to distract burn victims from painful procedures like daily wound caring and physical therapy in order to reduce feelings of pain. Thus, the game addresses the problem of severe pain during treatment and uses theories of distraction and attention as well as knowledge about brain functions related to pain in order to serve the purpose of pain relief. Based on these thorough considerations beforehand, the effectiveness of *SnowWorld* could be confirmed in a wide variety of studies with different focus groups, including pediatric and adult burn patients, e.g. [8,13,28], military patients with combat-related burn injuries, e.g. [20], as well as healthy volunteers who agreed to pain stimulating procedures [14,31]. In all studies, the use of the VR system and *SnowWorld* significantly reduced pain during painful treatment (up to 41 % of subjective pain relief and 50 % or greater reductions in pain-related brain activity), thereby strongly confirming its effectiveness both regarding subjective feelings of pain and objective assessments of pain-related brain activity (see [12] for a summarizing overview). Moreover, the development and evaluation process of *SnowWorld* is especially interesting in the context of serious games evaluation, as it demonstrates the advantages of several evaluation cycles. Due to repeated phases of (re-)developing and testing instead of conducting one big experiment at the end of the development process, the design of the game has been optimized and the underlying processes that lead to pain relief have been revealed to a great extent. The general system design allowed for several variations of different game and system parameters, which was used by researchers to test the influence of single design elements such as the level of presence and interactivity [31] and the VR display quality [14] (which both turned out to be important impact factors influencing pain relief). The wide variety of studies with different focal points concerning *SnowWorld* provides great insights into the functionality of the game and underlines the value of an iterative evaluation process as suggested by our framework.

5.3 Frequency 1550

The proper evaluation of serious games for health like *Re-Mission* and *SnowWorld* is especially important as they are supposed to positively influence the players' health and well-being without harming sensitive target groups, and this promise has to be proven. However, evaluation is also relevant in other application areas. One example of a serious game for learning which was evaluated in a comprehensive study is *Frequency 1550* by the *Waag Society*[2]. *Frequency 1550* is a location-based mobile educational game about the medieval city of Amsterdam designed for pupils in secondary school [15]. Its purpose is to convey knowledge about the history of places in an interactive way and to overcome a possible lack of motivation of pupils. Although the game had already won an award as the world's most innovative e-learning application due to it's design, researchers were aware of the fact that its effectiveness was still to be proven, hence they conducted a study with 458 pupils in Amsterdam [15]. Half of the participants

[2] http://freq1550.waag.org/.

played the game while exploring the city of Amsterdam during one school day while the other half learned the same educational content in a project-based lesson series in the classroom without the game, forming a reasonable control group. Measurement included engagement, motivation for history in general and the topic of Middle Ages in particular as well as knowledge of medieval Amsterdam. Results of the knowledge test that was conducted after the lessons demonstrate that *Frequency 1550* significantly increased learning outcomes compared to the control condition (with about 24 % more questions answered correctly), while topic-related motivation did not differ. Furthermore, the examination of possible influencing variables revealed that the prior knowledge level and level of education did moderate the effect. Hence, researchers were able to show an advantage of their game regarding learning outcomes compared to a non-gaming alternative, although long-term effects remain unclear.

In all those examples, interdisciplinary teams managed to successfully develop serious games based on grounded theory and to finally prove their efficacy in a proper evaluation process. However, many other serious games lack evidence of success due to minor or inconclusive evaluation activities or have not even been evaluated at all [6,9,11]. Moreover, in general there are still many open questions left and the empirical basis regarding the success of serious games is still far from being conclusive. In order to strengthen the trust in serious games and to help the industry grow, more studies are needed following a comprehensive evaluation approach. Therefore we introduced our framework of evaluation-driven design which offers guidance for future research activities.

6 Conclusion

This chapter gave an introduction into the evaluation of serious games focused on general evaluation procedures and models. We outlined the importance of evaluation during and after the design process and highlighted characteristic challenges that have to be met when the effectiveness of a serious game is to be tested. The framework of evaluation-driven design presented here comprises theories as well as practical experiences regarding the evaluation of serious games and is thus supposed to give guidance for all those who want to design, develop or research serious games, as evaluation is important to all these areas. The framework describes the important steps of the design and evaluation process and thereby highlights aspects that have to be considered like the definition and operationalization of the game's purpose as well as the iterative cycle between design process and evaluation.

While this chapter is supposed to introduce the evaluation process in general and to point out what has to be taken into account while planning and conducting a serious game study, certain aspects are just broached without going into detail. After gaining insight into the evaluation process, we recommend to continue with the following chapters about operationalization and experimental design to deepen the knowledge about serious games evaluation. Those chapters may answer questions that remain unanswered here and go into more detail

regarding single steps of the evaluation process. Moreover, we compiled the following list of further reading on related topics, that may be worth taking a closer look.

References

1. Baranowski, T.: Measurement method bias in games for health research. Games Health J. **3**(4), 193–194 (2014)
2. Chen, S., Michael, D.: Proof of learning: assessment in serious games (2005). http://www.gamasutra.com/features/20051019/chen_01.shtml
3. Chin, J., Dukes, R., Gamson, W.: Assessment in simulation and gaming: a review of the last 40 years. Simul. Gaming **40**(4), 553–568 (2009)
4. Coleman, J.S., Livingston, S.A., Fennessey, G.M., Edwards, K.J., Kidder, S.J.: The hopkins games program: conclusions from seven years of research. Educ. Res. **2**(8), 3–7 (1973)
5. Connolly, T., Stansfield, M., Hainey, T.: Towards the development of a games-based learning evaluation framework. In: Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces: Techniques and Effective Practices. IGI Global, Hershey (2009)
6. Connolly, T.M., Boyle, E.A., MacArthur, E., Hainey, T., Boyle, J.M.: A systematic literature review of empirical evidence on computer games and serious games. Comput. Educ. **59**(2), 661–686 (2012)
7. DeSmet, A., van Ryckeghem, D., Compernolle, S., Baranowski, T., Thompson, D., Crombez, G., Poels, K., van Lippevelde, W., Bastiaensens, S., van Cleemput, K., Vandebosch, H., de Bourdeaudhuij, I.: A meta-analysis of serious digital games for healthy lifestyle promotion. Prev. Med. **69**, 95–107 (2014)
8. Faber, A.W., Patterson, D.R., Bremer, M.: Repeated use of immersive virtual reality therapy to control pain during wound dressing changes in pediatric and adult burn patients. J. Burn Care Res. **34**(5), 563–568 (2013)
9. Girard, C., Ecalle, J., Magnan, A.: Serious games as new educational tools: how effective are they? a meta-analysis of recent studies. J. Comput. Assist. Learn. **29**(3), 207–219 (2013)
10. Göbel, S., Gutjahr, M., Hardy, S.: Evaluation of serious games. In: Bredl, K., Bösche, W. (eds.) Serious Games and Virtual Worlds in Education, Professional Development, and Healthcare, pp. 105–115. IGI Global, Hershey (2013)
11. Hainey, T., Connolly, T.: Evaluating games-based learning. Int. J. Virtual Per. Learn. Environ. (IJVPLE) **1**(1), 57–71 (2010)
12. Hoffman, H.G., Chambers, G.T., Meyer, W.J., Arceneaux, L.L., Russell, W.J., Seibel, E.J., Richards, T.L., Sharar, S.R., Patterson, D.R.: Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures. Ann. Behav. Med. **41**(2), 183–191 (2011)
13. Hoffman, H.G., Patterson, D.R., Seibel, E., Soltani, M., Jewett-Leahy, L., Sharar, S.R.: Virtual reality pain control during burn wound debridement in the hydrotank. Clin. J. Pain **24**(4), 299–304 (2008)
14. Hoffman, H.G., Seibel, E.J., Richards, T.L., Furness, T.A., Patterson, D.R., Sharar, S.R.: Virtual reality helmet display quality influences the magnitude of virtual reality analgesia. J. Pain **7**(11), 843–850 (2006)
15. Huizenga, J., Admiraal, W., Akkerman, S., Dam, G.T.: Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. J. Comput. Assist. Learn. **25**(4), 332–344 (2009)

16. Kato, P.M., Cole, S.W., Bradlyn, A.S., Pollock, B.H.: A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics **122**(2), e305–e317 (2008)
17. Kirkley, S.E., Tomblin, S., Kirkley, J.: Instructional design authoring support for the development of serious games and mixed reality training. In: Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) (2005)
18. Kirkpatrick, D.L., Kirkpatrick, J.D.: Evaluating Training Programs: The Four Levels, 3rd edn. Berrett-Koehler Publishers, San Francisco (2006)
19. Loh, C.S., Sheng, Y., Ifenthaler, D.: Serious games analytics: theoretical framework. In: Loh, C.S., Sheng, Y., Ifenthaler, D. (eds.) Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement. Advances in Game-Based Learning, pp. 3–29. Springer International Publishing, Cham (2015)
20. Maani, C.V., Hoffman, H.G., Morrow, M., Maiers, A., Gaylord, K., McGhee, L.L., DeSocio, P.A.: Virtual reality pain control during burn wound debridement of combat-related burn injuries using robot-like arm mounted VR goggles. J. Trauma: Injury, Infection Crit. Care **71**(supplement), 125–130 (2011)
21. Mayer, I.: Towards a comprehensive methodology for the research and evaluation of serious games. Procedia Comput. Sci. **15**, 233–247 (2012)
22. Michael, D.R., Chen, S.: Serious Games: Games That Educate, Train and Inform. Course Technology PTR, Boston (2005)
23. Mitgutsch, K., Alvarado, N.: Purposeful by design?: a serious game design assessment framework. In: Proceedings of the International Conference on the Foundations of Digital Games, FDG 2012, New York, NY, USA, pp. 121–128. ACM (2012)
24. Molenda, M.: In search of the elusive ADDIE model. Perform. Improv. **42**(5), 34–36 (2003)
25. O'Neil, H.F., Wainess, R., Baker, E.L.: Classification of learning outcomes: evidence from the computer games literature. Curriculum J. **16**(4), 455–474 (2005)
26. Papastergiou, M.: Exploring the potential of computer and video games for health and physical education: a literature review. Comput. Educ. **53**(3), 603–622 (2009)
27. Sanchez, A., Cannon-Bowers, J.A., Bowers, C.: Establishing a science of game based learning. In: Sanchez, A., Cannon-Bowers, J.A., Bowers, C. (eds.) Serious Game Design and Development: Technologies for Training and Learning, pp. 290–304. IGI Global, Hershey (2010)
28. Schmitt, Y.S., Hoffman, H.G., Blough, D.K., Patterson, D.R., Jensen, M.P., Soltani, M., Carrougher, G.J., Nakamura, D., Sharar, S.R.: A randomized, controlled trial of immersive virtual reality analgesia, during physical therapy for pediatric burn injuries. Burns: J. Int. Soc. Burn Injuries **37**(1), 61–68 (2011)
29. Shapiro, M.A., Peña, J.: Generalizability and validity in digital game research. In: Ritterfeld, U., Cody, M.J., Vorderer, P. (eds.) Serious Games: Mechanisms and Effects, pp. 389–403. Routledge, New York (2009)
30. Tate, R., Haritatos, J., Cole, S.: HopeLab's approach to re-mission. Int. J. Learn. Media **1**(1), 29–35 (2009)
31. Wender, R., Hoffman, H.G., Hunner, H.H., Seibel, E.J., Patterson, D.R., Sharar, S.R.: Interactivity influences the magnitude of virtual reality analgesia. J. Cyber Ther. Rehabil. **2**(1), 27–33 (2009)

Further Reading

32. Abt, C.C.: Serious Games. University Press of America, Lanham (1987)
33. Ainsworth, S.: Evaluation methods for learning environments. In: A Tutorial for the 11th International Conference on Artificial Intelligence Education, Amsterdam. www.psychology.nottingham.ac.uk/staff/Shaaron.Ainsworth/Evaluationtutorial.ppt
34. All, A., Castellar, E.P.N., Van Looy, J.: Towards a conceptual framework for assessing the effectiveness of digital game-based learning. Comput. Educ. **88**, 29–37 (2015)
35. Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., Berta, R.: Assessment in and of serious games: an overview. Adv. Hum. Comput. Interact. **2013** (2013). Article no. 1
36. Bernhaupt, R. (ed.): Evaluating User Experience in Games: Concepts and Methods. Human-Computer Interaction Series. Springer, London and New York (2010)
37. Bredl, K., Bösche, W. (eds.): Serious Games and Virtual Worlds in Education. Professional Development, and Healthcare. Premier reference source, Information Science Reference, Hershey (2013)
38. Dondi, C., Moretti, M.: A methodological proposal for learning games selection and quality assessment. Br. J. Educ. Technol. **38**(3), 502–512 (2007). http://dx.doi.org/10.1111/j.1467-8535.2007.00713.x
39. El-Nasr, M.S., Drachen, A., Canossa, A. (eds.): Game Analytics: Maximizing the Value of Player Data. Springer, London and New York (2013)
40. Göbel, S., Gutjahr, M., Steinmetz, R.: What makes a good serious game - conceptual approach towards a metadata format for the description and evaluation of serious games. In: Gouscous, D., Meimaris, M. (eds.) 5th European Conference on Games Based Learning, pp. 202–210. Academic Conferences Limited, Reading (2011)
41. Law, E.L.-C., Kickmeier-Rust, M.D., Albert, D., Holzinger, A.: Challenges in the development and evaluation of immersive digital educational games. In: Holzinger, A. (ed.) USAB 2008. LNCS, pp. 19–30. Springer, Heidelberg (2008). doi:10.1007/978-3-540-89350-9_2
42. Loh, C.S., Sheng, Y., Ifenthaler, D.: Serious Games Analytics: Methodologiesfor Performance Measurement, Assessment, and Improvement. Advances in Game-Based Learning. Springer International Publishing, Cham (2015)
43. Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., Lo, J., Kortmann, R., Wenzler, I.: The research and evaluation of serious games: toward a comprehensive methodology. Br. J. Educ. Technol. **45**(3), 502–527 (2014)
44. Schultz, C.P., Bryant, R.: Game Testing All in One, 2nd edn. Mercury Learning & Information, Dulles (2011)
45. Serrano, Á., Marchiori, E.J., Blanco, Á.D., Torrente, J., Fernández-Manjón, B.: A framework to improve evaluation in educational games. In: Global Engineering Education Conference (EDUCON), pp. 1–8. IEEE (2012)