Operationalization and Measurement of Evaluation Constructs

Katharina Emmerich[1(✉)], Natalya Bogacheva[2], Mareike Bockholt[3], and Viktor Wendel[4]

[1] Entertainment Computing Group, University of Duisburg-Essen, Duisburg, Germany
katharina.emmerich@uni-due.de
[2] Moscow State University, Moscow, Russia
bogacheva.nataly@gmail.com
[3] Graph Theory and Complex Network Analysis Group, TU Kaiserslautern, Kaiserslautern, Germany
mareike.bockholt@cs.uni-kl.de
[4] Multimedia Communications Lab, TU Darmstadt, Darmstadt, Germany
viktor.wendel@kom.tu-darmstadt.de

Abstract. This chapter deals with the operationalization and measurement of evaluation constructs, an important and challenging part of the serious games evaluation process. Hereby, advices will be given on what has to be measured and how to quantify an abstract concept. Thus, the chapter makes two main contributions. First, general data gathering methods are described and discussed in terms of advantages and disadvantages. Second, main psychological concepts and evaluation constructs relevant in the context of serious games, as well as their theoretical foundations, are introduced. In order to support the reader on planning future serious game evaluations, a list and description of concrete techniques and questionnaires addressing concepts like motivation, player experience, learning outcomes, health, well-being, and attitudes are compiled.

Keywords: Operationalization · Measurement methods · Player experience · Psychological constructs · Evaluation

1 Introduction: What Is Operationalization?

This chapter covers an important part of the serious games evaluation process, namely the operationalization of evaluation constructs. Operationalization is a substantial aspect of quantitative research and generally referred to as the process of defining how to quantify a phenomenon or concept which itself is not directly measurable. This applies to basic psychological constructs as for example motivation, well-being, and emotions, but also to seemingly more concrete concepts like health, intelligence, or learning progress. Though we all have some kind of understanding of all these constructs, we cannot tell a direct and distinct

R. Dörner et al. (Eds.): Entertainment Computing and Serious Games, LNCS 9970, pp. 306–331, 2016.
DOI: 10.1007/978-3-319-46152-6_13

way to quantitatively measure them. This is a great issue regarding the evaluation of serious games: Serious games are supposed to serve a certain purpose besides mere entertainment, but this purpose is mostly described in terms of such rather fuzzy concepts. This is where operationalization comes in.

Operationalization is based on the assumption that an abstract construct under examination can be inferred from its observable effects. For instance, if a serious game is supposed to address the player's attitude towards junk food, we cannot directly measure this attitude after a playing session, because an attitude itself is neither observable nor directly graspable by existing measuring tools. However, there are several indicators that are supposed to be closely related to attitudes and thus allow for drawing conclusions. Regarding the given example, players may be asked some questions about their opinion on junk food ("To what extent do you agree that junk food is unhealthy?"). In this case, the questions constitute the operational definition of a person's attitude towards junk food. But asking questions is not the only possibility to operationalize it. Attitudes are often also reflected by behavior. Hence, it could also be observed whether (and how often) a person consumes junk food in a certain period of time after playing the game. The number of times someone eats junk food in one month is another operationalization of one's attitude towards it.

1.1 The Challenge of Operationalization

As the example of attitude towards junk food illustrates, the process of operationalization is complex and ambiguous. Hence, the biggest challenge and the goal is to ensure that the construct is measured in a way that is "as accurate a representation of the construct as possible" [44]. Landers et al. [44] describe this issue with regard to classical test theory and explain that an operational definition is composed of two parts: a true score, meaning the correct inference to the construct, and an error score, meaning the proportion of mis-measurement of the construct. The challenge of operationalization is thus to minimize error scores by finding proper measurement approaches.

One strategy to reduce misleading results is to apply multilayered measures. If, for instance, the player's perceived autonomy should be assessed, instead of using just one single item like "I felt autonomous while playing the game", error score minimization can be achieved by asking for agreement on multiple statements, for example "the game provides me with interesting options and choices" and "I experienced a lot of freedom in the game" (cf. [61]). Moreover, different measurement methods like questionnaires and observations can be combined in order to check correlations and confirm results.

However, finding appropriate operational definitions remains challenging. Therefore, this chapter is supposed to provide assistance for the process of operationalization by addressing the core question on how evaluation constructs can be quantified. In the context of digital games and serious games, there are many constructs which are repeatedly under investigation, like motivation, player experience, learning outcomes, well-being and concepts regarding attitudes and personality traits. Accordingly, researchers already made experiences with different

approaches and in some cases already developed validated measurements. Those will also be discussed in the chapter in order to gather information, exchange experiences and give advise for future operationalization processes.

1.2 Overview

The chapter is divided into three main parts. First, general data gathering methods are introduced to give an overview of available means to assess data. Related characteristics, advantages, and disadvantages are shortly discussed. Subsequently, different constructs that are relevant in the context of serious games evaluation are addressed, as it is important to understand those constructs to be able to derive concrete operationalizations. Those constructs are motivation, player experience, learning outcomes, health and well-being, and attitudes. For each construct, common ways to operationalize them are presented. Finally, the chapter concludes with a summary of concrete method implementations, mainly standardized questionnaires and psychological test procedures, that are currently used in game research to assess the concepts discussed before. This list is not exhaustive, but provides a helpful starting point for the planning and conducting of evaluation in terms of operationalization.

2 User-Centered Data Gathering Methods

The main goal of serious games evaluation is to prove whether a game fulfills its designated purpose. Serious games are applied in very different contexts and thus can serve manifold purposes. However, they are always supposed to somehow influence the players by addressing their knowledge, emotions, attitudes, or behavior. This implies that the evaluation process should reasonably focus on user-centered data gathering methods. As the name indicates, those methods put the user in the center of the evaluation process. There are several core ideas that stand behind this methodology: First, the goal is to optimize the product (the game) to make it convenient for the users. Second, the game has to be evaluated from the users point of view (and not from the developers' one) to ensure that optimization is done in the right way. To put it in the simplest possible way, the developers know how their serious game is supposed to work but the players do not. Third, to achieve points one and two data has to be gathered directly from the users to validate hypotheses, while it needs to be done in a way that will not harm or interfere the players' gaming experience.

In this part of the chapter, we discuss the general methods that can be used to gather user-centered data while developing and evaluating a serious game. While the methods are generally all the same throughout different user-centered designs, their configurations differ according to the application field and the current stage of the game development process [7].

2.1 Types of Measurement Methods

Measurement methods that are applicable to evaluate serious games are diverse. Most of them can be assigned to one of the four main categories: self-reports, physiological measures, gameplay metrics, and observations.

Self-reports. Self-report measurements directly ask the players to answer concrete questions or to freely give their opinion on certain aspects. Hence, resulting data is always subjective. Common forms are questionnaires, focus group discussions, and interviews. The last two methods mainly result in qualitative data, as they allow for free (and long) answers. Focus groups are usually used at early stages of the development and evaluation process and help to investigate the general acceptance and opinions of players on the game.

The data gathered from interviews is quite similar, though questions are more focused on the individual. Gill et al. [27] differentiate three major types of interviews: structured, semi-structured, and unstructured. A structured interview is a verbalized questionnaire with predetermined questions and usually a set of possible answers. On the opposite, unstructured interviews do not strictly follow a predefined plan, but are adapted to the situation and previous answers. These interviews can last unlimited amount of time and, although they do not require much organization, the do require a lot of skill from the interviewer to make results useful. Semi-structured interviews are more common in most sciences and have the advantages of both other types. They do have a structure (a number of key questions, outlining important areas of the research), but they are also flexible enough to discover some essentially new data if the dialog goes that way. Semi-structured interviews also require more skill than fully structured ones, as well as more time and general efforts.

Standardized questionnaires, in contrast, consist of carefully selected items and provide quantitative data related to an abstract construct, and are thus most interesting in the context of operationalization. In general, a questionnaire is an instrument to quantify player-related constructs like feelings and thoughts. Unlike interviews, they are usually used if data from numerous participants has to be gathered with relatively small efforts. As it can be extracted from the name, questionnaires are made of questions or statements that participants are required to agree or to disagree with. The questions can be either closed or open-ended. Closed questions can be answered by "yes" or "no" or by choosing a response from several given alternatives. Open-ended questions require broader answers. The participants need to add something to the answers or create it by themselves. Open-ended questions can be more informative in some way, and in some studies they are shown to provide more valid data [43]. However, they are much more difficult to analyze and to be used in statistical calculations, as answers are not standardized.

Many questionnaires, especially those that use statements as their items, use interval scales (or Likert scales) to evaluate participant agreement or disagreement with the statement. Scales usually consist of an odd number of points (to have a neutral answer in the middle) [30], ranging in meaning from "Strongly

Disagree" to "Strongly Agree". Likert scales usually have from five to nine points; longer scales give the participants better opportunity to specify their attitude to the statement, but too many variants can be confusing. The same rule can be applied to the number of answers of closed questions – there should be enough answers to cover all possible or at least common alternatives, but not too many to make participants insecure.

Depending upon what data should be assessed, either already existing questionnaires made by some other scientists can be used, or an own questionnaire with a unique set of questions has to be developed. Once a questionnaire is compiled and proven to be valid and reliable, it is rather easy and convenient to use, but the development of a new questionnaire is difficult and complex. Unlike an interview, which allows to control the data gathering process and to change the pace and the manner to get information, the questionnaire is used as it is. Hence, it has to be easy to understand and well-conceived. One of the most important problems while creating a questionnaire is the actual wording. You cannot explain or reframe the questions during use, if the participant does not understand it, so the question or the statement must be as simple as possible. Characteristics of the target group have to be considered before creating the questionnaire: If the users of the serious game are going to be children or teenagers, probably a different language should be used compared to older participants. Moreover, if the question is quite simple for a psychologist or computer-science specialist, it is not necessarily understandable for a common user with no scientific background. This is the reason why pilot studies are needed before using new questionnaires on large groups of people. Goodwin et al. [30] give seven guidelines on phrasing good questionnaire items: They advise to prefer simplicity over complexity, to use complete sentences instead of some short phrase, to avoid any abbreviations that are not completely universal, to avoid slang and jargon, to avoid negatively phrased questions, and to make questions as balanced as possible, without favoring one position and without giving the participant any clues on the desired answer. This will help avoiding bias and making the results more objective.

Furthermore, questionnaires for evaluation purposes should be based on theories and conceptual frameworks regarding the construct under investigation (like the ones discussed in the second part of this chapter) in order to make proper interpretation possible. Most of the questionnaires are designed to measure more than one characteristic or parameter of a certain construct. Single scale questionnaires are also possible, but rare in psychological and related research because the reality that stands behind complex behaviors or attitudes is usually too complex to be described by a single latent variable.

Physiological Measures and Biometrics. Another approach to assess player data is to measure physical reactions of the player while playing the game. Based on the assumption that the processing of any stimuli (so also during play) always provokes physical reactions, those so-called biometrics [53] are used to draw conclusions about psychological phenomena [40]. This data is objective in contrast

to self-report measures, but psychophysiological data is sensitive to noise (for instance, facial EMG measurement can be confounded by related muscle activity such as speaking) and mostly ambiguous with regard to interpretation: Most relations between certain psychological states and physiological outcomes are not one-to-one but many-to-one relations, turning the interpretation of data into a challenging task [40,67]. There are various ways to assess physical reactions. The currently most common ones are shortly described in the following:

- Skin conductance: The electrodermal activity (EDA) can be assessed with sensors attached to the fingers. It is supposed to be related to arousal and stress [40].
- Cardiac activity: An electrocardiogram (ECG) or a peripheral pulse oximeter can be used to measure heart rate and pulse. These biometrics are indicators for arousal, attention, or stress [40].
- Electrical activity of facial muscles: Facial electromyography (EMG) measures facial expressions by means of the electrical activation of facial muscles. This data gives information about the valence and arousal of emotional reactions [40,53,58].
- Brain activity: Brain waves are measured by electroencephalography (EEG) and allow for deriving cognitive processes, the degree of attention and the use of mental resources [40,53].
- Respiration: The rate or depth of breathing provides an indication of relaxation, stress, or negative emotions [50].
- Eye movement: In order to investigate which elements are recognized, focused and paid attention to by the player, eye tracking systems can be used to record viewing direction and movement of gaze [53].

Gameplay Metrics. Gameplay metrics provide information about the interaction between the player and the game in terms of numerical data [53]. Compared to the other types of measurement, this method gathers data from the game system and not directly from the player. While playing, the player's in-game behavior is automatically tracked by the system and related to certain game events or locations (for instance, by time-stamps and labels). Results allow conclusions about the player experience as they offer insights into how people are actually playing the games under examination [53] (p. 50). The events and behavior that are relevant for the evaluation process differ with respect to the purpose of the game. Typical game metrics for example quantify how many times the player performs a certain action. However, this behavioral data does not inform about the reasons why players are acting the way they do. Thus, gameplay metrics are often used in combination with other methods to gain additional insights about the course of the game session and relevant events and actions. A proper visualization of gameplay data can reveal patterns of player behavior that otherwise would have been undetected. Hence, metrics are supposed to be highly relevant to serious games evaluation and can be applied in diverse contexts (see for instance [48] for further details).

Observations. Besides asking players directly, measuring their physiological reactions or assessing their in-game behavior, researchers might also gain valuable insights by simply observing players during play. In fact, observations are not that easy to conduct and to analyze properly, but nevertheless often used in games research. Body language, gestures, interactions, facial expressions as well as verbal communication are supposed to be rich data sources for evaluating player experience [50]. Especially video recording assesses many details that have to be edited and coded based on a predefined scheme, which mostly is a time-consuming process. In order to account for validity and reliability, observation plans and coding methods have to be tested and differences between observers have to be considered (inter-rater reliability). Often observation is focused on a few specific aspects in order to facilitate the process and to support the evaluation of other measures [50].

2.2 Advantages, Disadvantages and Challenges of Methods

The aforementioned general methods of data gathering for evaluating serious games all feature certain advantages, disadvantages and challenges. They mainly differ regarding aspects of objectivity, immediacy of measurement, obtrusiveness and gameplay interference, as well as effort and costs of conduction and data analysis. Table 1 gives an overview of those differences.

While data assessed by questionnaires and any other form of self-report is always more or less subjective, the other methods provide more objective data. In the case of observation, objectivity depends on the existence and quality of an observation scheme and whether game sessions are recorded or directly processed by present observers. Observer bias (different observers interpret occurring events differently) or ambiguous observation categories might harm objectivity. Gameplay metrics and physiological measures, in contrast, are hardly manipulable and thus, most objective. The advantage of objectivity

Table 1. Overview of the characteristics and differences of main measurement method types regarding objectivity, immediacy, interference and effort/cost.

	Objectivity of data	Immediacy of measurement	Interference of gameplay	Effort/Cost
Self-reports	subjective	post-hoc	no interference	low costs, easy processing
Observation	objective but prone to observer bias	immediate	possibly interfering	high effort in processing if not automated
Gameplay metrics	objective	immediate	no interference	implementation effort, low assessment costs
Psycho-physiological measures	objective	immediate	possibly interfering	high effort, huge datasets, tools needed

is that data is not blurred by personal sensitivities or social effects (like social desirability) and hence more reliable. However, objective data provides only limited insights regarding a person's thoughts, feelings and reasons to act the way that was detected. At this point, subjective data helps to find explanations and relations between observable behavior, physiological reactions, and psychological processes.

Another aspect that distinguishes self-reports from the other methods is the immediacy of the measurement. Questionnaires and interviews are obtained after the gaming session, thus participants have to rate the experience retrospectively. Hence, results may be influenced by memory effects, for instance primacy and recency effects, as the first and the last events of the session are recalled better than the rest of the experience. However, asking questions while participants play the game would interrupt the game flow and severely influence the experience. Post-hoc tests at least do not interfere with the gameplay. Observations, gameplay metrics, and physiological measures enable immediate assessment of data while players are interacting with the game. While data collection via metrics is unapparent for participants and thus not distracting, observations and physiological measures are not in any case unobtrusive. If the observation is not covert, the feeling of being observed might influence the behavior of players, and the attachment of measurement tools to the body might also have impact. That shows that the degree of interference depends on the way measures are applied in those cases.

Finally, methods differ regarding effort and costs of their implementation and analysis of data. Self-reports, especially questionnaires are rather easy to apply and provide clear scoring schemes. The effort of observations is variable: if observers are taking notes during the gameplay session, the assessment of data takes more time for researchers, while the use of video recording reduces personnel expenditure. However, the evaluation process of observational data is very time-consuming, as the whole video material has to be coded regarding predefined aspects. The implementation of gameplay metrics demands technical expertise and additional costs during the game development process, but once they are implemented the assessment is very easy. Psychophysiological measurements are most expensive and demanding. Some techniques require complex and expensive equipment and have to be applied by trained experimenters [40]. Their usage is time-consuming and the datasets provided by them are very extensive.

3 Evaluation Constructs and Their Operationalization

The core of each serious game is its purpose. A well-designed game is based on a theoretical ground related to this purpose. If the goal is to enhance the motivation to physically exercise, game mechanics and content should be designed with respect to theories of intrinsic and extrinsic motivation, self-efficacy, personal needs as well as self-determination. If, on the other hand, the game is supposed to impart knowledge, learning principles and theories of knowledge creation should be considered. Those considerations are not only relevant during

the game development process, but may also substantially influence the evaluation: Depending on the respective theories, the evaluation has to be designed with the knowledge of the corresponding theories in mind. In this context, the proper operationalizsation of theoretical constructs is one of the main challenges. Hence, this section introduces main constructs that are relevant in the context of serious games evaluation, as it is important to understand those constructs to be able to derive concrete operationalizations. It is shown how the methods described in Sect. 2 can be applied to this constructs in order to prove a serious game's effectiveness. While the first two sub-chapters deal with rather generic concepts that do apply to any kind of digital game, namely motivation and player experience, the other three refer to more specific concepts particularly relevant for the design and evaluation of serious games: learning outcomes, health and well-being, as well as attitudes.

3.1 Motivation, Player Models and Personality Traits

Often motivation is the core argument when deciding to use a serious game instead of other forms of educational tools or information material. It is commonly believed that games are highly motivating and a pleasant experience [25]. Diverse game mechanics like progress visualization, immediate feedback, rewards, and challenges constantly motivate the player to engage in the game and to master it [16,24]. Hence, the willingness of users to play games voluntarily is seen as vehicle to start the involvement of the targeted audience regarding a certain topic. A typical example is the use of digital games in schools to provide educational material in an innovative and motivating way. Furthermore, there are also some serious games that explicitly serve the purpose of fostering motivation at the core, for instance games that are supposed to enhance physical activity of players in general without referring to specific movement patterns or rehabilitation programs.

In any case, motivation is a relevant part of serious games evaluation, hence the question arises how this construct can be operationalized in order to measure it. In general, motivation is seen as a theoretical construct describing the active pursuit of an (individually) positively rated target state [59]. It can be distinguished between *intrinsic* and *extrinsic* motivation, which refers to the source of motivation. While extrinsic motivation emerges from external rewards and impact factors (for instance, if someone is paid for doing something), intrinsic motivation describes the individual, inherent willingness to do something that is rated as interesting or enjoyable [60]. Digital games are supposed to particularly trigger the latter form.

In the context of serious games evaluation, motivation can be seen from mainly two different points of view: It is both an antecedent of playing, because only persons who are motivated are supposed to decide to start (and keep) playing a certain game [21], as well as a part of the player experience itself, as game events and features can increase or decrease motivation. Moreover, the perceived motivational qualities of the game and the current motivational level of a player will probably influence the way the game is perceived and rated.

Accordingly, if the motivation of players should be measured, these two facets of player motivation have to be distinguished.

General Motivation to Play a Certain Game: Player Models. Regarding the general motivation to choose and start playing a certain digital game, diverse motives for playing have been researched and assembled into so-called *player models*, which classify players in terms of the main reasons why they play games. Popular player models are Bartle's player taxonomy [3,4], Yee's model of player motivation [73] and the BrainHex model [52]. Bartle more or less laid the foundation of player models in his early work by defining four main player categories based on the analysis of multi-user dungeon games [3]: *achievers*, *explorers*, *socializers*, and *killers*. The player types are characterized by different main interests and playing styles. For example, achievers are trying to master all game challenges and seek for success, while explorers like to immerse in the game world and explore it without focusing on the main goals. Yee [73] proposes three main categories of player motivation, namely achievement, immersion, and social. The BrainHex model differentiates further sub-categories, resulting in seven player types. However, the idea behind all those player taxonomies is the same. While such player categories are non-exclusive and simplify the construct of motivation to some degree, they provide a first understanding as to how motivation influences a player. And – even more important in terms of the topic of this chapter – they significantly help to operationalize the construct of player motivation. Based on those theories, questionnaires have been developed in which participants are asked to rate different game experiences and events, which are internally related to one or more of the defined player types. This way, it is possible to measure a player's general motivation to play digital games. Measuring player motivation in terms of player types may reveal different effects of the game on different groups of players and can thus be part of serious games evaluation.

There are also some more general methods to assess more generic player personality traits, which may also be interesting for investigating a game's effect on certain target groups. While a detailed discussion of general psychological models of personality traits is beyond the scope of this chapter, we suggest the very popular model of the Big Five personlity traits [57] as a starting point if you are interested in investigating such aspects. The Big Five is a five factor model describing five main dimensions of personality, namely *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*, which are established in psychological research and applied to diverse research fields, also to digital games. Here, too, the common form of operationalization is using a questionnaire. There are many different versions and inventories, for instance the short 10 Item Big Five Inventory (BFI-10) [57].

The Perceived Motivational Quality of a Game. As described above, besides general models of player motivation, which try to explain why players prefer certain kinds of games and show certain playing styles, another interesting

aspect regarding the construct of motivation is the actually perceived motivational quality of a game. In this context, a basic theory of motivation, namely the *self-determination theory* (SDT), has been applied to games by Ryan et al. [61]: SDT comprises main factors that are supposed to facilitate or undermine motivation, focusing especially on intrinsic motivation. Those factors are based on the psychological needs for *autonomy*, *competence*, and *relatedness*. According to the application of SDT to digital games by Ryan et al. [61], a player's intrinsic motivation is supported by

- perceived autonomy: a sense of own volition and freedom of choice;
- perceived competence: appropriate challenges and the sense of efficacy;
- perceived presence: a sense of actually being within the game world;
- intuitive controls;
- and perceived relatedness: the feeling of being connected with others.

In order to measure need satisfaction, the *Player Experience of Need Satisfaction (PENS)* questionnaire was developed [61], taking into account the aforementioned dimensions of player needs and experiences by integrating them as sub-scales. Based on the same idea, the Intrinsic Motivation Inventory (IMI) [51] can be used to assess dimensions of the intrinsic motivation related to a game. Based on grounded theory, those measures are supposed to be valid operationalizations of the motivation construct and thus offer possibilities to evaluate serious games in terms of motivational aspects.

3.2 Facets of Player Experience

Game user research is concerned with the user experience of players, often called game experience or player experience. Though there is no sole definition of player experience, the ISO standard definition of user experience allows for approximation: It defines user experience as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" (ISO Norm 9241) and thereby matches the conception of most researchers and practitioners [45]. In contrast to usability, which is focused on functionality and ease of use, the user experience is a psychological construct and primarily comprises the feelings, thoughts, and reactions of the user [6]. Accordingly, regarding digital games it can be specified as the individual, context-sensitive experience in terms of cognitions, emotions, and physical responses emerging from the interaction between a player and a game system [6]. In short, it deals with the direct effect that the game has on its players and thus is highly relevant in the context of serious games. The definition above underlines the complexity of player experience, as it comprises several concepts and research focuses from the fields of cognition, motivation, emotion and attention [14]. In order to be able to assess the player experience of a game, current research models (for instance [69]) break it down into single, more tangible sub-dimensions. The most prevalent and investigated sub-components are

- *fun*,
- *flow*,
- *immersion*,
- *presence*, and
- *social presence*

and will therefore be shortly introduced in the following.

Fun. Fun and flow are both concepts which are long-since established in game research. The evocation of fun, which means an intense and positively valenced emotion, is commonly seen as the main objective of entertainment games; it is commonly accepted that it is also a main element of serious games. However, fun is a rather fuzzy concept without clear definition and hardly to grasp or measure as such [22,35]. Therefore, researchers have established diverse theories about the emergence and manifestations of fun. For instance, Koster [42] relates fun to brain functionalities and learning processes and attributes feelings of joy to the relief resulting from mastering new patterns. Though this might be one aspect of fun, it does not account for individual preferences and differences. Lazzaro [46], in contrast, underlines the emotional facets of fun and proposed four basic types, called Four Keys, that lead to an enjoyable experience: *hard fun*, *easy fun*, *altered states* and the *people factor*. *Hard fun* arouses from meaningful challenges and a feeling of success and pride. On the other hand, *easy fun* is attributed to the investigation of interesting stimuli like a story, a game world or ambiguous details and is thus linked to surprise and curiosity. *Altered states* stands for the potential of games to arouse different emotions and to let people feel something new or different (also including the aspect of distraction). Finally, the *people factor* emphasizes that also the interaction with other players may lead to enjoyment. In a similar way, Dillon [17] connects fun to certain emotions and instincts, but goes a bit more into detail by differentiating six basic emotions and eleven related instincts (*6-11 framework*). These three representative examples of fun theories illustrate well that fun is not a single emotion but comprises several feelings as well as cognitive processes and can be induced in manifold ways. Hence, it should be considered as a multidimensional factor during evaluation.

Flow. A related concept, which is comparable to the idea of hard fun, is flow. Based on the research of Csikszentmihalyi, flow can be seen as a desirable effect in games, as it describes a mental state of total engagement [13] resulting from a proper balance of challenge and individual skills. The flow experience is characterized by an altered sense of time and a loss of self-consciousness due to a narrow focus on the game. Sweetser and Wyeth [68] transferred this general idea of flow to digital games and developed the *GameFlow model* for evaluating player enjoyment. For that purpose, they synthesized current game evaluation heuristics from the field of usability with those from the field of user experience research. They merged different heuristics (each focusing on single game

elements such as the interface, the gameplay or the mechanics) into one comprehensive model of player enjoyment. As result, they suggest eight interrelated core elements of *GameFlow*, namely

- concentration,
- challenge,
- skills,
- control,
- clear goals,
- feedback,
- immersion, and
- social interaction.

For each of those elements, they comprised a set of criteria for player enjoyment derived from game design heuristics. It is assumed that a game which supports all those criteria is most enjoyable [68].

Immersion and Presence. The constructs *immersion* and *presence* are often considered in game user research, as they describe the process of projecting one's thoughts into the game world and to immerse in the virtual environment and story. However, both terms are not uniformly defined, but often mixed up and sometimes even used synonymously [20].

Slater [65] presents an approach to disambiguate the concepts and defines *immersion* as the objective quality of an interactive system to deliver sensory cues to the human sensory system. That is to say, a system is more immersive if it occupies large parts of a person's perception. For instance, a system containing a head-mounted display and earphones shields a person's visual and auditory senses from stimuli of the real world and thus features higher immersion than an ordinary monitor display. *Presence*, in contrast, is described as the feeling of actually being in the virtual world, being part of it, interacting with it and perceiving it as real. It can be seen as a result of immersion, because high immersion is supposed to support higher feeling of presence [65].

As immersion and presence are both concepts that are not only important for games, but also in the larger context of virtual environments and tele-presence, a lot of approaches to operationalize and measure them can be found in the literature. As with the assessment of motivation and other facets of player experience, questionnaires are the commonly used methods to measure aspects related to presence and immersion. One example that has also been applied to games is the *Igroup Presence Questionnaire* (IPQ) [34]. It is used to assess the degree of presence, on four dimensions: *general presence*, *spatial presence*, *involvement* and *experienced realism*. Another popular questionnaire addressing presence is the *Presence Questionnaire (PQ)* by Witmer et al. [72]. Furthermore, it might also be interesting to assess individual differences of players regarding their interest and capability to immerse into a virtual world. For this purpose, the *Immersive Tendencies Questionnaire* (ITQ) [71] was developed, which includes questions regarding the general tendency of a person to immerse.

Social Presence. Besides the common concept of presence described above, which is mainly related to the physical gameworld, Takatalo et al. [70] argue that there is also the feeling of *social presence*, accounting for socially meaningful contexts of digital games. That social meaning does not only apply to multiplayer games, but more generally describes the feeling of being in a socially meaningful world (for instance elicited by non-player characters). The social presence construct is further divided into three components by Takatalo et al. [70]: *Social richness*, *social realism* and *co-presence*. With social richness they describe the degree to which a game is assigned certain social attributes such as being familiar or personal. Social realism refers to the similarity of in-game objects, events and behaviors with the ones the player is familiar to from the real world. Co-presence is defined as the sensation of being together and interacting with other persons (players or non-player characters) inside the game world.

Similarly, Biocca et al. [8] also elaborated on social presence and define it as the degree to which two (or more) persons are aware of each other (co-presence) and feel psychologically and behaviorally involved. Based on this concept, de Kort and colleagues [15] developed a self-report measure to assess social presence in digital games. The result is the *Social Presence in Gaming Questionnaire* (SPGQ) with three sub-scales: *Psychological Involvement – Empathy*, *Psychological Involvement – Negative Feelings*, and *Behavioral Engagement*. During the development of the questionnaire, de Kort et al. found that – in contrast to the concepts described above – co-presence should not be seen as a separate dimension, but that it instead relates to aspects of awareness and behavioral involvement, resulting in a scale the authors call behavioral engagement. Furthermore, the dimension psychological involvement was separated into empathy and negative feelings, thus accounting for positive and negative affect and feelings provoked by another social entity.

Another questionnaire that was recently developed and that tries to assess the relation between players in terms of social presence and related concepts is the *Competitive and Cooperative Presence in Gaming* scale (CCPIG) [33]. As the name indicates, it differentiates between competitive and cooperative gaming scenarios and is meant to be applied in multiplayer games. It assesses competitive social presence on the two dimensions awareness and engagement, while cooperative social presence is divided into perceived team cohesion and team involvement. Hence, if the interaction of players and their feelings related to each other have to investigated, this questionnaire is supposed to provide valuable insights and a proper operationalization of social presence.

Comprehensive Models of Player Experience. The aforementioned subdimensions of the player experience are not exclusive and do not represent all possible aspects. There is no consensus which dimensions constitute the overall player experience, but there are some comprehensive models that go even further into detail and distinguish some more dimensions. The purpose of those models is to gain a better understanding of player experience as a whole and also to make it measurable by defining sub-dimensions and concrete characteristics.

Accordingly, the two models described in the following both serve as the theoretical background for respective questionnaires that were developed to assess player experience as a comprehensive construct in experiments.

The first example is the *Presence-Involvement-Flow Framework (PIFF)* by Takatalo et al. [69]. It comprise all main aspects mentioned in different prior approaches and is built around the main constructs *presence*, *involvement* and *flow* (as its name already indicates). *Presence* is supposed to describe the intensity and extensity of the experience, and is subdivided into the concepts of physical presence and social presence, which in turn are subdivided into further subcomponents (quality of interaction, physical presence, attention, role engagement, co-presence and emotional arousal). *Involvement* is composed of the two factors importance and interest and describes the personal value (valence) and meaning (relevance) of the experience. Finally, seven subcomponents are related to the concept of flow: challenge, competence, playfulness, control, hedonic valence, impressiveness and enjoyment. Overall, this dimension is supposed to mainly describe the quality of player experience. This multidimensional model is based on literature review and a series of validating studies. The associated questionnaire is the Experimental Virtual Environment Experience Questionnaire-Game Pitk (EVEQ-GP), which is a very long questionnaire containing 180 items related to all of the sub-dimensions of the PIFF.

Another comprehensive model of player experience is the *Game Experience Model* by Poels et al. [56]. It is the result of a study based on focus group discussions and subsequent expert meetings, which were then combined with theoretical considerations. It defines nine main dimensions of player experience, namely enjoyment, flow, imaginative immersion, sensory immersion (comparable to our notion of presence), suspense, competence, negative affect, control, and social presence. The self-report scale belonging to this model of player experience is the *Game Experience Questionnaire* (GEQ) [36], which consists of a couple of questions for each of the eight dimensions *competence*, *sensory and imaginative immersion*, *flow*, *tension*, *challenge*, *negative affect* and *positive affect*. If the game under investigation is a multiplayer game or contains social entities, the GEQ is often combined with the SPQG (cf. section about social presence above). During the development of this questionnaire, the nine primary categories were further adapted: control is no dimension on its own anymore, and was integrated into the remaining ones.

In contrast to those comprehensive models, which define the constituents of player experience on a content-related level, Nacke et al. [53] have formulated the *Gameplay Experience Model*, which is a framework for the assessment of player experience and thus presents a more abstract view. The model emphasizes the importance of considering two taxonomical dimensions of gameplay, namely abstraction and time, and divides both of them into three layers resulting in a two-dimensional, three-layer gameplay experience model. The dimension of abstraction consists of the layers *Game System*, *Player* and *Context*. The game system itself is the most concrete layer and includes the apparent features of the game such as mechanics, interface and content. The player interacts with this system

and constitutes a more abstract layer, as the individual experience consisting of thoughts and feelings is not entirely externally visible. Finally, the most abstract layer of the gameplay experience is the current context, as it does not only refer to the contemporary interaction but also includes prior experiences, memories, knowledge and anticipated consequences. According to Nacke et al., all three layers include complex processes and interact with each other, thereby determining the experience. The second dimension accounts for the fact that those processes are not static but may change as a function of time due to different contexts, player attributes or changes in the game system. Accordingly, the three layers of time in this model are past, present and future. The *Gameplay Experience Model* is especially informative for the evaluation of games, because it points out that the dimensions of time as well as the three components *game system*, *player*, and *context* have to be considered in order to obtain a comprehensive impression of the experience. Applied to serious games, this underlines that evaluation approaches should not only focus on the player, but also on the context in which the game is used, as well as on possible time effects [53].

Summarized, the concept of player experience and its sub-components provide the basis of game user research in general, and thus should be considered in the evaluation process of serious games as well. Though aspects like fun, flow, immersion, presence, and involvement are often just indirectly associated with the purpose or desired outcome of a serious game, insights regarding the experience of players are supposed to explain their reactions to the game and may also give reasons for unexpected outcomes.

3.3 Learning Outcomes

After having discussed the more general constructs regarding motivation and player experience, this section focuses on a concrete purpose that many serious games are supposed to serve: learning. It addresses the question of how the efficacy of learning games can be evaluated, i.e. how the learning outcome of a serious game can be measured. To test whether the players learn what they are supposed to learn is not an easy task, since learning is a complex process, which is hard to assess and quantify. The research field in psychology that is concerned with the objective measurement of mental capabilities offers a broad range of theories and results. A broad overview of psychometrics for measuring human abilities, also in educational contexts, as well as its underlying models and theories can be found in Kline [41]. A general introduction into the different fields of psychological assessment is provided by Goldstein [29].

The term *learning* does not only contain the acquisition of simple facts, but also the acquisition of other cognitive skills. For instance, Connolly [12] mentions several possible learning outcomes for game-based learning:

- the improvement in knowledge acquisition which might be procedural, declarative, or general knowledge,
- the formation of meta-cognitive strategies,
- and the improvement in the formation of skills.

Accordingly, it has to be stated clearly what learning outcomes are intended by game designers in order to be able to find appropriate operationalizations.

The classical experiment design for evaluating the learning outcome of a serious game is the pre-post-test-design [10,19]. In this experimental design, participants are randomly divided into two groups of which the one is going to play the serious game of interest (treatment group), and the other will be taught by another comparable instruction technique (control group). The skills of each participant are tested before and after playing the game (or after getting the different instructions, respectively). This technique aims at measuring the acquisition of skills by measuring the difference of skills before and after playing the game. In order to analyze whether the knowledge gain is caused by the game and not by other reasons, the players knowledge gain is compared with the knowledge gain of the participants in the control group. The treatment of the control group is crucial for measuring a learning effect and has to be selected carefully and well-conceived [28]. For instance, the basic learning content should be similar in both groups to ensure that theoretically both groups have the same possibilities to access and acquire knowledge about it. Detailed information about experiment design will not be discussed here, but can be found in the chapter about experimental design in this book. In general, designing questions or tasks for the pre-test and the post-test is challenging. Theoretical background for test development can be found in Guilford [32]. Depending on which skill improvement is aimed at in the game, there are standardized and validated tests that might be appropriate to use. In the context of general educational assessment, there are for example the Collegiate Assessment of Academic Proficiency (CAAP)[1], the Collegiate Learning Assessment (CLA)[2], or the ETS Proficiency Profile[3]. Those tests are based on core skill areas and knowledge and supposed to help institutions to assess the general skill level of their students.

Additionally, it is important to be aware of the evaluation model of Kirkpatrick [39], which states that evaluation can take place on four different levels. An evaluation can measure whether the players liked the game (reaction level), to which extent the players improved their skills through playing the game, and can apply them in the game environment (level of learning), whether the players are able to transfer the gained skills outside the game environment (level of behavior), and how playing the game had an effect on the system in which the game was deployed (level of results). Hence, each of the levels needs to be evaluated in a different way.

There are several meta-studies of evaluations of serious games regarding learning [11,28,54]. The study of O'Neil et al. for example reviews all peer-reviewed papers in which learning outcomes in video games for adults are evaluated by quantitative and/or qualitative methods [54]. O'Neil et al. found that out of the 19 available studies, only three evaluated learning outcomes on level three or four of Kirkpatrick's model. Girard et al. [28] present a survey of the

[1] http://www.act.org/content/act/en.html.
[2] http://cae.org/participating-institutions/cla-references.
[3] https://www.ets.org/proficiencyprofile/about.

effectiveness of serious games for learning and conclude that the majority of analyzed games is evaluated by a classical pre-post-test-design of which some include a second post-test in order to measure long-term effects.

While the previous paragraphs describe experiment designs in order to evaluate the learning efficacy of a serious game, the following paragraph presents methods which can be used to concretely operationalize learning and effectively measure the learning outcome of a game. Maki [49] classifies the methods by which the learning outcome of students can be measured into two categories: direct and indirect methods. Direct methods ask students to "represent or demonstrate their learning or produce work so that observers can assess how well students' work or responses fit institution or program-level expectations". As examples for direct methods she mentions responses to questions, interactions within group problem solving, or observable performances. This also includes all kind of (standardized) tests of the particular skill, the players' performance in the game or – if in an educational context – their grades in assignments or exams, which follow the use of the serious game. Bellotti lists several tools for measuring players' game performance [5]. However, Chin et al. note that the usage of exam grades as assessment data is often not a good idea since assessment and grading are driven by conflicting motives [11]. While these methods take place after playing the game, there are also direct methods which take place during the game play and which give further insights. These include think-aloud-protocols, observations of the player by the instructor or video-recordings.

Indirect methods capture students' perceptions of their learning and include all measures in which the players are asked for their perceived learning outcome. Maki [49] mentions inventories, surveys, questionnaires, interviews, and focus group meetings as examples. It is clear that the players' self-reports of their learning efficacy are subjective and might over- or underestimate their actual learning performance. Nevertheless, accompanied by results of direct methods, indirect methods give valuable insights into the students' learning process, interpretations, and perceptions in order to improve the learning process and increase the learning outcome. Maki [49] is a good source for more information of direct and indirect methods as well as for an overview of standardized tests.

The described assessment methods either take place at the end of the learning process (and are summarized by summative evaluation methods) or are implemented and presented throughout the entire learning process (formative evaluation methods) [62]. Formative evaluation can give valuable insights in the players' learning process, however, it is desirable that the assessment methods do not disrupt the game flow. For this reason, assessment methods are incorporated in the game environment in a way that the game experience is not affected. Assessment methods which are "virtually invisible" for the player are called *stealth assessment* by Shute et al. [64]. An overview of embedded assessment in learning games and how game log data can be used for analyzing the learning process of a player can be found in the work of Plass et al. [55], Shute et al. [64], and Loh et al. [47]. Stealth assessment is particularly useful in the assessment of serious games for learning since assessment can be part of the game experience and allows to give

feedback to the user while playing in order to improve the learning process and increase the learning outcome.

3.4 Health and Well-Being

Besides learning, the promotion of health and well-being is one of the biggest application fields of serious games [26]. According to the World Health Organization, *health* is defined as a state of "complete physical, mental and social well-being and not merely the absence of disease or infirmity" [9] (p. 365). Serious games that somehow address the physical or mental health of players or health-related behavior are commonly called serious games for health or shortly health games (cf. the chapter on serious games for health in this book for a comprehensive overview and discussion of characteristics, design approaches and challenges). There are many aspects beside the purpose that differentiate those games from other serious games, like particularly sensitive target groups, specific risks and concerns as well as the involvement of various health-related disciplines. Furthermore, there is a special demand for verifiable effects: It has to be ensured that the game does not have any negative side effects like overstrain, and is at least as effective as any alternate intervention, because otherwise health or recovery may be at risk, especially in case of ill target groups. The goals of serious games for health are manifold and include many different approaches and purposes that are pursued by the games. They may:

- promote movement and physical exercise by giving reason, motivation and/or instructions,
- promote adherence to medical treatment such as the intake of pills by giving information and/or changing attitudes,
- promote healthy behavior like the abandonment of drugs and deleterious food,
- reduce negative thoughts and feelings and improve stress management in various situations (psychological components of health),
- provide knowledge about a disease or health-related issues, or
- support other medical interventions and procedures, e.g. distract from pain or frightening situations.

Depending on these goals the measures that are appropriate to evaluate the effectiveness of the game have to be chosen. As health as such is a highly complex construct impossible to assess directly, the main challenge is to identify sensible indicators of health and find ways to assess those. This will allow for drawing conclusions about how the game effects a player's state. Hence, the main question is what can be measured. As the change of attitudes, which is one of many possibly intended effects of health games, will also be discussed in the following section, this section will mainly focus on the assessment of physical attributes and indicators of health on the one hand, and of perceived well-being on the other.

Indicators of Physical Health and Healthy Behavior. As physical health is inherently related to bodily responses and physiological reactions, it is standing

to reason to use physiological measurement methods like the ones described in Sect. 2.1 to assess health indicators and to investigate the effectiveness of a health game. Similar to games for learning, a pre-post-test-design should be applied in most cases, in order to prove that the game affects the players' health. One sub-category of serious games for health are so-called exertion games or fitness games. Their purpose is to motivate players to physically exercise or to show increased physical activity in order to improve their overall constitution. To operationalize health improvement in those cases, classical medical tests to test stamina and resilience, like stress electrocardiograms (electrodes are attached to the participant's upper body and an ECG is recorded while he or she is riding an exercise bicycle). Furthermore, reaction tests can be applied or indicators like weight and the body-mass-index can be easily assessed.

Some serious games also serve the purpose of increasing medical adherence. Often, the intake of medicine (or its denial) can be tracked by blood analysis or saliva tests. In those cases, the specific indicators related to the medication should be identified and measured (see the popular evaluation of the serious game Re-Mission for children diagnosed with cancer for an example [38]). Of course self-report measures can be used for assessing adherence as well, though it has to be considered that participants may lie if they feel that they did not behave the way they should. See [23] for some guideline to the design of self-report measures regarding medical adherence.

If the serious game is meant to complement medical treatment, that is to say used as a therapy instrument, researchers may use standard evaluation methods, design features and measurement instruments of new medical therapies to prove the game's clinical impact. This was for instance done during the evaluation of the game Re-Mission [38], which is an interesting example of a comprehensive evaluation.

Besides direct physiological measures, game metrics also offer good opportunities to evaluate health games in some cases. If a serious game addresses a very specific part of physical health, its efficacy is best to be measured by assessing a player's performance on a given task over a certain period of time and check the results for significant improvements. For instance, if a game is supposed to improve a person's manual fine motor skills after being handicapped by stroke by training small and precise movements of the fingers/hand, then the accuracy and speed of participants should be tracked by the gaming system.

Indicators of Well-Being. Apart from direct physical consequences, games for health can also focus on improving a person's perceived health and well-being. Well-being is a complex and blurry concept. According to Dodge et al. [18], who considered and analyzed different approaches towards well-being, the term well-being describes a state in which "individuals have the psychological, social and physical resources they need to meet a particular psychological, social and/or physical challenge" (p.230). As this is very difficult to operationalize in terms of objective measures, aspects of well-being are often assessed by self-reports (see e.g. [37] for an overview).

One construct that is closely related to well-being in the case of ill persons is self-efficacy towards the disease [38]. It describes the feeling of being able to cope

with the current situation and the perception of having a chance to overcome the situation. A high perceived self-efficacy is assumed to be beneficial for recovery and thus can be addressed by health games as well, like in the game Re-Mission [38]. Self-efficacy is also often operationalized by using questionnaires. In this context, Bandura proposes a guide for constructing self-efficacy scales related to certain situations [2].

Stress. Finally, two aspects that are quite similar and both related to physical health and psychological well-being, are stress and anxiety. Those concepts are relevant to serious games for health as well, as their purpose can also be to reduce stress (or at least not to increase it). An established questionnaire for assessing stress is the *State-Trait-Anxiety Inventory* (STAI) [66], which is supposed to measure both trait and state anxiety and thus is considered to be an indicator for distress. Moreover, stress has been shown to be related to several physiological processes as well and thus can also be measured by psycho-physiological measurement methods. Here, blood pressure, heart rate, heart rate variability (HRV), skin conductance, cortisol measures as well as the pupil diameter have been used to detect stress levels in several studies [63].

3.5 Attitudes

Attitudes are the last evaluation construct that should shortly be discussed here, as some serious games also aim at changing attitudes of the players and their related behavior. An attitude in general is the evaluation of persons, objects or ideas with favor or disfavor [1]. Attitudes can mainly result from cognitive, affective or behavioral processes and thus are supposed to be modifiable to some degree by offering new cognitive, emotional or behavioral input [1]. However, the operationalization of attitudes is challenging, because they are very individual and subjective. Sometimes, people tend to hide their real attitude towards something if they feel some social pressure or discomfort. Hence, self-reports as well as observations can be applied but are prone to bias and social desirability effects.

As attitudes are an important object under investigation in the broad field of social psychology, researchers have invented more creative techniques to assess attitudes, which can also be applied to serious games evaluation. Those methods are projective, as they are indirect measures and participants are not aware that their attitude is measured. One classical test is the *Implicit Association Test* (IAT) [31]. Broadly speaken, in this test participants have to rate two different concepts (e.g. "junk food" and "vegetables") regarding a certain attribute (e.g. "pleasant"). Then the time they need to react is measured and indicates whether participants think the concepts fit the attribute or not. The whole setting of the IAT, which is too extensive to be described here, is described in [31] and worth a consideration if attitudes should be measured.

4 Summary of Concrete Measurement Methods

To conclude this chapter about operationalization and measurement methods, we present an overview of all the concrete questionnaires mentioned in the text

Table 2. Overview of taxonomies and questionnaires discussed in this chapter, which can be used to assess different evaluation constructs.

Evaluation construct	Name of questionnaire	Sub-dimensions/ Concepts	References
Player experience	Experimental Virtual Environment Experience Questionnaire-Game Pitk (EVEQ-GP)	Presence, Involvement, Flow	Takatalo et al. [69]
Player experience	Game Experience Questionnaire (GEQ)	Competence, Sensory and Imaginative Immersion, Flow, Tension, Challenge, Negative Affect, Positive Affect	IJsselsteijn et al. [36]
Flow	GameFlow Model Heuristics	Concentration, Challenge, Skills, Control, Clear Goals, Feedback, Immersion, Social Interaction	Sweetser and Wyeth [68]
Presence	Igroup Presence Questionnaire (IPQ)	general presence, spatial presence, involvement, experienced realism	Igroup Presence Consortium [34]
Presence	Presence Questionnaire (PQ)	ooehm	Witmer et al. [72]
Immersion presence	Immersive Tendencies Questionnaire (ITQ)	Focus, Involvement, Emotion, Games	UQO Cyberpsychology Lab [71]
Social presence	Competitive and Cooperative Presence in Gaming scale (CCPIG)	competitive social presence (awareness, engagement), cooperative social presence (perceived team cohesion, team involvement)	Hudson et al. [33]
Social presence	Social Presence in Gaming Questionnaire (SPGQ)	Psychological Involvement – Empathy, Psychological Involvement – Negative Feelings, Behavioral Engagement	De Kort et al. [15]
Personality	Big Five Inventory (short) (BFI-10)	Openness to experiences, Conscientiousness, Extraversion, Agreeableness, Neuroticism	Rammstedt et al. [57]
Player taxonomy	Bartle's Player Taxonomy	Achievers, Explorers, Socializers, Killers	Bartle [3,4]
Player taxonomy	Yee's Player Taxonomy	Achievement, Immersion, Social	Yee [73]
Player taxonomy	BrainHex	Seeker, Survivor, Daredevil, Master- mind, Conqueror, Socialiser, Achiever	Nacke et al. [52]
Stress/ Anxiety	State-Trait-Anxiety Inventory	State Anxiety, Trait Anxiety	Spielberger [66]

before and assume this collection to be useful for the planning and conduction of future evaluations of serious games or games in general (see Table 2). Moreover, a short list of further reading recommendations is provided.

References

1. Aronson, E., Wilson, T.D., Akert, R.M.: Social Psychology, 7th edn. Pearson, Boston (2010). Global edn
2. Bandura, A.: Guide for constructing self-efficacy scales. In: Pajares, F., Urdan, T.C. (eds.) Self-Efficacy Beliefs of Adolescents, pp. 307–337. Information Age Publishing, Charlotte (2006)
3. Bartle, R.: Hearts, clubs, diamonds, spades: players who suit muds. J. Virtual Environ. **1**(1) (1996). http://mud.co.uk/richard/hcds.htm
4. Bartle, R.A.: Designing Virtual Worlds. New Riders Pub., Indianapolis (2004)
5. Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., Berta, R.: Assessment in and of serious games: an overview. Adv. Hum. Comput. Interact. **2013**, 1–11 (2013)
6. Bernhaupt, R. (ed.): Evaluating User Experience in Games: Concepts and Methods. Human-Computer Interaction Series. Springer, London, New York (2010)
7. Bernhaupt, R.: User experience evaluation in entertainment. In: Bernhaupt, R. (ed.) Evaluating User Experience in Games. Human-ComputerInteraction Series, pp. 3–7. Springer, London, New York (2010)
8. Biocca, F., Harms, C., Burgoon, J.: Towards a more robust theory and measure of social presence: Review and suggested criteria. Presence Teleoper. Virtual Environ. **12**(5), 456–480 (2003)
9. Bircher, J., Kuruvilla, S.: Defining health by addressing individual, social, and environmental determinants: new opportunities for health care and public health. J. Public Health Policy **35**(3), 363–386 (2014). http://www.palgrave-journals.com/jphp/journal/v35/n3/pdf/jphp.201419a.pdf
10. Campbell, D.T., Stanley, J.C., Gage, N.L.: Experimental and Quasi-Experimental Designs for Research. Ravenio books, New York (2015)
11. 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)
12. Connolly, T., Stansfield, M., Hainey, T.: Towards the development of a games-based learning evaluation framework. In: Games-Based Learning Advancements for Multisensory Human Computer Interfaces: Techniques and Effective Practices (2009)
13. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper Perennial, New York (1990)
14. De Kort, Y.A.W., Ijsselsteijn, W.A.: People, places, and play: player experience in a socio-spatial context. Comput. Entertain. **6**(2), 18 (2008)
15. De Kort, Y.A.W., Ijsselsteijn, W.A., Poels, K.: Digital games as social presence technology: development of the social presence in gaming questionnaire (SPGQ). In: Proceedings of the 10th Annual International Workshop on Presence, pp. 195–203 (2007)
16. Dignan, A.: Game frame: Using Games as a Strategy for Success. Free Press, New York (2011). 1st free press hardcover edn
17. Dillon, R.: On The Way to Fun: An Emotion-Based Approach to Successful Game Design. A K Peters, Natick (2010)
18. Dodge, R., Daly, A., Huyton, J., Sanders, L.: The challenge of defining wellbeing. Int. J. Wellbeing **2**(3), 222–235 (2012)
19. Dugard, P., Todman, J.: Analysis of pre-test-post-test control group designs in educational research. Educ. Psychol. **15**(2), 181–198 (1995)

20. Ermi, L., Mäyrä, F.: Fundamental components of the gameplay experience: Analysing immersion. In: DiGRA 2005, Proceedings of the 2005 DiGRA International Conference: Changing Views: Worlds in Play (2005). http://www.digra.org/wp-content/uploads/digital-library/06276.41516.pdf
21. Fernandez, A.: Fun experience with digital games: a model proposition. In: Leino, O., Wirman, H., Fernandez, A. (eds.) Extending Experiences, pp. 181–190. Lapland University Press, Rovaniemi (2008)
22. Ferrara, J.: Playful Design: Creating Game Experiences in Everyday Interfaces. Rosenfeld Media, Brooklyn (2012)
23. Garfield, S., Clifford, S., Eliasson, L., Barber, N., Willson, A.: Suitability of measures of self-reported medication adherence for routine clinical use: a systematic review. BMC Med. Res. Methodol. **11**, 149 (2011)
24. Garris, R., Ahlers, R., Driskell, J.E.: Games, motivation, and learning: a research and practice model. Simul. Gaming **33**(4), 441–467 (2002)
25. Gee, J.P.: What video games have to teach us about learning and literacy. Comput. Entertain. **1**(1), 20 (2003)
26. Gekker, A.: Health games. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 13–30. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33687-4_2
27. Gill, P., Stewart, K., Treasure, E., Chadwick, B.: Methods of data collection in qualitative research: interviews and focus groups. Br. Dent. J. **204**(6), 291–295 (2008)
28. 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)
29. Goldstein, G., Hersen, M.: Handbook of Psychological Assessment. Elsevier, Hoboken (2000)
30. Goodwin, C.J., Goodwin, K.A.: Research in Psychology: Methods and Design, 7th edn. Wiley, Hoboken (2012)
31. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. J. Person. Soc. Psychol. **74**(6), 1464–1480 (1998)
32. Guilford, J.P.: Psychometric Methods, 2nd edn. McGraw-Hill, New York (1954)
33. Hudson, M., Cairns, P.: Measuring social presence in team-based digital games. In: Riva, G., Waterworth, J., Murray, D. (eds.) Interacting with Presence: HCI and the Sense of Presence in Computer-Mediated Environments. DE GRUYTER OPEN, Warsaw, Poland (2014)
34. Igroup Project Consortium: Igroup presence questionnaire (ipq)
35. IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., Bellotti, F: Characterising and measuring user experiences in digital games. In: ACE 2007 International Conference on Advances in Computer Entertainment Technology, Workshop 'Methods for Evaluating Games - How to measure Usability and User Experience in Games' (2007)
36. Ijsselsteijn, W.A., De Kort, Y.A.W., Poels, K.: The game experience questionnaire: Development of a self-report measure to assess the psychological impact of digital games. Manuscript in preparation
37. Kahneman, D., Krueger, A.B.: Developments in the measurement of subjective well-being. J. Econ. Perspect. **20**(1), 3–24 (2006)
38. 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). 18676516

39. Kirkpatrick, D.L.: Evaluating Training Programs. Tata McGraw-Hill Education, San Francisco (1975)
40. Kivikangas, J.M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., Ravaja, N.: A review of the use of psychophysiological methods in game research. J. Gaming Virtual World **3**(3), 181–199 (2011)
41. Kline, P.: Psychometrics and Psychology. Academy Press, Washington, D.C. (1979)
42. Koster, R.: A Theory of Fun for Game Design, 2nd edn. O'Reilly Media Inc., Sebastopol (2014)
43. Krosnick, J.A., Presser, S.: Question and questionnaire design. In: Marsden, P.V. (ed.) Handbook of Survey Research. Emerald Group Publishing, Bingley (2010)
44. Landers, R.N., Bauer, K.N.: Quantitative methods and analyses for the study of players and their behaviour. In: Lankoski, P., Björk, S. (eds.) Game Research Methods, pp. 151–174. ETC Press, Pittsburgh (2015)
45. Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P., Kort, J.: Understanding, scoping and defining user experience. In: Olsen, D.R., Arthur, R.B., Hinckley, K., Morris, M.R., Hudson, S., Greenberg, S. (eds.) The SIGCHI Conference, p. 719 (2009)
46. Lazzaro, N.: Why we play games: Four keys to more emotion without story (2004)
47. Loh, C.S., Anantachai, A., Byun, J., Lenox, J.: Assessing what players learned in serious games: in situ data collection, information trails, and quantitative analysis. In: 10th International Conference on Computer Games: AI, Animation, Mobile, Educational & Serious Games (CGAMES 2007), pp. 25–28 (2007)
48. Loh, C.S., Sheng, Y., Ifenthaler, D. (eds.): Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement. Springer, Heidelberg (2015)
49. Maki, P.L.: Assessing for Learning: Building a Sustainable Commitment Across the Institution. Stylus Publishing, LLC, Menlo Park (2012)
50. Mandryk, R.L., Inkpen, K.M.: Physiological indicators for the evaluation of co-located collaborative play. In: Herbsleb, J., Olson, G. (eds.) The 2004 ACM Conference, p. 102 (2004)
51. McAuley, E., Duncan, T., Tammen, V.V.: Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. Res. Q. Exerc. Sport **60**(1), 48–58 (1989)
52. Nacke, L.E., Bateman, C., Mandryk, R.L.: BrainHex: preliminary results from a neurobiological gamer typology survey. In: Anacleto, J.C., Fels, S., Graham, N., Kapralos, B., Saif El-Nasr, M., Stanley, K. (eds.) ICEC 2011. LNCS, vol. 6972, pp. 288–293. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24500-8_31
53. Nacke, L.E.: Affective ludology: scientific measurement of user experience in interactive entertainment, Blekinge Institute of Technology doctoral dissertation series, vol. 2009: 04. School of Computing, Blekinge Institute of Technology, Karlskrona (2010)
54. O'Neil, H.F., Wainess, R., Baker, E.L.: Classification of learning outcomes: evidence from the computer games literature. Cirriculum J. **16**(4), 455–474 (2005)
55. Plass, J.L., Homer, B.D., Kinzer, C.K., Chang, Y.K., Frye, J., Kaczetow, W., Isbister, K., Perlin, K.: Metrics in simulations and games for learning. In: El-Nasr, M.S., Drachen, A., Canossa, A. (eds.) Game Analytics, pp. 697–729. Springer, Heidelberg (2013)
56. Poels, K., de Kort, Y., Ijsselsteijn, W.: It is always a lot of fun! In: Kapralos, B., Katchabaw, M., Rajnovich, J. (eds.) The 2007 Conference, p. 83 (2007)
57. Rammstedt, B., Kemper, C.J., Klein, M.C., Beierlein, C., Kovaleva, A.: A short scale for assessing the big five dimensions of personality

58. Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., Kivikangas, M.: Spatial presence and emotions during video game playing: does it matter with whom you play? Presence Teleoper. Virtual Environ. **15**(4), 381–392 (2006)
59. Rheinberg, F.: Motivation, Kohlhammer-Urban-Taschenbücher, vol. 555. Kohlhammer, Stuttgart, 7, aktualisierte aufl. edn. (2008)
60. Ryan, D.: Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp. Educ. Psychol. **25**(1), 54–67 (2000)
61. Ryan, R.M., Rigby, C.S., Przybylski, A.: The motivational pull of video games: a self-determination theory approach. Motiv. Emot. **30**(4), 344–360 (2006)
62. Sadler, D.: Formative assessment and the design of instructional systems. Instr. Sci. **18**(2), 119–144 (1989). http://dx.doi.org/10.1007/BF00117714
63. Sano, A., Picard, R.W.: Stress recognition using wearable sensors and mobile phones. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp. 671–676 (2013)
64. Shute, V.J., Ventura, M., Bauer, M., Zapata-Rivera, D.: Melding the power of serious games and embedded assessment to monitor and foster learning. Serious Games Mech. Effects **2**, 295–321 (2009)
65. Slater, M.: A note on presence terminology. Presence Connect **3**(3), 1–5 (2003)
66. Spielberger, C.D.: State-trait anxiety inventory. In: Weiner, I.B., Craighead, W.E. (eds.) The Corsini Encyclopedia of Psychology. Wiley, Hoboken (2010)
67. Strube, M.J., Newman, L.C.: Psychometrics. In: Cacioppo, J., Tassinary, L.G., Berntson, G.G. (eds.) The Handbook of Psychophysiology, pp. 789–811. Cambridge University Press, Cambridge (2007)
68. Sweetser, P., Wyeth, P.: Gameflow: a model for evaluating player enjoyment in games. Comput. Entertain. **3**(3), 3 (2005)
69. Takatalo, J., Häkkinen, J., Kaistinen, J., Nyman, G.: Presence, involvement, and flow in digital games. In: Bernhaupt, R. (ed.) Evaluating User Experience in Games, pp. 23–46. Human-Computer Interaction Series, Springer, London, New York (2010)
70. Takatalo, J., Häkkinen, J., Komulainen, J., Särkelä, H., Nyman, G.: Involvement and presence in digital gaming. In: Mørch, A., Morgan, K., Bratteteig, T., Ghosh, G., Svanaes, D. (eds.) The 4th Nordic Conference, pp. 393–396 (2006)
71. UQO Cyberpsychology Lab: Immersive tendencies questionnaire (itq) (2004)
72. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: a presence questionnaire. Presence Teleoper. Virtual Environ. **7**(3), 225–240 (1998)
73. Yee, N.: Motivations for play in online games. Cyberpsychol. Behav. **9**(6), 772–775 (2006). The impact of the Internet, multimedia and virtual reality on behavior and society