Validation of Games for Behavioral Change: Connecting the Playful and Serious

Katinka van der Kooij[1*], Evert Hoogendoorn[2], Renske Spijkerman[3], Valentijn Visch[4]
[1*] *Technical University Delft, IDE Faculty, Delft; Vrije Universiteit; Faculty of Human Movement Sciences, Amsterdam, k.vander.kooij@vu.nl*
[2]*IJsfontein, Amsterdam, evert@ijsfontein.nl*
[3]*Parnassia Addiction Research Centre Brijder, The Hague, r.spijkerman@brijder.nl*
[4]*Technical University Delft, IDE Faculty, Delft, v.t.visch@tudelft*

Abstract

The application of games for behavioral change has seen a surge in popularity but evidence on the efficacy of these games is contradictory. Anecdotal findings seem to confirm their motivational value whereas most quantitative findings from randomized controlled trials (RCT) are negative or difficult to interpret. One cause for the contradictory evidence could be that the standard RCT validation methods are not sensitive to serious games' effects. To be able to adapt validation methods to the properties of serious games we need a framework that can connect properties of serious game design to the factors that influence the quality of quantitative research outcomes. The Persuasive Game Design model [1] is particularly suitable for this aim as it encompasses the full circle from game design to behavioral change effects on the user. We therefore use this model to connect game design features, such as the gamification method and the intended transfer effect, to factors that determine the conclusion validity of an RCT. In this paper we will apply this model to develop guidelines for setting up validation methods for serious games. This way, we offer game designers and researchers handles on how to develop tailor-made validation methods.

Keywords: *Persuasive game design, Serious games, RCT, Validation, Game Research;*

1. Introduction

Serious games: contradictio in terminis or treatment of the future? The answer depends on the type of evidence used. Ever since games keep teenagers locked to their game consoles, society has wondered whether these games have a lasting impact that reaches beyond entertainment. After an initial concern about potential negative effects of these games (e.g. [2]), society has turned towards potential positive behavioral change effects. As a result, so-called 'serious games' [3], (also referred to as 'applied' [4] or 'persuasive' [6]) have been developed for a seemingly unlimited variety of behavioral change purposes. Qualitative anecdotal evidence supports the idea that serious games can be useful in motivating patients for behavioral change (e.g. [7-11]). In contrast, quantitative evidence on games' behavioral change effects, is scarce [12-14] relative to the number of studies performed. A recent meta-analysis found 54 quantitative validation studies that produced heterogeneous results and only few showed small overall effects on behavior [12]. Multiple factors may underlie the discrepancy between qualitative and quantitative evidence on serious games' effects. One the one hand, subjective qualitative evaluations may be positively biased by the media hype around serious games beneficial effects. On the other hand, quantitative evaluation of serious games may have been flawed due to the lack of valid and appropriate evaluation tools. The development of appropriate scientific procedures and methods on a specific topic or field takes time and scientific methods may not have been fully adapted yet to the fast-paced developments in state-of-the-art game design. In this paper, we aim to facilitate the development of quantitative validation methods that are adapted to the properties of serious games. In doing so, we focus on the connection between games' entertainment and serious effects. We do this for methodological and practical reasons. Games' playful nature affects the tools with which we should evaluate their serious effects. Moreover, if we do not evaluate both the entertainment and serious aspects in relation to each other, users of serious games will have no solid arguments to choose a game-product over a traditional behavioral change program.

To facilitate the development of research methods that are adapted to the properties of serious game design we connect the main concepts in a research and game-design framework and discuss how design factors affect the ideal set-up of validation research.

As a framework for serious game design we take the Persuasive Game Design (PGD) model [1]. The framework was developed based on existing theory of game design and game experience and has been applied to describe and understand persuasive game design activities at the Technical University Delft – see for instance [4]. The PGD model [1] (Figure 1.B) is especially helpful in this context because it encompasses the full circle of game design and effects on the user without committing to a single theory on *how* games can invoke behavioral change. The model defines persuasive games as “games that invoke a game-world experience in the user to facilitate a behavioral change in the ‘real’ world”. In this sense, persuasive games overlap with most serious games although they can also be considered a subclass of serious games [6].

The PGD model focuses on four central concepts. The first concept is the ‘real-world’ context and subsequent user experience in which the game is applied. The second is the gamification (design) process in which game-elements are designed onto this real-world context to transport [15] the user towards a ‘game-world’ experience. It should be noted that the game-world and real-world experiences are two ends of an experience dimension that are almost never exclusively reached. The third concept is the game-world experience of the user with its protective and motivational qualities [16-18]. The final and fourth concept is the effect of the game-world experience on the user’s real-world behavior, which is called the transfer effect (Figure 1.A). A paper describing an empirical validation of the model in terms of user experiential categories of game versus real world and in the transporting effect of gamification is currently in preparation.

As a research framework we take the randomized controlled trial (RCT; Figure 1.B) which is a generally held golden standard in validation research [19-22] since this method is best suited to assess the causal relation between product and effect. In essence, the RCT is designed to statistically prove that a between-groups difference on an outcome measure or *dependent variable* (e.g. smoking abstinence) is larger than a coincidental difference and can therefore be attributed to the *independent variable* (i.e. product use). The RCT therefore generally includes two ‘conditions’: an experimental condition in which the independent variable (product use) is present and a control condition that does not contain the independent variable (in which the product isn’t used). Large numbers of participants are assigned randomly to these two conditions, minimizing the influence of individual differences on the group’s average outcome measurement. Participants are typically compared at two time-points: before product use and after (Figure 1.A). Finding that the measurements in the two groups were equal before product use, but different after one group had used the product would indicate that the product was effective in achieving change.

As we will discuss in the following paragraphs, the gamification and game-world experience determine control condition design, whereas the transfer effect and real-world context determine the measurement of dependent and mediating variables (Figure 1). In both aspects the RCT needs to be adapted to the properties of serious games. First, a basic assumption of the RCT is that the variables studied (e.g. game use and nicotine abstinence) can be isolated from their context [23], such that a proper control condition can be developed. Play however is intertwined with the users’ real-world context and difficult to control for. What is a game to one may be an annoying task to another: the designer can never know a priori what experiences he invokes in different users. The development of a control condition will thus require advanced methods in the development of placebo games and measurement of play experiences to check for an actual difference between the experimental and control condition. Second, the standard RCT tends to compare users at discrete moments in time. Change, in contrast, evolves dynamically over time and doesn’t occur in a vacuum [24,25]. Therefore validation of serious games for behavioral change will require a more fine-grained temporal analysis.

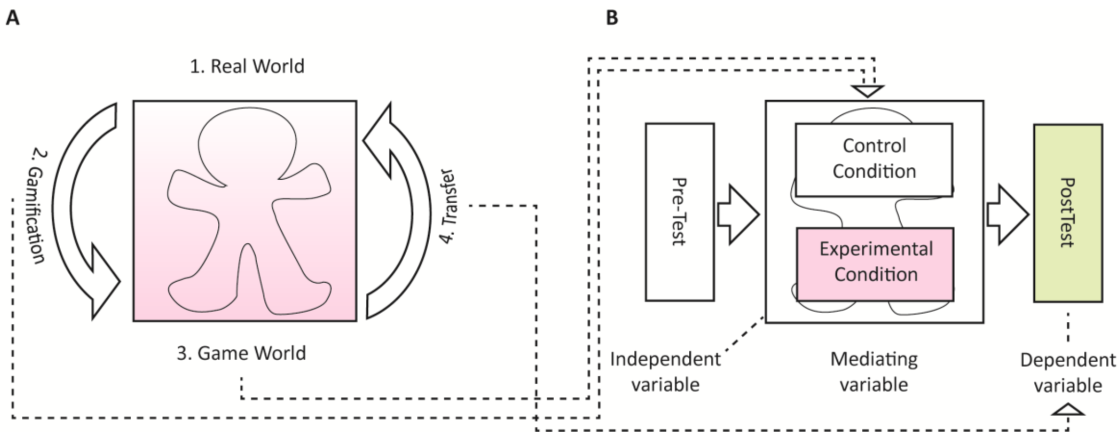

Figure 1. A) Persuasive Game Design (PGD) model. **B)** Randomized Controlled Trial (RCT) mapped onto the PGD model. Dotted lines indicate connections between the main concepts of the PGD model and RCT framework.

2. Gamification and game-world experience: Control condition design

The RCT's ability to provide information about a causal relationship between an independent and dependent variable depends on the formation of a proper control condition. Without a control condition, we cannot know whether an observed effect is due to the independent variables or to circumstantial factors (e.g. a newly installed smoking ban or user-expectations). Ideally, the control condition should be equal to the experimental condition in all aspects except the independent variable. Specification of the context of use, such as the importance of teacher involvement [26] helps in deciding on the population and place where a product should be tested. Yet what exactly is this independent variable in serious game research? The independent variable is not necessarily the full product. For medical drugs the active content of pharmacological ingredients is independent of the user's perception of a drug's capsule. Therefore a placebo drug can be developed whose appearance raises the same user experiences and expectations. For psychological therapies the form (e.g. conversation) and active ingredient (e.g. cognitive behavioral therapy) are both experience-based. Serious game products are somewhere in between medicine and therapy as they are physical products that derive their efficacy from the experiential and behavioral effects on the user (Figure 2). As their physical or digital nature may allow the researcher and designer to separate form and active content for the development of a control condition, it is useful to distinguish between the two. We name the active content, or the part of the game that the designers hold responsible for the transfer effect, the change catalyst. When designing the change catalyst, the designer may focus on different levels of a physical-experiential dimension. A popular division on this dimension in game theory is a hierarchical mechanics, dynamics, aesthetics (MDA) division of games in three levels: mechanics (the game rules), dynamics (interaction with the game rules) and aesthetics (the experiences that follow from the dynamics) [27]. Researchers of serious games have generally used a different two-level division of serious games: game elements (also called 'motivational affordances' [12], which are a combination of the game mechanics and dynamics or play behavior) and game [play] experiences that follow from the game elements [28-30]. In addition to game elements, serious games also contain serious elements that are drawn from the real-world context that the game is designed onto (for instance therapeutic information).

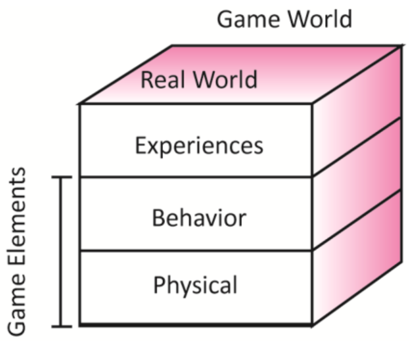

Figure 2. In addition to inducing a transformation from a ‘real’ to a game world experience (Figure 1.A) games consist of different levels: the physical product induces user-behavior, eventually resulting in experiences that can be transformed to a game-world experience.

2.1 Player experiences as change catalyst

When the change catalyst is designed as the full game experience that follows from the combination of serious and game elements, researchers will have to borrow from psychology in forming a control condition. In the validation of psychological therapies three types of control conditions have been frequently used: 1) a ‘waiting list control condition’ in which the control condition receives no treatment until the measurements have been finished, 2) a ‘treatment as usual’ (TAU) control condition in which regular treatment is offered and 3) a placebo control condition in which an intervention that is expected to have no effect is offered.

The waiting-list control condition does not control for the social interaction provided by therapist or researcher and is therefore especially appropriate for single-player digital games that are designed to play at home. Moreover there are ethical concerns in letting one group of patients wait for treatment. The ‘treatment as usual’ control condition, in contrast, does control for social interaction with a therapist and researcher and is a better fit for games that include social interaction or that are designed to be played under supervision of a professional. The main advantage of TAU paradigms is that conclusions can be drawn on the effectiveness of serious games relative to other forms of intervention. For example, in the validation studies of a game for adherence to asthma medication [31], a game to prevent coercion in adolescent sexual relationship [26] and a game to support diabetes treatment [32], special folders were developed that conveyed the same health information as the investigated games. These studies showed a clear benefit of using a serious game rather than a illustrated text medium. An additional advantage of using a ‘treatment as usual’ procedure is that users do not have to ‘waste’ time on a control condition that is designed to have no effect [23,31].

The final option of using a placebo-intervention that is expected to have no effect has been translated in game research in the use of ‘off-the-shelf’ entertainment games as ‘placebo’ games. For instance, in the validation study of Re-Mission, use of a serious first-person shooter game (Re-Mission) that was intended to increase adherence to cancer medication was compared to use of an entertainment first-person shooter game that had no intended behavioral change effects [33]. A similar procedure was followed in the validation of ‘Packy and Marlon’, a game designed to increase adherence to diabetes treatment. In this study, use of the game was compared to use of an entertainment ‘pin ball’ game [32]. We appreciate the intention to control for game experiences that are not directed at behavioral change, but argue that the field of game research is too young to label certain games as having no effect. Entertainment games have been found to have unintended effects on the user – both negative and positive [14,34,35]. The use of a placebo control condition is more feasible for game designs in which the change catalyst is designed onto a specific combination of game- and serious elements as we discuss below.

2.2 Elements as change catalyst – the modular approach

When a more modular approach is taken in which a specific combination of serious- and game elements is considered the change catalyst, a ‘placebo’ game can be designed that is used in the control condition. Such a placebo game is equal to the investigated serious game in all but the change catalyst, for instance by containing the same role-playing mechanics and graphic design but without rewarding self-reports of nicotine abstinence. This way we can investigate the efficacy of a specific game element (rewards for reports of nicotine abstinence) and derive information on whether we should use this game element in future serious games. Cole and colleagues [36] for instance, investigated the effect of interactivity on positive change in attitudes towards cancer medication. This study compared play of the serious game ReMission [33] with passive viewing of a recorded session of typical game play and provided support for the idea that interactivity contributes to serious games' behavioral change effects. However, comparison with a placebo game does not provide information on efficacy relative to other types of treatment. Moreover, the development of a control condition (placebo game) often requires the development of a product whose function is confined to the testing procedure. For the validation of the brain-training game Neuroracer, for instance, a 'placebo' game was developed that was identical to the tested game but did not involve the multi-tasking element that was held crucial for achieving the intended benefits in cognitive control [37]. On a more fundamental note, it is questionable whether it is possible to keep everything but the game-elements equal between the investigated and control product. Game experiences are determined by the interplay of game elements (mechanics & dynamics [27] and treating game elements as separate adjustable elements may imply a denial of the players' freedom in the interaction with the game mechanics.

2.3. Conclusion

Serious games are complex products that achieve their effects through behavioral and experiential effects on the user. Because only a subset of the product may be aimed at invoking behavioral change, we have termed this part the change catalyst. The change catalyst is the independent variable and thereby the object of validation. The design of the change catalyst determines the type of control condition that should be formed in validation research. If the change catalyst is designed on an experiential level, control conditions that are common in validation of psychological therapies can be used: the waiting-list or the 'treatment as usual' control condition. We would advocate against the use of an off-the-shelf entertainment game as a control condition: entertainment games have been found to have unintended effects on the user [14,34,35]. When the change-catalyst is designed onto a behavioral level of game- and serious-elements, the development of a placebo game that is equal in all but the change catalyst is the most valuable option. For instance when a game is intended to increase therapy adherence by means of rewarding adherence to therapeutic assignments (e.g. diary registrations), with the game narrative and visual elements being unrelated to therapy adherence. In this case, a placebo game can be developed that is equal to the investigated game in all aspects but the change catalyst. Doing so requires extra time and budget but does generate generic knowledge on how game elements contribute to behavioral change. This generic knowledge can be used for the improvement of future game designs. However, if we want to know whether a particular game product is more effective than an existing therapy a treatment-as-usual control condition will have to be used.

3. Transfer effect: Measurement of mediating and dependent variables

The ultimate goal of a serious games' validation trial is to demonstrate the behavioral change effect on the user. In validation terminology: we are looking for the product's effect on the dependent variable (Figure 1.B). If there were no limits to the measurements that could be taken and these measurements would perfectly reflect the underlying variable (validity), add no variability (reliability), and have infinite resolution, any existing effect could be observed. Reality is fuzzier however: some variables can be measured only indirectly, causing validity problems, some variables are protected by user privacy, causing access problems, and for other variables we may not yet have developed proper measurement tools. The quality of the dependent variable's measurements determines the effect size that can be observed in a validation trial. Invalid measurements are problematic because they threaten the objectivity of a validation trial. Unreliable measurements make it difficult to determine whether an effect was found by chance (due to the natural variation) or due to an effect of product use. Poor precision, finally, renders us blind to small effects. Naturally, the quality of measurements is determined by the type of variable measured and the available measurement tools. The type of measurements and moments at which they are conducted are determined by the design of the transfer effect. In the following paragraphs we discuss how the level and timing of change affect the quality and methods by which change can be measured.

3.1 The level of change affects the availability of measurement tools

Serious games can aim to impact the user on different levels of functioning [13]. Even when the overarching aim is a change in the user's real-world behavior (e.g. not smoking), the developers of a serious game can chose to lay the health claim on a different level of functioning than the eventual target behavior. In a first draft of the Persuasive Game Design model, borrowing from persuasive technology literature, we made a distinction between the effect on attitudes, behavior or compliance [1]. To ground the model firmer into the variety of behavioral change theories, we propose to adjust it into a division of impact on the users' cognition, behavior and real-world context. Cognitions include attitudes, knowledge and outcome expectancies that have been frequently pinpointed as important determinants of human behavior [38-41]. The behavior category encompasses all forms of user behavior, without distinguishing between self-determined and compliance behavior [42]. The final, 'context' category accommodates all forms of primary impact that do not occur on the user himself but on his real-world context, for instance by reducing barriers or norms for performing a behavior [39,41] or by influencing the user's social circle [40,43]. Below we discuss the measurement tools that exist for the different levels of impact.

First, a serious game may be designed to evoke behavioral change by impacting the cognitive level of attitudes, associations, knowledge and beliefs. This may seem attractive: during game-play, the design may immediately impact the users' mental world whereas accessing the users real-world behavior involves an extra step (Figure 3). However, as cognitions reside in the users' mind, we can only access them indirectly. Either by the users' ability to report on his mental states (introspection) or by performing cognitive tests. Introspective measurement tools derive their quality from repeated use, by which it can be demonstrated that different items consistently measure the same concept and are hence valid and reliable. If a game impacts the user on a cognitive level it is therefore important to assess whether validated measurement tools exist for this specific variable [12].

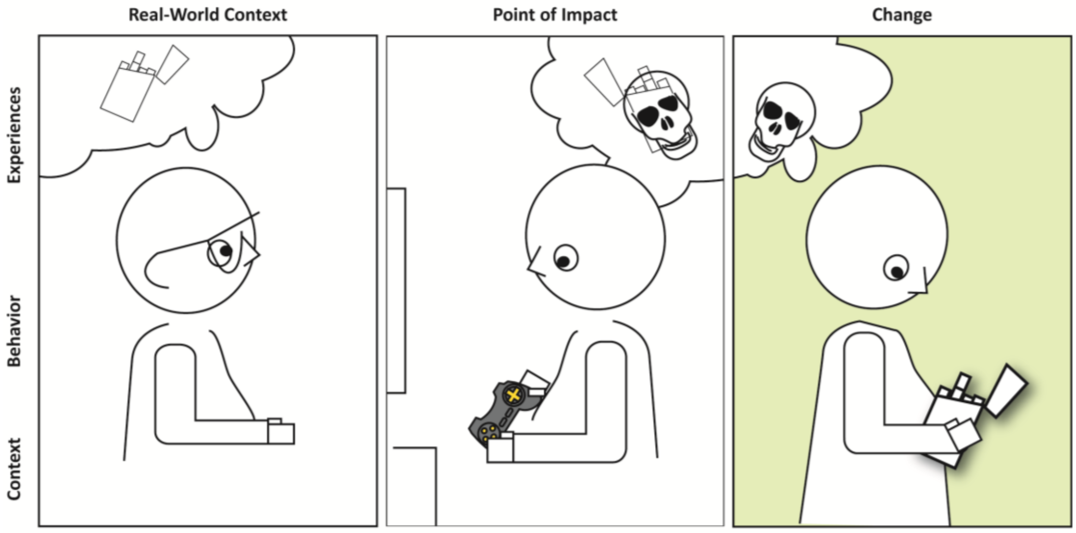

Figure 3. Games can facilitate a behavioral change by first aligning the game product with the user's real world context, secondly by introducing the change catalyst in the user experience during interaction with the game product and thirdly by letting the change catalyst affect the user' post-play behavior.

A second level at which a serious game can aim to affect the user is by directly impacting his real-world behavior. For instance compliance to a medication schedule [32,33,44], reduced sexual risk taking [45,46] or emotion recognition [47]. Behavior seems easier to measure because it can be observed directly: especially for digital games, the tools to gather behavioral data from sensors are virtually unlimited. Difficulties arise however in the capturing of change beyond the episode of game play. Because we cannot see the full context in which data were recorded, measurement errors are difficult to interpret. Moreover, it isn't always possible to predict when a behavior will occur (think for instance about sexual risk taking in contrast with buying cigarettes in a cigarette shop) and when measurements should be taken. Validation research is therefore facilitated by designs that impact the user's behavior at a pre-defined time and place, allowing the researcher to limit pervasiveness of measurements. An example of such a pre-defined behavior is compliance to doctor's visits [32] or completing therapy assignments [48]. In addition, measurements may be difficult to interpret because we cannot see the full context in which data were recorded. A measurement tool with greater sensitivity to contextual information than sensor data is letting the user complete diary reports [19,50]. However, even though digital reminders allow reports to be completed relatively close to the actual behavior, which reduces event fading in the user's memory [51], diary reports generally do have lower resolution and precision than sensory data and rely on introspection, which may cause validity problems. A final, indirect, type of behavioral measurements that we consider is the assessment of a physical reflection of the user's behavior. For instance blood levels of a drug that the user is supposed to take [33] or body weight as a reflection of dietary habits [19]. An advantage of these types of measurements is that they can provide relatively precise information about behavior that is difficult to capture in sensor data (for instance medication ingestion). A nice example of how such a physical reflection of behavior can be more reliable than introspective reports is provided by the validation study of the game Re-Mission. Kato et al. [33] found that self-reports of medication adherence did not differ between the control and experimental group whereas blood levels of the prescribed cancer medication were higher in the experimental compared to the control group.

A third type of impact a serious game can aim for is an impact on the user's interaction with his social context, for instance by strengthening interactions between the user and therapist [50]. Using games to affect the user's social circle may take optimal advantage of games' natural quality in strengthening social relations [17], but it adds an overwhelming complexity to the research table. First of all, finding enough participants can already be a difficult task when single users are tested, when aiming to measure an effect on the user's social circle these individuals will have to be recruited as well. Moreover as players are influenced by a combination of co-player behavior and game mechanics, their experiences will be much more varied compared to when all players respond to the game mechanics only. To reduce such complexity issues, games that affect the user's social context are probably best researched in clinical settings where social contexts are well defined and users can be studied together.

In short, the validity, reliability, precision and accessibility of measurement tools determine the size of an effect we are able to detect in a validation trial. Different types of measurement tools exist for cognitive versus behavioral levels of impact and these measurement tools bring their own challenges and opportunities that we have sketched above. When designing the level of impact (cognitive, behavioral, contextual) for a to-be-validated game, we would recommend the use of a measurement checklist: the smaller the expected effect, the more quality ticks one should be able to set for available measurement tools.

3.2 The process of change / the moment of measurement

The traditional RCT compares an average user at two discrete moments in time: before and after product use (Figure 1.A). The reason for assessing an effect at these pre-defined moments is that this prevents the researcher from using natural variation to 'fish' for a chance effect. When assessing a serious game's effects, however, multiple moments in time may be important because change is invoked in steps (Figure 3) and may unfold over time. In this paragraph we discuss how unfolding change may be captured in an RCT.

Serious games aim to invoke a behavioral change beyond the episode of game play. To be able to evoke this change, the game world needs to connect with the variables that determine the user's behavior in the real world, a moment within the episode of game play that we call the 'point of impact' (Figure 3). For instance an attitude that has been affected by role-playing [46] or neural mechanisms that have been strengthened during game play [36,37]. The intended transfer effect often lies far from this moment, think for instance about a change in sexual risk taking behavior as a result of changed attitudes in contrast with rejecting a presented cigarette. Validation studies tend to focus on the intended transfer effect and pay little attention to the point of impact, which may be difficult to measure as it involves a collision of mental worlds that cannot easily be observed. Yet, measuring both the point of impact and intended transfer effect increases the quality of conclusions drawn from an RCT. First, analyzing the variable that we hold accountable for the invoked change (for instance engagement), enables understanding of how change occurs. This knowledge can be applied to the development of design recommendations for future products. Second, measuring both the point of impact and intended transfer effect allows one to trace back how individual differences in change arise. This knowledge can then be used to supplement information on a game's general efficacy with information on how different individuals were impacted by the game. Interestingly, although quite some studies measure multiple dependent variables, few take the relation between these variables into account in their analyses [12]. One interesting study used brain imaging methods to measure the point of impact where interaction with the serious game Re-Mission promoting adherence to cancer medication, affected activity in brain areas that according to the researchers were related to attitude formation [36]. This study found that activity in these brain areas was correlated with reported attitudes towards cancer medication immediately following game play whereas there was no statistical evidence for a relation with attitudes towards cancer medication one month later. Thus, although the serious effect dissolved over time, there was a significant effect immediately following gameplay which provides fundamentally different information on the game's efficacy than if the game's effects would have only been measured 1 month following gameplay.

A second problem in capturing change is that change is a dynamic process that may take time and can take many shapes [25]. Hence, change may be maximal long after it was first evoked by the game. On the other hand, because users each have their own unique set of cognitions, habits and context that mediate the change process, the shape of change will also increasingly differ between individual users. Therefore the reliability of the causal connection between gameplay and behavioral change decreases with the timespan between play and measurement of behavioral change. Measuring change at a single moment therefore involves a significant risk of drawing conclusions that do not properly reflect the eventual product-use effect on the user. To get a better picture of the shape of change, multiple measurements should be taken [25]. The advantage of studying the shape of change for serious game research is that the benefits of play may be better captured then when a change in a variable interest is captured at a single moment in time. Kato et al. [33] for instance, showed that medication-adherence benefits of serious game play increased with time. When the change in the data points is expected to be linear, the repeated measurements can be analyzed in a repeated measures analysis of variance as in [33]. When change is expected to have a more complex shape, studying the shape of change requires pre-processing of the repeated measures before differences between the experimental and control condition can be tested in analysis of variance. Regression analysis can be used when change is expected to smoothly unfold with time. This is done by fitting a mathematical model of change to the data that have been observed, for instance a linear model if the user is expected to step-by-step acquire self-control skills in controlling a smoking habit. The type of model fit is preferably determined by predictions on how behavioral change will unfold. Unless the transfer effect is designed based on known behavioral change principles, this will generally be difficult. Moreover, those looking for examples of how change may evolve will soon recognize that change is not always best characterized by a smooth curve. Developing minds, like all complex systems, are characterized by nonlinear interactions among system components, phases of sensitivity to and insensitivity to external influences and rapid transitions between states [51]. Change can be triggered by a sudden event and just as suddenly be undone by another event. One may for instance quit smoking upon confrontation with a tarred lung and relapse due to sudden work-stress. When dealing with such erratic changes, a dynamic systems approach can be more effective [25]. Such an approach studies how behavior moves through different states being attracted and repelled by behavioral determinants [24,51]. Rather than fitting a curve through a time-series of measurements, the dynamic systems approach classifies behavior onto a state-space-grid with ordinal behavioral determinants plotted on the axes. Once behavior has been plotted into the state-space grid, the researcher can quantify how behavior moves through different states, for instance taking the number of transitions between cells on the grid as a measure of flexibility [24]. An advantage of this approach is that it cuts a middle path between the mathematical and descriptive approaches remaining descriptively 'real' yet quantitatively faithful to dynamic systems principles [52].

4. Future Directions: into the black box of gameplay

In this paper we have described how game design features, such as the gamification method and intended transfer effect, determine the challenges and opportunities that arise when performing validation research. To facilitate tailoring of validation protocols to game designs, we have developed a vocabulary by which game design and research factors can be connected. The change catalyst is the part of the design to which the transfer effect is attributed and is the object of validation. The scope and level of the change catalyst determine the control condition that should be formed in validation research. The point of impact is the first contact between the user's game-world and real-world and is a mediating variable in validation research. The transfer effect, finally, determines the independent variable that should be compared on the pre- and post-measurements.

The RCT is the procedure that is best suited for providing information on a product's causal effect on the user and is therefore a preferred method of choice in validation research [19-22]. However, the basic RCT is also limited in providing information about individual users and in uncovering experiential effects. The RCT studies an average user to isolate the product-induced effect from an effect that was induced by circumstantial or chance factors [23] but in serious game research, the average user may not reflect the underlying population. The designer can manipulate the game mechanics, but the user is free to use them and let play experiences evolve. People interact with game-like systems in different ways and for different reasons [12]. As a result, users have different experiences, develop at their own rates and may be affected differently by the game. So should the RCT's rigid protocol be dismissed in serious game research to do justice to the individual users, and should perhaps a more qualitative approach be taken? Because the foundation of validation research is a demand for evidence on causality that cannot be met without the RCT [23], we propose that the RCT shouldn't be dismissed and instead tools should be developed to assess additional game-world experiences, taking advantage of digital games' potential for recording fine-grained data on player behavior [53]. By opening the black box of game play, users may become more predictable then if they were studied only before and after game play. Moreover, by adapting game mechanics to player behavior, games can be designed that induce more homogenous game experiences for different users.

Assessing the duration and intensity of contact with the change catalyst (exposure) facilitates interpretation of this variance and an RCT's outcomes. First, even when on average a game didn't cause the intended behavioral change effect, it may still be demonstrated that it did achieve the intended change effect for those users that experienced the intended game-world experiences. This information can be used in decision-making on the application of serious games as well as for the development of future design recommendations and iterative game design aiming to expose a maximal number of users to the change catalyst. Second, by assessing the relation between exposure to the change catalyst and behavioral change, for instance in mediation analysis, general conclusions may be drawn on the efficacy of certain play experiences in facilitating change. Assessing play-experiences properly will require work on the development of proper assessment tools [13]. Whereas psychology has a long tradition of developing tools for the assessment of cognitive variables such as attitudes and memories, relatively few validated tools exist for the assessment of play-specific experiences [54] such as exploration, discovery, competition, immersion, fantasy and sensation [29]. Some recent questionnaires have been developed that capture game experiences such as enjoyment [55], immersion [56] and engagement [57]. These questionnaires have been used in few studies however and need further evidence to proof their validity, to fine-tune them and to allow standardization of test-scores. Moreover, measurements of user experience are usually limited in temporal resolution: a player cannot be expected to continuously rate his experiences and play as he would otherwise. Temporal resolution is important however because individual users may differ in the timing and duration of exposure to the change catalyst, with different behavioral change effects as a result. Since questionnaires are limited in temporal resolution, a promising direction in game-research studies how fine-grained behavioral and psychophysiological measures of play such as physical movement [58], player performance [59], reinforced actions [60] and EEG or EMG [61] relate to playful experiences such as engagement, frustration [60], self-efficacy [59] and positive affect [60]. Incorporating in-game assessments takes us away from the predominant, classic form of assessment compromised by questionnaires, questions and answers and that usually interrupts and negatively affects the learning process [53]. Designing proper in-game assessment that does not interfere with game experience and can be related to the outcome of behavioral change measurement, is a challenge that requires collaboration between designer and researcher [52]. Online (formative [52]) measurement of game-play experiences allows validation research to take behavioral markers for the moment at which playful experiences are expected to have occurred. This information can subsequently be used to supplement information on a game's general behavioral change effects and to improve game design. Validation research isn't necessarily the end of game design: it may serve as input for subsequent design iterations. In the validation study of brain-training game Neuroracer, for instance, task difficulty was adapted to measurements of initial player behavior, allowing for a fairer comparison across players of different skill and age [37].

Finally, would it be enough for a validation study of a serious game to demonstrate only the game's entertainment effects? As serious games, by definition, are intended to change the serious by the playful, a serious game cannot be validated by demonstrating the entertainment effects without the serious effects. However, studying the entertainment value of games is an interesting field of study on its own.

5. Conclusion

The RCT procedures that have been applied in most validation studies of serious games are not optimally suited to measure the full experiential and behavioral potential of serious games. To demonstrate a game's relative effectiveness compared to other forms of treatment, a 'treatment-as-usual' control condition should be used. However to better demonstrate that serious games can have added value because they are games, advances can be made in the development of true 'placebo' games rather than using entertainment games as a control condition. The quality of data on the transfer effect and thereby the sensitivity to small effects can be enhanced by recording time-series rather than single time-points. The interpretation of results, finally, can be improved by measuring change catalyst exposure focusing on additional measurements such as play behavior and experience questionnaires. Challenges for designers will reside in the implementation of online assessment tools that do not interfere with play experiences [52] and challenges for researchers will lie in temporal analysis of fine-grained data on play behavior and serious effects as well as in the linkage of experiential, behavioral and psychophysiological data to play experiences. By facing these challenges, a better understanding can be developed of the added value of using play for serious goals. When the results of the proposed combination of measurements converge, they will be valuable for game designers to improve future serious game designs, for healthcare institutions as they make better-informed decisions on where and when to implement a serious game, and finally for end-users that benefit from more effective products and information on their effects.

References

[1] Visch, V.T., Vegt, N., Anderiessen, H., van der Kooij, K., Persuasive game design: a model and its definitions. CHI. Paris, France, 2014.

[2] Anderson C.A., Shibuya A., Ihori N., Swing E.L., Bushman B.J., et al. (2010) Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. Psychological Bulletin Vol. 136, pp. 151-173, 2014. http://dx.doi.org/10.1037/a0018251

[3] Abt, C.C., Serious Games. New York: University Press of America, Inc, 1987.

[4] Cadamuro A., & Visch V., 'What Remains?': A Persuasive Story Telling Game. In Games for Health. Springer Fachmedien Wiesbaden, 2013.

[5] Graafland M., Dankbaar, M., Mert, A., Lagro, J., De Wit-Zuurendonk, L., et al. How to systematically assess serious games applied to health care. JMIR Serious Games, Vol. 2, 2015.

[6] Bogost, I., Persuasive games: The expressive power of videogames: MIT Press, 2007.

[7] Brezinka V., Treasure Hunt - a serious game to support psychotherapeutic treatment of children. Studies in health technology and informatics, Vol., 136, pp. 71-76, 2008.

[8] Coyle D., Matthews, M., Sharry, J., Nisbet, A., Doherty, G., Personal Investigator: a therapeutic 3D game for adolescent psychotherapy. Journal of Interactive Technology & Smart Education, Vol., 2, pp. 73-88, 2005. http://dx.doi.org/10.1108/17415650580000034

[9] Gamberini, L., Barresi, G., Majer, A., Scarpetta, F., A game a day keeps the doctor away: a short review of computer games in mental healthcare. Journal of CyberTherapy & Rehabilitation Vol., 1, pp. 127-149, 2008.

[10] Shandley, K., Austin, D., Klein, B., Kyrios, M., An evaluation of 'Reach Out Central': an online gaming program for supporting the mental health of young people. Health Education Research, Vol., 25, pp. 563-574, 2010. http://dx.doi.org/10.1093/her/cyq002

[11] Thieme, A., Wallace, J., Johnson, P., McCarthy, J., Lindley, S., et al. Design to promote mindfulness practice and sense of self for vulnerable women in secure hospital services. CHI: changing perspectives. Paris, France, 2013. http://dx.doi.org/10.1145/2470654.2481366

[12] DeSmet. A., van Ryckeghem, D., Compernolle, S., Baranowksi, T., Thompson, D., et al. A meta-analysis of serious digital games for healthy lifestyle promotion. Preventive Medicine, Vol., 69, pp. 95-107, 2014. http://dx.doi.org/10.1016/j.ypmed.2014.08.026

[13] Hamari, J., Koivisto, J., Sarsa, H., Does gamification work? - A literature review of empirical studies on gamification. In System Sciences (HICSS), 47th Hawaii International Conference on. IEEE, 2014.

[14] 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. Computer and Education Vol., 59, pp. 661-686, 2012. http://dx.doi.org/10.1016/j.compedu.2012.03.004

[15] Green, M.C., Brock, T.C., Kaufman, G.F., Understanding media enjoyment: The role of transportation into narrative worlds. Communication Theory, Vol., 14, pp. 311-327, 2014. http://dx.doi.org/10.1111/j.1468-2885.2004.tb00317.x

[16] Caillois, R., Les jeux et les hommes. Paris: Librairie Gallimard, 1958.

[17] Huizinga, J., Humo Ludens - Proeve eener bepaling van het spel-element der cultuur. Amsterdam: Amsterdam University Press, (1938 / 2008). http://dx.doi.org/10.5117/9789089640031

[18] Przybylski, A.K., Rigby., C.S., Ryan, R.M., A motivational model of video game engagement. Review of General Psychology Vol., 14, pp. 154-166, 2010. http://dx.doi.org/10.1037/a0019440

[19] Baranowski, T., Buday, R., Thompson, D.I., Baranowski, J., Playing for real: video games and stories for health-related behavior change. American Journal of Preventive Medicine Vol., 34, pp. 74-82, 2008. http://dx.doi.org/10.1016/j.amepre.2007.09.027

[20] Baranowski, T., Baranowksi, J., Thompson, D., Buday, R., Jago, R., et al. Video game play, child diet, and physical activity behavior change. A randomized clinical trial. American Journal of Preventive Medicine Vol., 40., pp. 33-38, 2011. http://dx.doi.org/10.1016/j.amepre.2010.09.029

[21] Bartholomew, L.K., Parcel, G.S., Kok, G., Gottlieb, N.H., Fernandez, M.E., Planning health promotion programs: an intervention mapping approach. New York, NY: John Wiley & Sons, 2011.

[22] Kato, P., Evaluating efficacy and validating health games. In Games for Health: Research, Development, and Clinical Applications; Amsterdam; 2013.

[23] Brahmajee, K., Nallamothu, R., Hayward, E., Bates, E.R., Beyond the randomized clinical trial: the role of effectiveness studies in evaluating cardiovascular therapies. Circulation Vol., 118, pp. 1294-1303, 2008. http://dx.doi.org/10.1161/CIRCULATIONAHA.107.703579

[24] Granic, I., O'Hara, A., Pepler, D., Lewis, M.D. A dynamic systems analysis of parent-child changes associated with successful "Real-world" interventions for aggressive children. Journal of Abnormal Child Psychology, Vol., 35, pp. 845-857, 2007. http://dx.doi.org/10.1007/s10802-007-9133-4

[25] Laurenceau, J.P., Hayes, A.M., Feldman, G.C., Some methodological and statistical issues in the study of change processes in psychotherapy. Clinical Psychology Review, Vol., 27, pp. 682-695, 2007. http://dx.doi.org/10.1016/j.cpr.2007.01.007

[26] Arnab, S., Brown, K., Clarke, S., Dunwell, I., Lim, T., et al. The development approach of a pedagogically-driven serious game to support Relationship and Sex Education (RSE) within a classroom setting. Computer and Education Vol., 69, pp. 15-30, 2013. http://dx.doi.org/10.1016/j.compedu.2013.06.01

[27] LeBlanc, M., Tools for creating dramatic game dynamics. The game design reader: A rules of play anthology, pp. 438-459, 2006.

[28] Deterding, S., Dixon, D., Khaled, R., Nacke, L., From game design elements to gamefulness: defining "gamification". Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9-15, 2011. http://dx.doi.org/10.1145/2181037.2181040

[29] Garris, R., Ahlers, R., Driskell, J.E., Games, motivation and learning: a research and practice model. Simulation gaming, Vol., 33, pp. 441-467, 2002. http://dx.doi.org/10.1177/1046878102238607

[30] Korhonen, H., Montola, M., Arrasvuori, J., Understanding playful user experience through digital games. International Conference on Designing Pleasurable Products and Interfaces, pp. 274-285, 2009.

[31] McLay, R.N., Wood, D.P., Webb-Murphy, J.A., Spira, J.L., Wiederhold M.D., et al. A randomized, controlled trials of virtual reality-graded exposure therapy for post-traumatic stress disorder in active duty service members with combat-related post-traumatic stress disorder. Cyberpsychology, behavior and social networking, Vol., 14, pp. 223-229, 2011. http://dx.doi.org/10.1089/cyber.2011.0003

[32] Brown, S.J., Lieberman, D.A., Gemeny, B.A., Fan, Y.C., Wilson, D.M., et al. Educational video game for juvenile diabetes: results of a controlled trial. Informatics for Health and Social Care, Vol., 22, pp. 77-89, 1997. http://dx.doi.org/10.3109/14639239709089835

[33] Kato. P.M., Cole, S.W., Bradly, A.S., Pollock, B.H., A video game improves behavioral outcomes in adolescents and young adults with cancer: a randomized trial. Pediatrics, Vol., 122, e305, 2007.

[34] Granic, I., Lobel, A., Engels, R.C.M.E., The benefits of playing video games. American Psychologist Vol., 69, pp. 66-78, 2013. http://dx.doi.org/10.1037/a0034857

[35] Green, C.S., Bavellier, D. Action video game modifies visual selective attention. Nature, Vol., 423, pp. 534-537, 2003. http://dx.doi.org/10.1038/nature01647

[36] Cole, S.W., Yoo, D.J., Knutson, B. Interactivity and reward-related neural activation during a serious videogame. PLOS One, Vol., 7, 2012.

[37] Anguera, J.A., Boccanfuso, J., Rintoul, J.L., Al-Hashimi, O., Faraji, F., et al., Video game training enhances cognitive control in older adults. Nature, Vol., 501, pp. 97-101, 2013. http://dx.doi.org/10.1038/nature12486

[38] Starks, K., Cognitive behavioral game design: a unified model for designing serious games. Frontiers in Psychology, Vol., 15, 2014.

[39] Azjen, I., The Theory of Planned Behavior. Organizational behavior and human decision processes, Vol., 50, pp. 179-211, 1991. http://dx.doi.org/10.1016/0749-5978(91)90020-T

[40] Bandura, A., Social foundations of thought and action: a social cognitive theory. Englewood Cliffs, NJ: Prentice Hall, 1986.

[41] Rosenstock, I.M., Strecher, V.J., Becker, M.H., Social learning theory and the health belief model. Health Education Quarterly, Vol., 15, pp. 175-183, 1988. http://dx.doi.org/10.1177/109019818801500203

[42] Oinas-Kukkonen, H., A foundation for the study of behavior change support systems. Pers Ubiquit Comput Vol., 17, pp. 1223-1235, 2013. http://dx.doi.org/10.1007/s00779-012-0591-5

[43] Fogg, B.J., Persuasive technology: using computers to change what we think and do. Ubiquity, Vol., 5, 2002

[44] McPherson, A.C., Glazebrook, C., Forster, D., James, C., Smyth, A., A randomized, controlled trial of an interactive educational computer package for children with asthma. Pediatrics, Vol., 117, pp. 1046-1054, 2006. http://dx.doi.org/10.1542/peds.2005-0666

[45] van der Stege, H.A., van Staa, A., Hilberink, S.R., Visser, A., Using the new board game SeCZ TaLK to stimulate the communication on sexual health for adolescents with chronic conditions. Patient Education and Counseling Vol., 81, pp. 324-331, 2010. http://dx.doi.org/10.1016/j.pec.2010.09.011

[46] Cense, I.M., van der Werf, W., Haastrecht, P., Maak seks lekker duidelijk. Tijdschrift voor gezondheidswetenschappen, Vol., 90, pp. 205-208, 2012.

[47] Serret, S., Hun, S., Iakimova, G., Lozada, J., Anastassova, M., et al. Facing the challenge of teaching emotions to individuals with low- and high-functioning autism using a new serious game: a pilot study. Molecular Autism, Vol., 5, pp. 37, 2014. http://dx.doi.org/10.1186/2040-2392-5-37

[48] van der Kooij K., Hoogendoorn, E., Spijkerman, R., Visch, V.T., In Games for Health, Schouten B, Schijven M, Gekker A, Fedtke S, Vosmeer M, (Eds.) Utrecht. Springer Vieweg, 2014.

[49] Shiffman, S., Stone, A.A., Hufford, M.R., Ecological Momentary Assessment. Annual Reviews in Clinical Psychology, Vol., 4, pp. 1-32, 2008. http://dx.doi.org/10.1146/annurev.clinpsy.3.022806.091415

[50] Wilkinson, N., Ang, R.P., Goh, D.H., Online video game therapy for mental health concerns: A review. International Journal of Social Psychiatry, Vol., 54, pp. 370-382, 2008. http://dx.doi.org/10.1177/0020764008091659

[51] Lewis, M.D., Lamey, A.V., Douglas, L., A new dynamic systems method for the analysis of early socioemotional development. Developmental Science, Vol., 2, 457-475, 1999. http://dx.doi.org/10.1111/1467-7687.00090

[52] Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., Berta, R., Assessment in and of serious games: an overview. Advances in Human Computer Interaction, Vol., 1, pp. 1-11, 2013. http://dx.doi.org/10.1155/2013/136864

[53] Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., et al., The research and evaluation of serious games: Toward a comprehensive methodology. British journal of educational technology, Vol., 45, pp. 502-527, 2013. http://dx.doi.org/10.1111/bjet.12067

[54] IJsselstijn, W., van den Hoogen, W., Klimmt, C., de Kort, Y., Lindley, C., et al. Measuring the experience of digital game enjoyment. Proceedings of Measuring Behavior; Netherlands: Maastricht, 2008.

[55] Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., et al., Measuring and defining the experience of immersion in games. International journal of human-computer studies, Vol., 66, pp. 641-661, 2008. http://dx.doi.org/10.1016/j.ijhcs.2008.04.004

[56] Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M., The development of the Game Engagement Questionnaire: a measure of engagement in video game-playing. Journal of Experimental Social Psychology, Vol., 45, pp. 624-634, 2009. http://dx.doi.org/10.1016/j.jesp.2009.02.016

[57] Shaker, N., Asteriadis, S., Yannakakis, G.N., Karpouzis, K., Fusing visual and behavioral cues for modeling user experience in games. IEEE Transactions on cybernetics, Vol., 43, pp. 1519-1542, 2013. http://dx.doi.org/10.1109/TCYB.2013.2271738

[58] Trepte, S., Reinecke, L., The pleasures of success: game-related efficacy experiences as a mediator between player performance and game enjoyment. Cyperpsychology, behavior and social networking, Vol., 14, pp. 555-559, 2011.

[59] Chumbley, J., Griffiths, M., Affect and the computer game player: the effect of gender, personality, and game reinforcement structure on affective responses to computer game-play. CyberPsychology & Behavior, Vol., 9, pp. 308-316, 2006. http://dx.doi.org/10.1089/cpb.2006.9.308

[60] Kivikangas, J.M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., et al., A review of the use of psychophysiological methods in game research. Journal of Gaming and Virtual Worlds, Vol., 3, pp. 181-199, 2011. http://dx.doi.org/10.1386/jgvw.3.3.181_1