The Experimental Method as an Evaluation Tool in Serious Games Research and Development

Nataliya V. Bogacheva(✉)

Pedagogics and Medical Psychology Department, Faculty of Higher Nursing Training,
Psychology and Social Work, Sechenov First Moscow State Medical University,
8-2 Trubetskaya street, Moscow 119991, Russia
bogacheva.nataly@gmail.com

Abstract. This chapter aims to provide the reader with basic knowledge about the experiment as a general method that can be applied towards serious games research and evaluation. It explains the main terms and rules of the experimental design, as well as points out the main risks and difficulties to avoid. The chapter also explains the differences in possible conclusions between true experiments, quasi-experiments, and correlational studies.

Keywords: Experiment · Experimental design · Variables · Validity · Biases · Samples · Quasi-experiment · Correlational study

1 Introduction

Serious games are commonly defined as games, “designed for a primary purpose other than pure entertainment” [5]. In this chapter, we are not going to discuss whether this definition is good or not, but it definitely brings up an evaluation problem, specific for all serious games. If the game has some “serious” purpose, then in must not only be entertaining as any other game, but also effective in some definite area. With different types of serious games, developed for various purposes, such as post-traumatic and post-stroke physical rehabilitation [36, 39], therapy of phobias [8, 54], autism treatment [56], cognitive training [4, 52], cognitive disability treatment [48], pain and stress management [6, 13, 55], training programs for surgeons, odonatologists, nurses and other specialists [45], pedagogical and educational reasons [16, 17, 23], business training [7], sports [46], military purposes [17, 32], global project planning [38], racing car design [2] and even research of protein sequences in biology [25] possible outcomes indicate the games effectiveness in different ways. In all these games, one can measure different parameters and there are different criteria of effectiveness, so there is no common receipt for serious game evaluation. In general, while developing or researching a serious game, we eventually face such questions, as “Does this serious game really work? Does the game fulfil its purpose?” There are even more questions that are important: “Does this serious game work as it was supposed to? Is this serious game more effective than some other instrument for the same purpose? Which one of two (or more) serious games, designed for the same purpose works better?” and so on. The final question might be as global as “Why do we actually are making this game?”

R. Dörner et al. (Eds.): Entertainment Computing and Serious Games, LNCS 9970, pp. 284–305, 2016.
DOI: 10.1007/978-3-319-46152-6_12

With entertaining games (games, designed for entertaining purposes in the first place), we can say the game is "effective" if many people buy it, play it and give positive reviews on it. Predicting game success in advance can be difficult, but at least the parameters of the success are rather obvious [33, 34]. With serious games, the effect is sometimes neither obvious nor rapid – and many people are prone to under- or overestimating this rather new and sometimes even exotic educational, training, awareness raising and treatment tool [37, 53]. Here comes one more reason for evaluation in serious games development: it is the way to acquire strong evidence about the effectiveness of the game. First, it can tell the game developers whether they are doing right and second, it helps to convince doctors, psychologists, teachers, parents and other reference groups that this particular serious game and serious games in general are useful (and worth paying for).

When we say that the use of a serious game leads to an improvement (cognitive learning, motor skills development, awareness raising, collaboration, pain reduction, etc.) we suppose that the relationship between the game and its effect is causal (also called "cause and effect"). When we want to check causality between some factors, the most relevant method is ***an experiment***.

1.1 What Is an Experiment?

While most people have some basic ideas about what an "experimental method" is, the definitions of this term may vary greatly. Merriam-Webster online dictionary explains the experiment as "an operation or procedure carried out under controlled conditions in order to discover an unknown effect or law, to test or establish a hypothesis, or to illustrate a known law" [20]. This definition is broad and does not give us much information about the method, except the idea that the conditions of our experimental study must be controlled. This is how the experiment differs from another important research method - the observation. The observation, on the contrary, generally implies that there is no interference with the observable reality.

Another common definition [14] gives us a more concrete idea of what an experiment is: "an orderly procedure carried out with the goal of verifying, refuting, or establishing the validity of a hypothesis." This suggests that **(1)** we should have a hypothesis that requires some verification before setting up an experiment; **(2)** this is the procedure that follows some predesigned order; **(3)** the definition brings up an important concept – the ***validity***. It will be discussed in paragraph 4 of this chapter. The experiment is used when we need to find or prove that there is a causal connection between something – and it is the only scientifically approved method, that can test causality. The simplest example of a causal connection in the field of serious games design is the suggestion that playing a certain serious game really leads to an increase in some skill or knowledge. In the social sciences and psychology, causal connections between factors are known to be the most difficult to set up and the experiment is the only method that provides the researcher with required arguments for this [1]. To specify the possibility of causal interpretations, three rules of the causal conclusion can be introduced: **(1)** variable X changes before variable Y; **(2)** the linkage between the variables is consequential; **(3)** there is no other possible explanation for the causal relationship between X and Y [30].

The third definition to discuss states that an experiment is “a systematic research study in which the investigator directly varies some factor (or factors), holds all other factors constant and observes the results of the variation” [24]. This definition includes the idea of the “controlled conditions” from the first definition and uses the terms “factor” and “variation”. Those concepts lead us towards another important term that requires further discussion – the ***variable*** (see Sect. 3 of this chapter).

While the first two definitions relate to experiments in general, the third definition describes the so-called “controlled” or “laboratory” experiment – the experimental design with the most controlled conditions, where the researcher tries to manipulate or control as many factors as possible. Another type of the experiment that usually occurs in practice is a “field experiment”. In this type of the experiment, the researcher manipulates some parts of the reality outside the laboratory, for example, when the educational serious game is being researched while already introduced into the educational process. The researcher still controls as many factors as possible, without disturbing the educational process. However, the situation is by far not as controllable as in a laboratory experiment, thus field experiments need a lot of caution to avoid confounds.

2 Experiment as a Scientific Method

2.1 The Rules of the Scientific Thinking

All the experiments in every scientific field more or less share the same rules of scientific thinking. First, all the scientists and sciences assume that events around us have a causal effect. Scientific methods help to discover these causes. These two rules of thinking are known as ***determinism*** and ***discoverability***. Without these two assumptions, no science would be possible. Second, all the sciences are based on so-called ***paradigms***. The term “paradigm” was intensively developed in the works of T. Kuhn. According to him, a paradigm in science is a set of “universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners” [31]. The paradigm includes the ways scientists are required to build up their theories and to use empirical methods to prove those theories true. Although another famous philosopher of science, K. Popper argues with Kuhn’s idea that a “normal” scientist is usually bound to use the paradigm in his works [41], it is impossible to build scientific knowledge without following some shared rules.

While paradigms in particular sciences seem to shift in a rather fast pace (a serious games related example: S. de Freitas and F. Liarokapis suggest that the extensive use of serious games for learning can lead to a paradigm shift in education [15]), the basic rules of scientific thinking in general stay rather constant.

C.J. Goodwin [24] summarizes them into five statements: **(1)** scientific knowledge must be ***objective*** (free from the scientists expectations and other biases); **(2)** scientific knowledge must be ***data-driven***; **(3)** scientific conclusions are ***never absolute***, but tentative; **(4)** sciences ask empirical questions, which means that these questions can be answered through ***empirical research*** and **(5)** scientific theories ***can be disproven***. The last point represents K. Popper’s concept of falsifiability of theories – a theory can be considered a scientific one only if it is at least hypothetically possible to falsify that theory [42]. In this case, "falsification" does not refer to any kind of fraud. Instead, it means that a theory is open for possible disproof. What is more, methodological implications from the famous K. Gödel's incompleteness theorems state that in every formal system (and every science is a formal system) there are statements that cannot be proven within this system [12], so scientific knowledge is always incomplete and open for further development.

As it was mentioned above, scientific theories use empirical data as a resource for development, growth, and possible falsification. The relationship between theoretical and empirical knowledge is built by deductive and inductive thinking. Through induction, we reason numerous events (e.g. the results of the experiments) into general theories, while through deduction we state some theoretically based hypotheses about the events (results) that would possibly occur. At this point, we face real difficulties, as the theories and the reality do not use identical elements.

2.2 Theoretical and Experimental Hypotheses

For example, one wants to develop a serious game for a medical purpose such as distracting a child patient from pain and discomfort at the dentist's (Dutch scientists developed a game with this underlying idea [6]). Discussing the possibilities to develop the game, the authors suggest that the key point to relaxation (a required state of patient) is immersion. However, immersion (as well as relaxation itself) is not something from the objective reality. It is a hypothetical construct, which belongs to the theoretical level of science. You can measure someone's heart rate to see if the person is relaxed or not. You can run an IQ test to find out something about the intelligence of the student or use an academic test to measure his or her knowledge in a certain area, but this is all possible only because we have some theory about what relaxation, intelligence and knowledge are. Psychological laws, concepts, and terms belong to the theoretical level of thinking. Heart rate, true or false answers, and behavior patterns happen in the reality. The link between those two worlds sometimes seems obvious, and we jump from one level to another without much thinking.

However, it is more difficult, when we plan an experiment. First, we need to develop a hypothesis, based on our theory (a simple definition for a theory is "an existing knowledge that scientists use to explain and predict events" [24]). When we suggest that the use of a serious game leads to an improvement (cognitive learning, motor skill development, awareness raising, collaboration, pain reduction, etc.) we suppose that the relationship between the game and its effect is causal. As mentioned above, when we need to check causality between some factors, the most relevant method is an experiment.

A link between the theory and the research is the hypothesis, "a reasoned prediction about some empirical result that should occur under certain circumstances" [24]. Another definition for a hypothesis, retrieved from the online dictionary is: "a tentative assumption made in order to draw out and test its logical or empirical consequences" [26]. With both these definitions, we can see that a hypothesis is a statement and not a question. An empirical question usually precedes the hypothesis, but we need a theoretical background to make a hypothesis. And we need the hypothesis to conduct an experiment. Hypotheses, however, can differ. If we suggest that a particular serious game raises human awareness of ecological problems, we have a causal hypothesis, a hypothesis that predicts a cause and effect relationship. This hypothesis is a theoretical one and uses theoretical terms, like “awareness”. To conduct an empirical study, we will need to transform this hypothesis into an experimental one, where we establish the empirical evidence for ecological awareness.

On the other hand, a hypothesis can be formulated like “people, who play our serious game have higher awareness about ecological problems” – a theoretical hypothesis as well, which is not causal, but correlational, a hypothesis about a connection. The linkage, however, says nothing about cause and effect relationship. Such hypotheses are proved through so-called correlational studies. For example, we found out that higher use of the serious game coexists with higher knowledge of ecological problems. Then there will be at least two possible explanation. The first one is that our serious game develops ecology-oriented thinking (something that we really want to prove as the game developers) and the second one is that people, who already are anxious about ecology, are more likely to play our ecology-oriented game. Maybe, they think they can learn from it. One of the explanations, or both of them, or none of them can be true, but we are unable to prove it empirically before we conduct an experiment. Well, actually, we can provide strong theoretical reasons to promote the explanation that we think is more liable (and/or desirable) but there will always be a possibility for counter arguments.

Nevertheless, even if we have a causal hypothesis, there is still a lot of work to be done before we can conduct an experiment. As the world of theories and the world of objective reality merge in the experimental study, we need to “translate” our hypothesis into the terms of measurable parameters. This process is called ***operationalization*** and the parameters that substitute theoretical constructs are ***variables***.

3 Variables

3.1 Dependent and Independent Variables

A ***variable*** is an operationalized parameter or attribute of an object. Wikipedia describes the process of operationalization as “a process of defining the measurement of a phenomenon that is not directly measurable, though its existence is indicated by other phenomena” [40]. We can suggest that a serious game increases the users’ knowledge in some field, but to prove this in an experiment we initially need to operationalize this knowledge or, in other words, find some measurable attribute, that represents the knowledge. For example, we can measure someone’s knowledge with an academic test. We can operationalize this knowledge in terms of behavior – for example, we suggest that ecology-oriented person will not ignore a kicked down trash can. Therefore, we can organize this condition and see what is happening… In this example, we showed that the same concept from the hypothesis can be operationalized in different ways. Almost the same theoretical hypotheses can be possible proven by very different (in the terms of variables) experiments.

As the name states, variables do vary, or, in other words, have levels. For example, with the academic test, we do not usually use the exact test scores. More often, we subdivide the group into subgroups, like “those who successfully passed the test” or "those who did not pass the test" (two levels of the variable "academic knowledge"); "those who scored low", "those who scored medium", "those who scored high" (three levels of the variable) and so on. Playing a serious game can be operationalized with variable levels: "played the game" and "did not play the game" (two levels) or "played a short amount of time", "played a lot of time", and "did not play the game" (three levels). This seems obvious, but we must be extremely accurate with variables levels operationalization. We cannot voluntary assign users, who play 10 h a week to "played a lot of time" group without theoretical or statistical explanation, why this is "a lot of time" and not "a moderate amount of time". The operationalization is also needed to transfer the collected data (scores and measures) into the variable's levels.

In every experimental study, we meet at least two types of variables: ***dependent*** and ***independent***. In a correlational study, instead, the variables are equal to each other, so there are no dependent or independent variables.

Independent variable is the variable that we can control directly. J.S. Goodwin describes the independent variable as "the factor of interest to the experimenter, the one that is being studied to see if it will influence behavior" [24]. This definition, however, includes so-called ***subject variables***. A subject variable is a variable that differentiates subjects from one another, but it exists prior to our research. Gender, age, intelligence, sometimes – educational level are all subject variables. The research, where we cannot influence independent variables directly and use subject variables instead is a quasi-experiment. This research scheme will be discussed at the end of this chapter.

The independent variable must have two or more levels. The two-level variable is called bivalent and a variable with more than two levels – a multivalent [30]. In general, we need to pick as many experimental groups (of participants), as there are levels of the independent variable. If we have more than one variable, we need enough groups to test each level of each variable separately. For example, with two bivalent independent variables we need at least four groups to cover all possible combinations of variables' levels, etc. This type of experimental design is called between-group design. In serious games research, we sometimes prefer to expose the same group of participants to different levels of the independent variable. This design is called within-subjects experimental design. Both types of the experiments are further discussed in Sect. 6 of this chapter.

According to J.S. Goodwin [24], there are three main types of independent variables in psychology and social science experiments: **(1)** ***Situational variables*** – the different environmental features, that the participants encounter; **(2)** ***Task variables***, which occur when participants are asked to complete different tasks and **(3)** ***Instructional variables***, where the participants are asked to perform the same task in different ways or under different circumstances like different payment for solving tasks. We can meet all these types of independent variables in serious games research. For example, in the Dutch research of game-based cognitive control training for elders [52], the independent variable had two levels: one group of the participants used video games for their training, while people from the other group were watching documental films and were completing quizzes. This is the example of the task independent variable.

If one of the independent variable's level is a zero level (the participants in one of the experimental groups receive no treatment, do not play the serious game, etc.), the group, receiving this zero level independent variable is called the ***control group***, opposed to other, ***experimental groups***. Depending on the hypothesis you want to check and the empirical question you want to answer, the control group may appear or may not appear in your experimental design. For example, if you want to compare two different serious games, developed for the same purpose – then you have two experimental groups. However, if you want to compare future surgeons who played the training serious game with their fellow students, who did not – there are an experimental group and a control group.

The variables that are influenced by the levels of our independent variables are called ***dependent variables***. J.S. Goodwin [24] describes these variables as "those behaviors that are the measured outcomes of experiments". We do not manipulate the dependent variables directly, but we can measure them if they are properly operationalized. The experimental hypothesis usually includes our assumption about the behavior of the dependent variable. For example, we assume that our participants' test performance in history would raise after they played our educational serious game. The test score is the dependent variable. The hypothesis, though, can be either confirmed or denied.

It is possible to measure several dependent variables in the same experiment. One independent variable may influence more than one dependent variable. However, such experiments require more complex statistical analysis [22] on further stages of the research, as simple statistical procedures, commonly used in experimental research, ignore interactions between the variables [27].

3.2 Other Types of Variables in the Experimental Research

While we only need independent and dependent variables to imagine an experiment, in the real experiment, we can never separate them from many other factors. When we deal with people of certain age, background, personalities, experience we can never ignore the fact that there are many variables, influencing their behavior and responses. The variables, which can influence the results of the experiment and therefore must be controlled are known as ***extraneous variables***. Extraneous variables, that have not been properly controlled and appear alongside the independent variable's levels are called ***confounding variables***. These variables give us an alternative explanation for the relationship between the independent and the dependent variable and therefore limit our possibility to verify the causal relationship. In digital game research, the participant's gender is generally a confounding variable, as in general population men are more likely to play video games in their everyday lives, comparing to the women, and there are age differences between male and female gamers as well [18]. This means, that if we experiment with a serious game, designed for young adults, we are more likely to have male participants with more video games experience, comparing to the female participants of the same age. The gaming experience, as well as gender specified differences in spatial thinking, working memory, etc. can influence the results of serious game training as well if we do not control those parameters [28].

The most obvious way to control extraneous variables is to hold them fixed. Such variables are called ***control variables*** or constants [35]. It might seem that we need to control as many variables as possible, but with too many constants, our experimental condition becomes extremely artificial. We can get a clear causal relationship between the variables in the laboratory with many constants, but that condition would be artificial and impossible in the real world where no one holds control variables fixed. With the serious games research, that point is crucial, as we develop them mostly for practical reasons. Therefore, control variables are very important, you should not try to maximize the number of them if you do not want to get the result only applicable in a laboratory.

Such variables as age, educational level of the participants, their psychological characteristics, IQ level, gender and many others can deeply influence the experimental results, but if we use them as control variables, we can end up with the results, adequate only for, e.g. highly intelligent male participants age from 25 to 30. This condition is certainly not generalized.

Another way to deal with such variables is ***randomization***. In this case, participants with different levels of uncontrolled variables are randomly assigned to different groups. With large enough groups, the possible side effects of different variables would compensate each other with no significant impact on the main experimental effect. If the groups are rather small, however, it is statistically possible that one group will differ greatly from another – for example, the participants from one group might be older or there might be significantly more female or male participants in one of them. In this case, the randomization is still possible, but with some constraints to make the groups equal. For example, the male and female participants are assigned to groups at random but we keep the number of them equal in each group. In our example from van Muijden and al. study [52] the groups are randomized, but equal by age, level of education, IQ, and psychological state. These variables are rather general random variables for most of the studies in psychological and social research, but you might want to consult a specialist in the particular area you develop your serious game for to find what is important for your research.

Alongside with variables, that appear due to participants differences, there are also variables inside the experimental design, which provide some additional circumstances. When the researcher does not recognize such variables or does not properly control other variables, the experimental results are influenced by ***biases***.

3.3 Biases

In general, we say that someone's viewpoint is biased when a person views things from a partial perspective and refuses to consider alternative points of view. In a scientific research framework we speak about biases when the research or the researcher's conclusions are incorrect due to some inner mistakes, intentional or not. There are different biases that may occur during the experimental evaluation process, but most of them can be subdivided into several groups:

- ***Sampling biases*** - occur when our sample does not match the referencing population. General population and samples are discussed further in the chapter, but in short, the sample we use must adequately represent the population or target group we are referring to. If we design a serious game for schoolchildren, we should not base our evaluation on adults and vice versa.

- ***Selection biases*** - occur when for some intentional or unintentional reasons the control and the experimental groups in a between-group experimental design are different. Accurate randomization and variables control could help with these biases.
- ***Response biases*** - self-selection of the respondents according to some implicit variables. Some people are more willing to take part in the research, while the others are not. These participants might not represent the general population.
- ***Performance biases*** – occur when participants from one group behave differently due to some reasons. Sometimes performance biases occur due to inequality in the experimenter behavior (for example, more attention towards the experimental group). In a critical article about entertaining video games related cognitive training T. Shubert and T. Strobach [47] suggest that commitment to training and motivational state might affect the results of the experimental study and artificially enhance the experimental effect. Another example of the performance bias is the placebo effect. The participants from the experimental group under certain condition believe that might perform better and they really do, but not because of our treatment or playing the game. Such performance biases can be avoided with a blind experiment – a scheme, where the participants do not know, which group they belong to. In the double-blind experiment neither participants nor the experimenter knows where the participants do belong to. Thus, these schemes require more efforts and resources to conduct.

4 Validity and Reliability

Avoiding biases is an important problem of the experimental research. There are two other concepts that a researcher must keep in mind when conducting any type of the empirical research: ***validity*** and ***reliability***.

The word "***valid***" is defined as a synonym to "justifiable" and "logically correct" [50], while "***reliability***" means "the extent to which an experiment, test, or measuring procedure yields the same results on repeated trials" [44], a synonym for ***repeatability***. So, an experiment is considered reliable, if anyone can repeat it, using the same variables and matching population. As for the validity, J.C. Goodwin [24] suggests four types of validity in an experimental design.

1. ***Statistical validity*** – is determined by accurate and adequate use of statistical methods. The threats to this type of validity are wrong analysis tactics and deliberate analysis, where the researcher describes only the results that match his or her experimental hypothesis.
2. ***Construct validity*** – is determined by the accurate and adequate operationalization of independent and dependent variables. In psychology and social sciences, the use of some constructs inevitably threatens construct validity of the research. For example, one of the most controversial topics in modern cyberpsychology is the relationship between violent digital games and aggression. The research group under the leadership of C. Anderson [10] sees the violent video games as the proved source of aggressive behavior and thoughts in children and adults, while other researchers, including C. Ferguson [21] point out that Anderson's methods of aggression

measurement lack the construct validity and this is a shortcoming to the whole experimental research's conclusion. However, sometimes it is equally difficult to prove both the construct validity and the lack of the construct validity of the research, mostly due to ethical reasons. It is very difficult to maintain ethics in true experiments dealing with aggression, violence, discrimination, etc. The researchers need either to perform correlational studies instead or find some non-obvious and ethical ways to operationalize those important parameters. For example, in C. Anderson and K. Dill research [3] aggression was operationalized through "noise blast", a noxious blast of white noise with changeable intensity and duration. Participants, who used longer and more intense noise blast, than the others, were supposed to be more aggressive and violent. Though the linkage between the noise punishment and real aggression is arguable, the experiment itself shows a creative way to operationalize a difficult concept.

3. ***External validity*** – is the degree to which research findings can be applied out of the experimental sample. Most of the time for obvious reasons we deal with samples that do not resemble the general population. The most acquirable and willing participants for many researchers are students, but the question is whether we can distribute their results to people of different age or background. Aside from their age, students are likely to have higher mean intelligence than the rest of their contemporaries and on certain faculties, they might have specifics abilities and psychological characteristics as well [30]. This means that there are numerous risks for external validity when we use students' samples. Response biases also threaten this type of the validity alongside with mistreatment of such confounding variables as gender.
4. ***Ecological validity*** belongs to external validity but relates not to the samples, but to the experimental environments. Many laboratory experiments while being perfectly reliable often lack this type of validity. J.C. Goodwin also points out that historical context influence the external validity of classical experimental research as well, as they might not be valid in modern society [24]. D. King, P. Delfabbro and M. Griffiths point out that playing digital games at home or in the laboratory is a very different experience for the player [29]. Besides the environment, laboratory experiments are usually time-bound while some of digital games effects require a lot of time to develop, so we might fail to prove our serious game works due to lack of time or participants being nervous. On the other hand, a serious game that worked in a laboratory with few distractions might not work in a crowded classroom with old and slow computers or with smartphones instead of 10" tablets.
5. ***Internal validity*** – apparently the most important type of the validity. It qualifies the complete experimental research as being valid. An experiment is internally valid if its methodology is adequate and confounding variables properly controlled.

Different types of validity apply to different types of experiments. Internal validity should be evaluated in every experimental research, regardless of its type. The same could be applied to the statistical validity, as the use of statistical analysis is common for the experimental research. External validity is important for the experiments with broad and practical conclusions, while construct validity is crucial for experiments with a highly theoretical background [30].

5 The Verification of a Statistical Hypothesis

5.1 H1 and H0 Hypotheses

Earlier in this chapter, we discussed that the connection between the independent and dependent variables forms our experimental hypothesis. In the research design, the experimental hypothesis stands between the theoretical hypothesis level and the statistical hypotheses level. We have already discussed the nature of theoretical hypotheses. As for statistical hypotheses level, we require it when we use inferential statistical methods (we use inferential statistics when we want to make interferences about the population while working with our samples; to describe the characteristics of our sample we use descriptive statistics) to prove that the differences between some groups of variables are statistically significant. In the experimental setting, we usually need to prove that there is a difference between dependent variables levels in an experimental and control conditions. In the other words, we need to prove that our independent variable really affects the dependent one though the methods of inferential statistics. Most of these tests are based on null hypothesis significance testing. The null hypothesis (often referred as H0 for short) states that the levels of the independent variable have no effect on the dependent variable level [34]. In our example with the ecologically oriented serious game, the H0 hypothesis says that people who played the game and those who did not will have just the same levels of ecological consciousness. Even if the results are slightly different in numbers, it does not mean anything, as they are insignificant. The opposite of the null hypothesis is the alternative hypothesis, also known as the H1. The H1 hypothesis suggests that there are significant differences between the levels of dependent variable, affected and unaffected by the manipulations with the dependent variable. If we reject H0 hypothesis and accept the H1 hypothesis, we state that there is ***a significant experimental effect***.

However, any conclusions on the statistical hypotheses are only made with a certain confidence degree – thus, there is always a probability for a mistake. In fact, there are two different types of the mistakes, occurring while operating the statistical hypotheses.

If we falsely reject the H0 hypothesis, when it was true, we face the Type I error. If we falsely reject the H1 hypothesis, we face the Type II error. In psychology, we usually set the confidence interval for Type I error as 0.05. This means, that the chance to reject the H0 hypothesis falsely is 5 %. With confidence interval equal to 0.01, this chance is reduced to 1 %. This level usually depends on the sample size (with relatively small samples 0.05 confidence interval is more common, while with large samples 0.01 interval is more accurate).

5.2 Basic Inferential Statistical Methods

Talking about inferential statistics in general, we cannot avoid discussing some of its methods. There are not so many serious games studies, involving multivariate analysis, structural modeling, and other advanced statistics methods, so we relegate this part to further reading [22]. On the other hand, such methods as ANOVA or

Student's T-test appear in many studies, including experimental research (for formulas see [35] or any other textbook in statistics).

There are two types of data that can be gathered through the research: quantitative (deals with numbers, measures something) and qualitative (describes something, but does not measure it). In the experimental research, we usually deal with quantitative data, but some qualitative data can be obtained as well. Gender, nationality, preferences are qualitative characteristics of people, while their reaction time, IQ score or heart rate are quantitative. At the same time, there are different types of data inside those groups. Thus, qualitative data can be measured by a nominal scale or by an ordinal scale. Quantitative scales are either interval or ratio (with the statistics being mostly the same).

Gender is an example of a nominal scale. The only thing that you can do is to count the number of people with each gender. The same rule applies to nationality, skin or hair color, etc. "gamer" and "non-gamer" also belong to nominal scale, while the amount of time spent in games is not. The only inferential statistics procedures, applicable for nominal data is Chi-square. If you know, that among gamers there are 59 % males and 41 % females [18] and you have a sample of 40 gamers, 13 males, and 27 females among them, you can use Chi-square to evaluate, whether you sample reflects the general gamers population or not (the answers will be "no" with confidence interval around 0.004).

If you ask your participants to rate your serious game with such parameters as entertainment, difficulty or immersion and ask them to use a five-item Likert-like scale (e.g., 1 stands for a completely boring game while 5 stands for a very interesting game), you gather ordinal data. You can never measure the amount of interest between 4 and 5 points on this scale, and you cannot say for sure that a person, who rated the game with 5 received more positive emotions than a person, who rated it with 4. Thus if you want to compare two games or two groups, you can do it, using ***Mann-Whitney U test,*** and if you want to see, how the scores changed, for example on the first and the last level of difficulty, you can use ***Wilcoxon matched-pairs signed-ranks test***. Note, that those two tests show the difference between groups or conditions, either separate like two different groups (Mann-Whitney) or related like the same group on different stages of the game (Wilcoxon). If you want to see how subjective entertainment is linked to subjective immersion, use non-parametric correlation test (***Spearman rank-order correlation coefficient*** seems to be the most common). If you have three or more levels of the variable to compare, there ***is Kruskal-Wallis test***. Note that it only tells you that there is a significant difference somewhere between the groups, but it does not mean that all of them differ significantly from one another.

In the experiment, we are more likely to use parametric statistics. Those methods can be applied towards interval or ratio scales only. Temperature is an example of interval scale data (every single °C is equal, so you can say that +10°C is 5°C warmer, then +5°C, but it does not mean that it feels twice as warm) [35]. Time is an example of ratio scale data, as 20 s are twice as long as 10 s, etc. With such scales, you can use ***Student's T-test for dependent or independent samples*** for two-level variables and one-way ***ANOVA*** for multivalent variables. Note that you will still need Bonferroni's or Scheffe's method to compare separate groups, after ANOVA showed that there are significant differences. For correlational research, ***Pearson's correlation coefficient*** is the most common. Note also, that you are not supposed to use parametric statistics if your data is not normally distributed (does not have that well-known "bell" shape and/or specific mean and standard deviation parameters; it can be checked with Kolmogorov-Smirnov test). Additional correction is also needed if the measurement has different dispersion in the groups you compare. This might happen with relatively small or ungeneralized samples, but you still can use non-parametric statistics.

6 The Participants in an Experiment

6.1 Sampling and Sample Sizes

As it was described earlier in this chapter, the recruitment of the participants is an extremely important part of the experimental research. Adequate sampling influences the external validity of the research and helps to avoid many types of biases.

In a perfect experiment, we would be able to access and test an unlimited number of participants. Of course, this is impossible. In the real world, we can only access samples, more or less representative. A representative sample is a sample that is formed out of the general population and copies its general internal structure.

J.C. Goodwin [24] points out that the psychologists often use a so-called ***convenience sample.*** A convenience sample can be recruited by different ways – for some studies they can be students, while for other research you might need to place ads in the newspapers or use a so-called "snowball" sampling. Although convenience samples are very common in the experimental studies, these samples often lack the external validity and are prone to sampling and response biases.

Simple random sampling – participants are drawn from some general population at random, usually with the help of random numbers generator. While this type of sampling is sometimes used for survey studies, especially in sociology, there is also a statistical chance for biases.

Stratified sampling – unlike the random sample, stratified sample represents the adequate proportions of important subgroups in the population. It is important to plan this type of sampling relying on those factors that can influence the results of the research.

Cluster sampling – is used when it is impossible to acquire a complete list of individuals to run a random or stratified sampling.With a cluster sampling, a few of relatively identical groups are selected, like school classes or students. Cluster sampling can be combined with stratified sampling for better results.

It is worth to remember, that experimental research design requires a number of identical groups, determined by the number and levels of independent variables. Therefore, not only we need to create a more or less representative sample, but to divide it into equal groups as well. With a big enough sample a random assignment will do, with a procedure of block randomization used to ensure that every group gets an equal number of participants (in each block a participant is assigned to each condition). However, if there are only a few participants, randomization possibly leads to biases. In this situation, matching is a preferred alternative for randomization. Matching means that the experimenters choose matching variables and pick up pairs/triplets/etc. of participants with the nearest scores in these variables. One participant from each pair will belong to one group, while the second one – to the other one.

As for the general amount of the participants, there is always "the more the better rule", as the bigger sample usually tends to be more representative and we are more likely to get statistically significant results.

To give a more precise answer to the question "How many participants do we need?" we need to introduce the concept of experimental effect. The experimental effect in the population, the preferred statistical method, confidence interval and the sample size form the power of the research. In a good research, we try to achieve the power of at least 0.80 with a confidence interval level of 0.05. With the medium population effect, we will need at least 64 participants in each group (if we use Student's T-test for independent samples) or 33 participants in each group (if we use Student's T-test for dependent samples) [30]. With stronger effects, smaller samples are required, but we do not face such effects on a regular basis. Anyway, if the size of the effect is known, it is possible to use one of the numerous online calculators to evaluate the required sample.

6.2 Considering Ethics

Serious games are developed for people and the experiments we conduct involve people. That means that research ethics in serious game research is basically the same as in psychological and social science research.

APA Code of Ethics [19] states five general principles, applicable to all the fields of psychology. There are (A) Beneficence and Nonmaleficence; (B) Fidelity and Responsibility; (C) Integrity; (D) Justice and (E) Respect for People's Rights and Dignity. Applied towards the experimental research paradigm, these rules can be summarized as follows: the researcher must respect the participants' wellbeing and the participation in the experiment must be physically and psychologically harmless, until the participant knowingly and willingly accepts the risk, if any. No force or threats are allowed to involve or keep the participant in the experiment. People must be allowed to discontinue their participation at any time if they want to.

The researcher is responsible for every possible outcome of the experiment as well as for the accuracy, honesty, and truthfulness of his research and scientific conclusions. In the research that involves digital games in general and serious games in particular, the participants tend to be less suspicious about the possible effects and side-effects and thus more prone to them, so the researcher takes the responsibility for the psychological outcomes. The experimenter needs to respect dignity and worth of all people and the rights of individuals, including privacy, and confidentiality. This means that the results of any experimental research should not involve any personal data of the participants without their informed consent.

A written informed consent might be useful for both the experimenter and the participant. Some experiments require the experimenter to actively hide the aim of the study or to withhold some principal information, but this must be done in the least harmful way. An ethic committee must be consulted and must approve the research plan to avoid harmful effects.

7 Experimental Designs

7.1 Between-Groups and Within-Subjects Experiments

There are two main types of the experimental designs: between-subjects design where we compare different groups of participants and within-subjects design, where the same participants are tested more than once. In both variants, there can be one or more than one independent variable with two or more levels in it. In the simplest case, however, there is a single independent variable with two levels in it.

Between-groups designs are necessary when certain levels of the independent variable give the participants some experience that would influence further research with other levels of the variable. In the serious game testing, this might be important if participants receive plot or strategy-related information, which can influence their further gameplay tactics. The main advantage of this scheme is that all the participants are so-called "naive subjects" with no previous experience with our game. The main disadvantages are the large amount of the participants required and the problem of the equivalent groups, which was discussed in the section about samples [24].

Within-subject design requires fewer participants than the between-groups scheme. The group equality is not a problem, due to the self-equality of all the participants. In this type of the experimental design, each participant meets all levels of the independent variable so that we can measure how his condition or knowledge changes. In cases other than experimental plus zero condition (bivalent independent variable) we need to keep in mind that there might be different types of sequence effects, which appear through trials. Speaking about the sequence effects, we mean that the order in which different experimental conditions are presented may influence the outcome of the research. Apart from practice and fatigue effects, this involves many other possible factors. For example, if we want to compare two educational serious games and we also have a boring online lecture as a control condition. If the order in which we present those three conditions in always the same, one of the games might benefit not because it is more effective, but because it contrasts the lecture. At the same time the lecture might lose some of its effectiveness simply because the participants are too aroused to concentrate after all those games.

With more than two levels of the independent variable, we need to use counterbalancing schemes to apply all possible sequences of variable level for at least once. Complete counterbalancing is possible for 3 or 4 conditions as there are six and twenty-four sequences respectively, but for more levels of the independent variable partial counterbalancing is needed. These schemes can be represented with the use of a balanced Latin square. A Latin square is a way to ensure that every condition of the study occurs equally often, precedes and follows every other condition exactly once.

An example of a balanced Latin square with 4 conditions is, as follows (different letters represent different conditions):

A	**B**	**D**	**C**
B	**C**	**A**	**D**
C	**D**	**B**	**A**
D	**A**	**C**	**B**

Other counterbalancing technics, such as reverse counterbalancing are also possible, though it seems rather hard to introduce a serious game experiment, requiring those experimental designs.

7.2 Examples of the Real Experimental Schemes in Serious Games Studies

Let's discuss a couple of real examples of experimental schemes.

The first one is the experiment from J. van Muidjen and colleagues' study of cognitive control in elderly people [52]. The study aimed first, to show that cognitive training games can improve cognitive control functions and second, to compare training with games to training with documentaries and quizzes. In this study, both experimental designs are applied in different parts of the research, due to multiple dependable variables. In the experiment, there were two independent groups of participants (the between-group scheme; bivalent variable) and controlled randomization was used to form the groups out of the general population, with groups being equal in age, educational level, IQ, and cognitive heath scores. There were two levels of the independent variable – the game condition and the documentary film condition. There were a pretest and posttest with nine cognitive tests (the scores form dependable variables). As the tests could interfere with each other results, the scheme was introduced to counterbalance the battery across the participants (like in within-subjects design). While the statistical hypothesis examined the differences between the groups in terms of test scores, the discussion and the final conclusion were made in terms of cognitive controls theory.

The second example is retrieved from the E.D. van der Spek's dissertation [51]. The author describes a variety of experiments held on different stages of a training serious game development. In one of the experiments three independent variable level were introduced (with no cues, auditory cues or visual cues). As the researcher used between-group scheme, there were three groups, with two balanced extraneous variables – gender and gaming experience of the participants. There were four dependent parameters to measure – three learning tests and an engagement questionnaire with pretest and posttest made. The research is especially interesting due to the results, which were unpredicted by the researcher and led to the experimental hypothesis rejection (though, alternative explanations were made out of the theoretical background).

8 If an Experiment Cannot Be Conducted

8.1 Quasi-Experiment

The quasi-experiment occurs, when instead of usual, manipulated independent variable we use a so-called ***subject variable*** [24]. Subject variables are already existing characteristics of the individuals, participating in the study. Gender, age, culture, level of intelligence, personality attributes and so on can be used as subject variables. For sure, we cannot manipulate sex or personality of our participants. Instead, we can select people with different levels of these variables into different groups to compare, if their reaction towards our serious game would be the same or not.

Alongside with the quasi-experimental scheme above, D.T. Campbell [11] introduces two more variants of the quasi-experimental design. One of these schemes is a between-group comparison without group randomization. The second one is the quasi-experimental scheme with a single group, where experimental condition changes might blend with a time factor.

In general, quasi-experiment shares most of the true experiment's advantages but has lesser control over any additional variables and influences. This means, that we should be especially accurate with extraneous variables on the one hand and that sometimes we will not be able to prove a causal hypothesis for sure on the other hand, as less control leads to the possibility of alternative explanations. Some quasi-experimental schemes, though, have better ecological validity than true experiments.

Quasi-experimental schemes sometimes can involve numerous groups and conditions. For example, in the flow and anxiety research in serious games [9] participated six different groups and there were five different conditions as well. A true experiment with such amount of variable levels would require enormous efforts. The authors introduce their research as an exploratory study, which means that the discussed problem is not clearly defined. In this case, the quasi-experimental scheme is more reasonable, as the research itself is not intended to prove causalities. Instead, it searches for relationships and succeeds. Please note, that quasi-experimental scheme still requires a lot of variable control, with pre- and posttest, different sampling technics and many efforts to manage equal timing for different groups and participants.

8.2 Correlational Study

While quasi-experiments (with some additional reservations and strict control schemes) can still be used to verify a causal hypothesis, the correlational study only shows the positive or negative linkage between certain parameters. Due to its simplicity, correlational studies are very common in psychology and social sciences. M.L. Raulin [43] notices that in the experimental design correlations of demographical variables are often used to point out possible confounding variables to enhance the control.

The correlational study requires to be mentioned in the experimental design chapter due to the fact that sometimes people tend to discuss correlations in the term of cause and effect. Such assumptions are methodologically wrong and considered to be non-scientific, as not only we cannot determine which variable precede the other, but also we cannot exclude the possibility of the third variable that is linked to both variables or causes their correlation in some other way.

In a validation study of serious games for clinical assessments [49], different correlational coefficients were used to obtain as much data as possible with different types of scales. While the authors use Mann-Whitney test as well, correlations are more important in this particular study, as it aims to find linkage between cognitive assessments and serious games. One of the goals is described as “develop a method for predicting the presence of delirium, using serious game”. As the correlation describes that some variables coexist with each other, it can support such a notion, though it does not show why this happens.

9 Conclusions

The chapter discusses the use of the experimental methodology in serious games research. The main advantage of the experimental method, that it does not share with any other empirical methods in modern science, is the possibility to testify causal hypotheses according to an approved scientific paradigm. This means, that the only way to prove empirically that a serious game causes some changes in the users knowledge, skills or awareness is to conduct an experiment within one the discussed schemes. The advantages and disadvantages of between-groups and within-subjects experimental designs are mentioned alongside with some real life examples of experimental serious games research, retrieved from scientific publications. The chapter encourages the readers to use experiments in their own research projects and helps them to understand the terminology of experimental research. It also points out some general mistakes, that students should avoid while practicing in serious games evaluation.

While scientifically the experiment is one of the best evaluation tools, it is also one of the most labor- and time-consuming ones. Without proper control of the variables, the researcher can overlook the serious game’s effect. What is more important (as it will be Type I error), the experimental research, influenced by biases, leads to a poor evaluation of the serious game. Faulty experiments (and “cause-and-effect” conclusions based on the wrong methods) cannot only damage the particular researcher’s reputation but the reputation of the serious games in general.

Many serious games are developed to help vulnerable groups of people, such as children and adults with disabilities, people with phobias, medical patients, elders with dementia, etc., so it is very important to foresee not only the positive effects but also the negative once. This means that experiments with people playing serious games must maintain the highest ethical standards and that is why the researcher needs to work together with an ethical committee.

You need to remember that before starting the experiment, you need to pass the whole way from theoretical background to the experimental hypothesis. Evaluation through the experiment is sometimes a very long process, and you will probably need more than one experiment to prove your hypotheses. You also need to find suitable statistical methods for your research and you need to understand how these methods work to avoid misinterpretations. If researchers are unfamiliar with these methods a cooperation with experts from psychology or Human Computer Interaction is recommended. However, the evidence of your serious game effectiveness obtained through the adequate experiment absolutely worth the efforts.

Further Reading

For more experiments on serious games, see:

- Van der Spek, E.D.: Experiments in Serious Game Design: a Cognitive Approach. Utrecht University Repository (Dissertation). Utrecht University, Utrecht (2011)

For deeper knowledge about experimental and other research types in social sciences and psychology, as well as for common statistical procedures see:

- Kantowitz, B.H., Roediger, H.L.III, Elmes, D.G.: Experimental Psychology. Wadsworth, Belmont (2009)
- Martin D.W.: Doing Psychology Experiments. Thompson Higher Education, Belmont (2008)
- Goodwin C.J. Research in Psychology: Methods and Design. Wiley, Danvers (2010)

To learn about online experiments' possibility, see:

- Reips, U.-D.: Standards for Internet-based Experimenting. In: Experimental Psychology. 49(9), 243–256 (2002)
- Reips, U.-D., Krantz, J.H.: Conducting True Experiments on the Web. In: Gosling, S., Johnson, J. (eds.) Advanced Methods for Conducting Online Behavioral Research, pp. 193–216. American Psychological Association, Washington DC (2010)

For basic knowledge about latent variables (variables, which can only be discovered through statistical procedures, very common in psychology and social sciences) see:

- Bollen, K.A.: Latent Variables in Psychology and the Social Sciences. In: Annu. Rev. Psychol. 53, 605–634 (2002)

For advanced knowledge in statistical methods (multivariate analysis of variance, multiple regression, etc.), see:

- Foster J., Barkus, E., Yavorsky, C.: Understanding and Using Advanced Statistics: A Practical Guide for Students. SAGE Puclications, London, Thousand Oaks, New Delhi (2006)

References

1. Adams, K.A., Lawrence, E.K.: Research Methods, Statistics and Applications. SAGE Publications, London (2014)
2. Adejumobi, B., Franck, N., Janzen, M.: Designing and testing a racing car serious game module. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 192–198. Springer, Heidelberg (2014)
3. Anderson, C.A., Dill, K.E.: Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. J. Person. Soc. Psychol. **78**(4), 772–790 (2000)

4. Anguera, J.A., Boccanfuso, J., Rintoul, J.L., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., Gazzaley, A.: Video game training enhances cognitive control in older adults. Nature **501**, 97–101 (2013)
5. Baek, Y., Ko, R., Marsh, T. (eds.): Trends and Applications of Serious Gaming and Social Media. Gaming Media and Social Effects. Springer, New York (2014)
6. Bidarra, R., Gambon, D., Kooij, R., Nagel, D., Schutjes, M., Tziouvara, I.: Gaming at the Dentist's–serious game design for pain and discomfort distraction. In: Schouten, B., Fedtke, S., Bekker, T., Schijven, M., Gekker, A. (eds.) Games for Health 2013. Proceeding of the 3rd Conference on Gaming and Playful Interaction in Healthcare, pp. 207–215. Springer, Heidelberg (2013)
7. Boinodiris, P., Fingar, P.: Serious Games for Business: Using Gamification to Fully Engage Customers, Employees and Partners. Meghan-Kiffer Press, Tampa (2014)
8. Botella, C., Breton-Lópeza, J., Queroa, S., Bañosb, R.M., García-Palaciosa, A., Zaragozac, I., Alcanizc, M.: Treating cockroach phobia using a serious game on a mobile phone and augmented reality exposure: a single case study. Comput. Hum. Behav. **27**(1), 217–227 (2011). Current Research Topics in Cognitive Load Theory. Third International Cognitive Load Theory Conference
9. Brom, C., Buchtová, M., Šisler, V., Děchtěrenko, F., Palme, R., Glenk, L.M.: Flow, social interaction anxiety and salivary cortisol responses in serious games: a quasi-experimntal study. Comput. Educ. **79**, 69–100 (2014)
10. Bushman, B.J., Anderson, C.: Violent video games and hostile expectations: a test of the general aggression model. Pers. Soc. Psychol. Bull. **28**(12), 1679–1686 (2002)
11. Campbell, D.T., Cook, D.T.: Quasy-Experimental Design and Analysis Issues for Field Setting. Rand McNally, Chicago (1979)
12. Casti, J.L.: Reality Rules, Picturing the World in Mathematics–The Fundamentals, vol. I. Wiley, New York (1997)
13. Choo, A., Tong, X., Gromala, D., Hollander, A.: Virtual reality and mobius floe: cognitive distraction as non-pharmacological analgesic for pain management. In: Schouten, B., Fedtke, S., Schijven, M., Vosmeer, M., Gekker, A. (eds.) Games for Health 2014 Proceeding of the 4th conference on gaming and playful interaction in healthcare, pp. 8–12. Springer Fachmedien Wiesbaden, Heidelberg (2014)
14. Adams, K.A.: Cram101 Textbooks Reviews, Just the facts101 Textbook Key Facts: Research Methods, Statistics, and Applications. Content Technologies (2014). http://cram101.com
15. De Freitas, S., Liarokapis, F.: Serious games: a new paradigm for education? In: Ma, M., Oikonomou, A., Jain, L.C. (eds.) Serious Games and Edutainment Applications, pp. 9–23. Springer, London (2011)
16. De Gloria, A., Bellotti, F., Berta, R.: Serious games for education and training. Int. J. Serious Games. **1**(1). http://dx.doi.org/10.17083/ijsg.v1i1.11
17. Djaouti, D., Alvarez, J., Jessel, J., Rampnoux, O.: Origins of Serious Games. In: Ma, M., Oikonomou, A., Jain, L.C. (eds.) Serious Games and Edutainment Applications, pp. 25–43. Springer, London (2011)
18. Essential facts about the computer and video game industry. 2016 Sales, Demographic and Usage Data. ESA (2016). http://www.essentialfacts.theesa.com/Essential-Facts-2016.pdf
19. Ethical Principles of Psychologisys and Code of Conduct. APA (2010). http://www.apa.org/ethics/code/
20. Experiment. In: Merriam-Webster.com. Merriam-Webster. http://www.merriam-webster.com/dictionary/experiment

21. Ferguson, C., Kilburn, J.: Much ado about nothing: the misestimation and overinterpretation of violent video game effects in eastern and western nations: a comment on Anderson, et al. Psychol. Bull. **136**(2), 174–178 (2010)
22. Foster, J., Barkus, E., Yavorsky, C.: Understanding and Using Advanced Statistics: A Practical Guide for Students. SAGE Puclications, London, Thousand Oaks (2006)
23. 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)
24. Goodwin, C.J.: Research in Psychology Methods and Design. Wiley, New York (2010)
25. Hess, M., Wiemeyer, J., Hamacher, K., Goesele, M.: Serious games for solving protein sequence alignments - combining citizen science and gaming. In: Göbel, S., Wiemeyer, J. (eds.) GameDays 2014. LNCS, vol. 8395, pp. 175–185. Springer, Heidelberg (2014)
26. Hypothesis. In: Merriam-Webster.com. Merriam-Webster. http://www.merriam-webster.com/dictionary/hypothesis
27. Kantowitz, B.H., Roediger, H.L., Elmes, D.G.: Experimental Psychlogy, vol. III. Wadsworth, Belmont (2009)
28. Kaufman, S.B.: Sex differences in mental rotation and spatial visualization ability: can they be accounted for by differences in working memory capacity? Intelligence **35**, 211–223 (2007)
29. King, D.L., Delfabbro, P., Griffiths, M.D.: The psychological study of video game players: methodological challenges and practical advice. Int. J. Mental Health Addict. **7**(4), 555–562 (2009)
30. Kornilova, T.V.: Experimental Psychology. Jurajt, Moscow (2014). (in Russian)
31. Kuhn, T.S.: The Structure of Scientific Revolutions. International Encyclopedia of United Science, vol. 2. The University of Chicago Press, Chicago (1962)
32. Lim, C., Jung, H.: A study on the military serious game. Adv. Sci. Technol. Lett. **39**, 73–77 (2013). Proceedings. SERSC 2013
33. Marchand, A., Hennig-Thurau, T.: Value creation in the video game industry: industry economics, consumer benefits, and research opportunities. J. Interact. Market. **27**, 141–157 (2013)
34. Marsden, J.: The essential checklist for making an awesome video game, According to Futurlab (2013). http://indiegames.com/2013/07/the_essential_checklist_for_ma_1.html
35. Martin, D.W.: Doing Psychology Experiments. Thompson Higher Education, Belmont (2008)
36. Martins, T., Araújo, M., Carvalho, V., Soares, F., Torrão, L.: PhysioVinci – a first approach on a physical rehabilitation game. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 1–9. Springer, Heidelberg (2014)
37. Martin-SanJosé, J.-F., Juan, M.-C., Segui, I., Garcia-Garcia, I.: The effects of computer-based games and collaboration in large groups vs collaboration in pair or traditional methods. Comput. Educ. **87**, 42–54 (2015)
38. Mayer, I., Zhou, Q., Keijser, X., Abspoel, L.: Gaming the future of the ocean: the marine spatial planning challenge 2050. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 150–162. Springer, Heidelberg (2014)
39. Omelina, L., Jansen, B., Bonnechère, B., Van Sint Jan, S., Cornelis, J.: Serious games for physical rehabilitation: designing highly configurable and adaptable games. In: Proceeding 9th International Conference on Disability, Virtual Reality and Associated Technologies Laval, France, 10–12 September 2012 (ICDVRAT 2012), pp. 195–201. ICDVRAT (2012)
40. Operationalization. Wikipedia. https://en.wikipedia.org/wiki/Operationalization
41. Popper, K.: Normal Science and its Dangers. In: Lakatos, I., Musgrave, A. (eds.) Criticism and the Growth of Knowledge, pp. 51–58. Cambridge University Press, Cambridge (1970)

42. Popper, K.: The Logic of Scientific Discovery. Routledge, London (2002)
43. Raulin, M.L., Graziano, A.M.: Quasi-experiments and correlational studies. In: Colman, A.M. (ed.) Companion Encyclopedia of Psychology, vol. 2, pp. 1124–1141. Routledge, London (1994)
44. Reliability. In: Merriam-Webster.com. Merriam-Webster. http://www.merriam-webster.com/dictionary/reliability
45. Ricciardi, F., De Paolis, L.T.: A Comprehensive review of serious games in health professions. Int. J. Comput. Games Technol. (2014). http://dx.doi.org/10.1155/2014/787968
46. Senevirathne, S.G., Kodagoda, M., Kadle, V., Haake, S.J., Senior, T., Heller, B.W.: Application of serious games to sports, health and exercise. In: Proceedings of the 6th SLIIT Research Symposium, Sri Lanka, vol. 4, pp. 6–9 (2011)
47. Shubert, T., Strobach, T.: Video game experience optimized executive control skills—on false positives and false negatives: reply to boot and simons. Acta Psychol. **141**, 278–280 (2012)
48. Tomé, R.M., Pereira, J.M., Oliveira, M.: Using serious games for cognitive disabilities. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 34–47. Springer, Heidelberg (2014)
49. Tong, T., Chignell, M., Tierney, M.C., Masella, C.: A Serious game for clinical assessment of cognitive status: validation study. JMIR Serious Games **4**(1), e7 (2016). doi:10.2196/games.5006
50. Validity. In: Merriam-Webster.com. Merriam-Webster. http://www.merriam-webster.com/dictionary/validity
51. Van der Spek, E.D.: Experiments in serious game design: a cognitive approach. Utrecht University Repository (Dissertation). Utrecht University, Utrecht (2011)
52. Van Muijden, J., Band, G.P.H., Hommel, B.: Online games training aging brains: limited transfer to cognitive control functions. Front. Hum. Neurosci. **6**, 221 (2012). doi:10.3389/fnhum.2012.00221
53. Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E.D.: A meta-analysis of the cognitive and motivational effects of serious games. J. Educ. Psychol. **105**, 249 (2013). doi:10.1037/a0031311
54. Wrzesien, M., Alcañiz, M., Botella, C., Burkhardt, J.-M., Lopez, J.B., Ortega, A.R.: A pilot evaluation of a therapeutic game applied to small animal phobia treatment. In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 10–20. Springer, Heidelberg (2014)
55. Yoo, K., Ahn, J., Lee, W.: A design of the stress relief game based on autonomic nervous system. In: Park, J.H., Jeong, Y.-S., Park, S.O., Chen, H.-C. (eds.) EMC 2012. LNEE, vol. 181, pp. 371–376. Springer, Heidelberg (2012)
56. Zakari, H.M., Ma, M., Simmons, D.: A review of serious games for children with autism spectrum disorders (ASD). In: Ma, M., Oliveira, M.F., Baalsrud Hauge, J. (eds.) SGDA 2014. LNCS, vol. 8778, pp. 93–106. Springer, Heidelberg (2014)