Replication of Software Engineering Experiments

Natalia Juristo and Omar S. Gómez*

Facultad de Informática,
Universidad Politécnica de Madrid,
Boadilla del Monte 28660, Madrid, España
natalia@fi.upm.es, ogomez@ieee.org

Abstract. Experimentation has played a major role in scientific advancement. Replication is one of the essentials of the experimental methods. In replications, experiments are repeated aiming to check their results. Successful replication increases the validity and reliability of the outcomes observed in an experiment.

There is debate about the best way of running replications of Software Engineering (SE) experiments. Some of the questions that have cropped up in this debate are, "Should replicators reuse the baseline experiment materials? Which is the adequate sort of communication among experimenters and replicators if any? What elements of the experimental structure can be changed and still be considered a replication instead of a new experiment?". A deeper understanding of the concept of replication should help to clarify these issues as well as increase and improve replications in SE experimental practices.

In this chapter, we study the concept of replication in order to gain insight. The chapter starts with an introduction to the importance of replication and the state of replication in ESE. Then we discuss replication from both the statistical and scientific viewpoint. Based on a review of the diverse types of replication used in other scientific disciplines, we identify the different types of replication that are feasible to be run in our discipline. Finally, we present the different purposes that replication can serve in Experimental Software Engineering (ESE).

Keywords: Experimental Replicaction, Types of Replication, Experimental Software Engineering, Empirical Software Engineering.

1 Introduction

Experimentation should be an indispensable part of SE research. As Tichy says [1], "Experimentation can help build a reliable base of knowledge and thus reduce uncertainty about which theories, methods, and tools are adequate". Basili [2] claims that "Experimental SE is necessary, common wisdom, intuition, speculation, and proofs of concepts are not reliable sources of credible knowledge".

* This work has been performed under research grant TIN 2008-00555 of the Spanish Ministry of Science and Innovation, and research grant 206747 of the México's National Council of Science and Technology (CONACyT).

B. Meyer and M. Nordio (Eds.): LASER Summer School 2008-2010, LNCS 7007, pp. 60–88, 2012.

Voices in favour of experimentalism as a way of research about software development have recently grown stronger. DeMarco [3] claims that "The actual software construction isn't necessarily experimental, but its conception is. And this is where our focus ought to be. It's where our focus always ought to have been". Meyer [4, 5] has also joined the line of researchers to point to the importance of experimentation in SE.

A key component of experimentation is replication. To consolidate a body of knowledge built upon experimental results, they have to be extensively verified. This verification is carried out by replicating an experiment to check if its results can be reproducible. If the same results are reproduced in different replications, we can infer that such results are regularities existing in the piece of reality under study. Experimenters acquainted with such regularities can find out mechanisms regulating the observed results or, at least, predict their behaviour.

Most of the events observed through experiments in SE nowadays are isolated. In other words, most SE experiments results have not been reproduced. So there is no way to distinguish the following three situations: the results were produced by chance (the event occurred accidentally); the results are artifactual (the event only occurs in the experiment not in the reality under study), or the results really do conform to a regularity of the piece of reality being examined.

A replication has some elements in common with its baseline experiment. When we start to examine a phenomenon experimentally, most aspects are unknown. Even the tiniest change in a replication can lead to inexplicable differences in the results. In immature experimental disciplines, which experimental conditions should be controlled can be found out by starting off with replications closely following the baseline experiment [6]. In the case of well-known phenomena, the experimental conditions that influence the results can be controlled, and artifactual results are identified by running less similar replications. For example, using different experimental protocols to verify the results correspond to experiment-independent events.

The immaturity of ESE has been an obstacle to replication. As the mechanisms regulating software development and the key experimental conditions for its investigation are yet unknown, even the slightest change in the replication leads to inexplicable differences in the results. However, context differences oblige experimenters to adapt the experiment. These changes can lead to sizeable differences in the replication results that prevent the outcomes of the baseline experiment from being corroborated. In several attempts at combining the results of ESE replications, Hayes [7], Miller [8–10], Hannay et al. [11], Jørgensen [12], Pickard et al. [13], Shull et al. [14] and Juristo et al. [15] reported that the differences between results were so large that they found it impossible to draw any consequences from the results comparison.

ESE stereotype of replication is an experiment that is repeated independently by other researchers at different sites to the baseline experiment. But some of the replications in ESE do not conform to this stereotype: either they are jointly run, or replicators researchers reuse some of the materials employed in the baseline experiment or they are run at the same site [16–25]. How replications should be

run has moved a debate in ESE. There are researchers that recommend reusing some of the baseline experiment materials to run replications [2, 26] with the aim of assuring that the replications are similar and results can be compared. There are researches who advise the use of different protocols and materials to those employed in the baseline experiment [10, 27] with the aim of preserving the principle of independence and preventing error propagation in replications that use the same materials. Others suggest using alternative ways of verifying the experimental results [28] with the aim of understanding the problems that replication have had to date in SE experiments. This debate can probably be put down to the fact that replication has still not satisfactorily tailored to ESE.

In this chapter we study the concept of replication with the aim of getting a better understanding of its use in ESE. This chapter is organized as follows. Section 2 describes the statistical perspective of replication. Section 3 discusses replication in science. Section 4 reviews different types of replication accepted in different experimental disciplines. Section 5 discusses the differences between the concepts of replication and reproduction. Section 6 describes adequate variations in replication. Section 7 discusses some types of replications in SE. Section 8 presents the purposes that a replication can serves. Section 9 presents the conclusions. Finally, Annex A lists and describes replication typologies found in other disciplines.

2 Statistical Perspective of Replication

Sample size is an essential element in a controlled experiment. An adequate sample size increases the possibilities of the effect observed in the sample occurring in the real population. The accuracy level of the results grows in proportion to the sample size.

One of the commonly used coefficients for representing effect size observed in an experiment is Cohen's d [29]. This coefficient is used to measure the differences between the treatments studied in the experiment. The effect size indicates how much better one treatment is compared to another. This coefficient is usually used with one-digit accuracy. For example [29], d=0.2 represents a small effect, d=0.5 indicates a medium effect or d=0.8 is a large effect. The sample size required to satisfy a one-digit accuracy level can be calculated from (1): the function in (1) is derived from (2) and (3), where the differences in the confidence intervals (left and right) are equal at the specified accuracy level, in this case 0.1.

$$N = \frac{2 + d^2}{2(0.0255102)^2} \tag{1}$$

$$2 \times 1.96 \times deviation(d) = 0.1 \tag{2}$$

$$deviation(d) = \sqrt{\frac{n1 + n2}{n1n2} + \frac{d^2}{2(n1 + n1)}} \tag{3}$$

For effect sizes d=0.2, d=0.5 and d=0.8, a sample size of N=1,567, N=1,729 and N=2,028 is required, respectively. Fig. 1 shows the graph of the resulting function in (1).

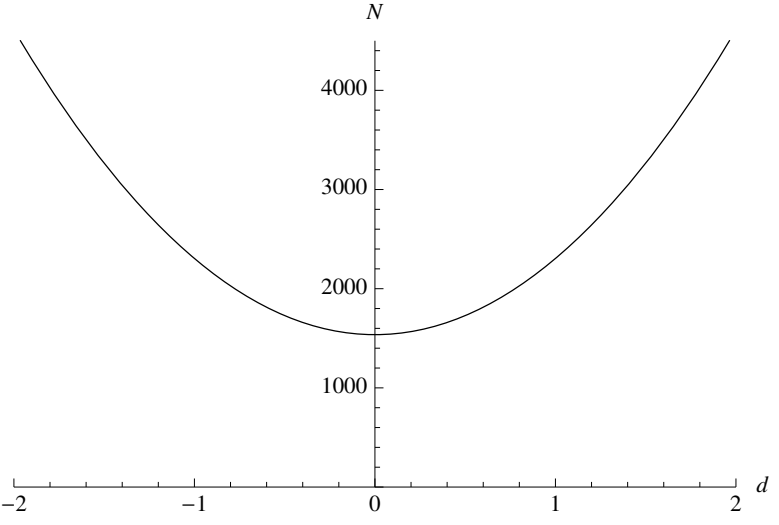

Fig. 1. Sample (N) necessary for a particular effect size (d) with one-digit accuracy

To be able to estimate effect sizes with one-digit accuracy, we need to repeat the same experiment to increase the sample size and reach the required level showed in Fig. 1. In the set of controlled SE experiments examined by Dybå et al. [30], the average sample size of the samples used in these experiments is N=55 (55 observations per experiment).

For an average sample size of 50 observations, the same study would have to be repeated 31 times to satisfy the sample size required for an effect size of d=0.2; the same study would have to be repeated 34 and 40 times, respectively, to get an effect size of d=0.5 and d=0.8. Consequently, experiment repetitions have to be equal. For increasing sample size the replications have to measure the independent and dependent variables in exactly the same manner, using exactly the same experimental protocol, and they should all sample the same populations [31].

Since experimental conditions are hard to control in ESE, one option worth considering to satisfy the statistical requirement of identical repetitions is running internal replications (at the same site and by the same experimenters) of SE experiments. Through internal repetitions, the sample comes closer to the interval of observations [1,537; 2,305] required to be confident that the observed effect (from 0, none; to 1[1], very large) occurs not only in the sample used in the experiment but also in the real population.

[1] Note that effect size over 1 is possible. In fact Kampenes et al. [32] show that 32% of the experiments published in SE have an effect size greater than 1. The bigger the effect size the bigger the sample.

The results of a single execution of an experiment is threatened by type I error[2]. Having more (internal) replications of the same experiment considerably reduces this type of error. For example, if an experimenter establishes the significance level α of an experiment at 0.05, which represents a 1:20 probability of obtaining a chance result, the likelihood of again obtaining an accidental result drops to 1:400 ($p = 0.05 \times 0.05 = 0.0025$) if the experiment is identically internally repeated again.

The sample size of experiments run in SE is not large enough to accurately estimate the effect size under study. Therefore, identical replications are required to be able to estimate the effect size with any accuracy. However, identical replications are virtually impossible when they are carried out in other sites [25].

3 Replication in Science

In science, replication refers to the repetition of a previously run experiment. Some definitions of replication in science are:

1. "Replication refers to a conscious and systematic repeat of an original study" [33].
2. "Replication is traditionally defined as the duplication of a previously published empirical study" [34].
3. "Replication is a methodological tool based on a repetition procedure that is involved in establishing a fact, truth or piece of knowledge" [35].
4. "Replication – the performance of another study statistically confirming the same hypothesis" [36].
5. "Replication is the repetition of the basic experiment. That is, replication involves repeating an experiment under identical conditions, rather than repeating measurements on the same experimental unit" [37].
6. "The deliberate repetition of research procedures in a second investigation for the purpose of determining if earlier results can be repeated" [38].
7. "Is the process of going back, or re-searching an observation, investigation, or experimentation to compare findings" [39].

The value of replication has been widely recognized in a number of scientific disciplines. Popper [40] claimed that "We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable". Hempel [41] realized the importance of reckoning with more than one

[2] Type I error occurs when the null hypothesis is rejected while it is true, i.e. when there is believed to be a significant difference between the treatments that the experiment compares and there is, in actual fact, no such difference.

study to increase the robustness of the gathered evidence. Campbell and Stanley [42] claim that "The experiments we do today, if successful, will need replication and cross-validation at other times under other conditions before they can become an established part of science, before they can be theoretically interpreted with confidence". Other widely accepted claims about replication are that it "is the Supreme Court of the scientific system" [43], it is considered the cornerstone of science [36], "it is the crucial test whereby theories and experiments in science are judged" [44], and "it is at the heart of (any) experimental science" [35].

From a scientific viewpoint, not having sufficient replications of an experiment can lead to the acceptance of results that are not robust enough. Fahs et al. [45] gave a good example of this problem in an article concerning the retinopathy of prematurity (ROP). Nurses working in neonatal intensive care units (NICUs) tend to place premature babies in incubators or try to somehow protect their eyes from the light, as this practice is believed to reduce the rate of ROP. This practice apparently dates back to a study by Glass et al. [46], concluding that ROP was possibly caused by the bright lighting in NICUs. Years later, however, Ackerman et al. [47] replicated this study and provided evidence contrary to the results published by Glass et al. [46]. Later another two replications of this study were run [48, 49] and corroborated the results reported by Ackerman et al. [47], i.e. NICU lighting is not a factor causing ROP.

Replications of experiments have proven the need to be careful about accepting evidence that has not been subject to strict checks. The evidence provided by a single study or experiment can be weak. Several replications have to be run to strengthen the evidence. In the field of SE, many of the empirical studies published have low statistical power [30]. Failure to replicate these experiments can lead to the belief that there is no significant effect when there probably is.

Even though replication is an important experimental mechanism, we have to be aware of its limits. It is not possible to completely verify a theory based on a finite series of observations. For example, someone observing three black crows at different times cannot conclude that all crows are black. To do this, s/he would have to observe all the crows of all times. Replication is closely related to induction[3], which has been used since ancient times as a way of inferring general rules from repeated past regular observations (instances) [43]. As Restivo [50] says, "replication is the experimental equivalent of induction; what is regularly reproducible is generalizable" or, as Collins [51] argues "experimental replication is the experimental equivalent of inductive inference".

Induction has a catch [40, 52–55] or logical defect, as the general conclusion is reached without individually evaluating all the cases. The problem of proving something inductively is that the gathered knowledge cannot be fully verified. Using probabilistic approaches [56–60], however, we can be somewhat confident about a conclusion reached based on a finite number of observations, that is, a hypothesis can be verified with some level of confidence based on a set of replications.

[3] Also known as inductive reasoning or inductive logic.

4 Replication Types in Other Disciplines

With the aim of discovering how to run a replication, we examined several types of replication used in other disciplines. We identified the different types of replication after running Google®, Google scholar®, ScienceDirect® and JSTOR® searches with different keywords (*types of replications*, *types of experimental replications*, *typology of replications*, *replication types*, *replication typologies*, *replication types* and *classification of replications*).

After running the searches on the four search engines and examining all the results returned, we located an initial set of 10 replication typologies [31, 35, 61–68]. This initial set of typologies served as a source for locating more replication types. Following the references in this initial set, we were able to locate another 8 [33, 69–75]. This way, we ended up with 18 replication typologies shown in Annex A. Altogether the typologies contain a total of 79 replication types. These typologies belong to the fields of social science (61%), business (33%) and philosophy (6%). Table 1 lists the typologies grouped by field.

Table 1. Typologies grouped by discipline

Area	Number of Typologies	References
(Social Science)		
Psychology	5	Lykken [69]; Hendrick [70]; Hunter [31]; Schmidt [35]; Kantowitz et al. [65]
Sociology	3	Finifter [71]; La Sorte [33]; Bahr et al. [62]
Economics	1	Mittelstaedt and Zorn [67]
Human Communication	1	Kelly et al. [72]
Human Development	1	Van IJzendoorn [63]
(Business)		
Marketing	3	Leone and Schultz [73]; Easley et al. [61]; Monroe [74]
Accounting	1	Lindsay and Ehrenberg [68]
Management	1	Tsang and Kwan [66]
Forecasting	1	Evanschitzky and Armstrong [64]
(Philosophy)		
Philosophy of Science	1	Radder [75]
TOTAL	18	

Lykken's [69] is the most often cited typology, followed by Hendrick's [70]. Lykken's [69], Hendrick's [70] and Bahr et al.'s [62] typologies have been referenced not only within their disciplines, but also in some business areas. We have counted citations where the author somehow uses the typology rather than referring to other questions that the above articles address.

In most typologies, the authors give the replication types an original name. They tend, therefore, to use their own terms to refer to a replication type. There are some exceptions, like Kelly et al. [72], who use the same terms as are applied in Lykken's [69] typology. In the identified typologies, we also find that there is no intra- or inter-disciplinary standardization for naming replication types.

The identified typologies were found to have two purposes: 1) some authors developed the typology to classify existing sets of replications; 2) other authors generated the typology for no particular purpose. Within this purpose, some authors illustrate the replication types using a number of existing replications, whereas others develop the typology and use examples to describe the replication types. Table 2 shows the possible usage of typologies.

Table 2. Typologies usage

Typologies generated to classify existing sets of replications	Typologies generated for understanding replication types	
	With examples of real replications	With imaginary examples
Bahr et al. [62]	Lindsay and Ehrenberg [68]	Hendrick [70]
Kelly et al. [72]	Tsang and Kwan [66]	Monroe [74]
Leone and Schultz [73]	Kantowitz et al. [65]	Radder [75]
Evanschitzky and Armstrong [64]	Lykken [69]	Easley et al. [61]
	Hendrick [70]	Hunter [31]
	La Sorte [33]	Schmidt [35]
	Van IJzendoorn [63]	Finifter [71]
		Mittelstaedt and Zorn [67]

Examining the typologies, we found that experiment results were not always verified by running the experiment over again. Neither did the replication always repeat the experimental protocol of the baseline experiment. We have identified three major groups of methods for verifying findings:

1. Follow the **same experimental protocol** used in the baseline experiment. The degree of similarity between the replication and the baseline experiment vary. For verification purpose some of the elements of the baseline experiment can be changed or modified in the replication. For example, Tsang and Kwan [66] use the term *empirical generalization* when the study is repeated on different populations. Monroe [74] uses the term *independent replication* when the study is repeated by different researchers.

   This type of replication is used for different purposes. According to Lykken [69], for example, the purpose of *operational replication* is to check that the experimental recipe produces the same results with another researcher. Tsang and Kwan's [66] *empirical generalization* purpose is to test the extent to which the study results are generalizable to other populations.

Most researchers use the term *replication* accompanied by an adjective to refer to this method of verification, e.g. *real replication*, *strict replication*, *close replication*. The adjective denotes the degree of change made to the structure of the experiment. Table 3 shows the replication types in this category.

Table 3. Using the same experimental protocol

Term	Author(s)
Close Replication	Lindsay and Ehrenberg [68]
Conceptual Replication	Hunter [31]; Monroe [74]
Demonstrated Replication	Monroe [74]
Differentiated Replication	Lindsay and Ehrenberg [68]
Direct Replication	Schmidt [35]; Kantowitz et al. [65]
Empirical Generalization	Tsang and Kwan [66]
Exact Replicarion	Van IJzendoorn [63]; Tsang and Kwan [66]
Experimental Replication	Leone and Schultz [73]
Generalization and Extension	Tsang and Kwan [66]
Independent Replication	La Sorte [33]; Monroe [74]
Instrumental Replication	Kelly et al. [72]
Literal Replication	Lykken [69]; Kelly et al. [72]
Nonexperimental Replication	Leone and Schultz [73]
Nonindependend Replication	Monroe [74]
Operational Replication	Lykken [69]; Kelly et al. [72]
Partial Replication	Hendrick [70]; Monroe [74]
Real Replications	Evanschitzky and Armstrong [64]
Reproducibility of an experiment under a fixed theoretical interpretation	Radder [75]
Reproducibility of the material realization of an experiment	Radder [75]
Retest Replication	La Sorte [33]
Scientific Replication	Hunter [31]
Sequential Replication	Monroe [74]
Statistical Replication	Hunter [31]
Strict Replication	Hendrick [70]; Monroe [74]
Systematic Replication	Kantowitz et al. [65]; Finifter [71]
Types 0, I, II	Easley et al. [61]
Types A..H	Bahr et al. [62]
Varied Replication	Van IJzendoorn [63]
Virtual Replication	Finifter [71]

2. Use a **different experimental protocol** to the baseline experiment. In this type of verification, the only thing the replication has in common with the baseline experiment is that they are both based on the same theoretical structure, i.e. they share the same constructs. This verification is used to corroborate previously observed findings through a different path. Hendrick [70], Schmidt [35] and Kantowitz et al. [65] call this type of verification *conceptual replication*, whereas Finifter [71] names it *systematic replication*. Radder [75], describes it as the *reproducibility of the result of an experiment*. Table 4 shows the replication types that adhere to this kind of verification.

Table 4. Using a different experimental protocol

Replication Type	Author(s)
Conceptual Extension	Tsang and Kwan [66]
Conceptual Replication	Hendrick [70]; Schmidt [35]; Kantowitz et al. [65]
Constructive Replication	Lykken [69]; Kelly et al. [72]
Corroboration	Leone and Schultz [73]
Differentiated Replication	Lindsay and Ehrenberg [68]
Generalization and Extension	Tsang and Kwan [66]
Reproducibility of the result of an experiment	Radder [75]
Systematic Replication	Finifter [71]
Theoretical Replication	La Sorte [33]
Type III	Easley et al. [61]
Types I..P	Bahr et al. [62]

3. Use **existing data sets** from a previous experiment to reanalyse the data employing either the same analysis procedures or others. This modus operandi is useful for verifying whether errors were made during the data analysis stage or whether the outcomes are affected by any particular data analysis technique. Some replication types reanalyse the statistical models instead of the existing study data. Different names are used for this type of verification. For example, La Sorte [33] calls it *internal replication*; Finifter [71] terms it *pseudoreplication*, and Tsang and Kwan [66] describe it as *checking of analysis* and *reanalysis of data*. Table 5 shows the replication types we identified that fall into this category.

Table 5. Reanalyzing existing data

Replication Type	Author(s)
Checking of Analysis	Tsang and Kwan [66]
Complete Secondary Analysis	Van IJzendoorn [63]
Data Re-analyses	Evanschitzky and Armstrong [64]
Internal Replication	La Sorte [33]
Pseudoreplication	Finifter [71]
Reanalysis of Data	Tsang and Kwan [66]
Restricted Secondary Analysis	Van IJzendoorn [63]
Types I, II	Mittelstaedt and Zorn [67]

If we want one term to identify each of the three forms of verification, we would surely refer to the third one as *re-analysis*, because the descriptions clearly allude to this term. However, the naming of the other two forms causes some confusion. Do both forms adhere to the concept of replication, or does each one introduce a different concept? The authors of some of the articles that we consulted to identify the typologies use the terms replication and reproduction indistinctly. This led us to examine whether these two concepts are equivalent or different.

5 Replication vs. Reproduction

According to the typologies we found, most researchers use the term replication to refer to the repetition of an experiment, although some use the term reproduction or reproducibility to describe this repetition. So it seems that many researchers consider the two terms to be synonyms. Likewise, Wikipedia uses these terms indistinctly and defines reproducibility as "one of the main principles of the scientific method, and refers to the ability of a test or experiment to be accurately reproduced, or replicated, by someone else working independently" [76].

Some researchers, however, do make a distinction between the two terms. Cartwright [77], for example, suggests that replicability "doing the same experiment again" should be distinguished from reproducibility "doing a new experiment". For Cartwright [77] the replication of an experiment refers to repeating a new experiment very closely following the experimental protocol used in the previous experiment, whereas reproduction refers re-examining a previously observed result using a different experimental protocol to what was employed in the previous experiment.

According to Cartwright [77], replication does not guarantee that the observed result represents the reality under observation. The result can be artifactual, i.e. a product of the materials or the instruments used in the experiment. To guarantee that the result is consistent with the reality under observation, we have to undertake a reproduction using different experimental protocols to ensure that the observed result is independent of the procedure, materials or instruments used in the experiments that arrived at the result.

When the results are repeatable using the same experimental protocol, the experimenters can be confident that they have observed some sort of phenomenon that is stable enough to be observed more than once. But, as it was observed using the same experimental protocol, there could be a very close relationship between the protocol and the phenomenon. As Radder put it [78], "[this result] does not imply any agreement about what the phenomenon is. Some interpreters may even argue that the phenomenon is an artifact, because, though it is stable, it is not to be attributed to the object under study but to certain features of the apparatus", where the term apparatus refers to the instruments, materials or procedures used, i.e. the experimental protocol. Cartwright [77] claims that "reproducibility, then, is a guard against errors in our instruments" in such a situation. According to Cartwright [77], though, reproduction is not absolutely necessary, as the better designed the instruments (apparatus) are, the less likely it is to have to use reproducibility.

Reproduction can be seen as a sort of triangulation, where the experimenters use different experimental protocols in an attempt to validate or corroborate the findings of the previous experiment [79]. According to Park [80], "These triangulation strategies can be used to support a conceptual finding, but they are not replications of any degree".

In this respect, the concept of replication given by Cartwright [77] would fit the first form of verification described in the previous section, whereas the concept of reproduction adheres to the second form of verification that we identified in the replication typologies.

6 Variation among Replications

Based on the different replication types that we have found, replications appear to fall into three groups:

1. Replications that vary little or not at all with respect to the baseline experiment.
2. Replications that do vary but still follow the same experimental protocol as the baseline experiment.
3. Replications that use different experimental protocol to check the baseline experimental results i.e. reproductions.

Tables 6 and 7 list the replication types that fall into these first two groups. The third group corresponds with the second type of verification presented in section 4 (Use a different experimental protocol to the baseline experiment).

Table 6. Replications with few or no variations that adhere to the baseline experiment

Replication Type	Author(s)
Close Replication	Lindsay and Ehrenberg [68]
Direct Replication	Schmidt [35]; Kantowitz et al. [65]
Exact Replication	Van IJzendoorn [63]; Tsang and Kwan [66]
Experimental Replication	Leone and Schultz [73]
Literal Replication	Lykken [69]; Kelly et al. [72]
Real Replications	Evanschitzky and Armstrong [64]
Reproducibility of the material realization of an experiment	Radder [75]
Sequential Replication	Monroe [74]
Statistical Replication	Hunter [31]
Strict Replication	Hendrick [70]; Monroe [74]
Type 0	Easley et al. [61]
Types A..D	Bahr et al. [62]
Type I	Easley et al. [61]

Based on the descriptions of the replications, it appears that a replication can have different levels of similarity to the baseline experiment. In other words, the elements of the experiment structure do not necessarily have to be the same in the replication. Table 8 shows some experimental elements that, according to the typologies we have found, do not necessarily have to be the same in each replication. Note that the type (or aim) of the replication differs depending on this change.

Table 7. Replications with variations that adhere to the same experimental protocol

Replication Type	Author(s)
Conceptual Replication	Hunter [31]; Monroe [74]
Demonstrated Replication	Monroe [74]
Differentiated Replication	Lindsay and Ehrenberg [68]
Direct Replication	Schmidt [35]
Empirical Generalization	Tsang and Kwan [66]
Generalization and Extension	Tsang and Kwan [66]
Independent Replication	La Sorte [33]; Monroe [74]
Instrumental Replication	Kelly et al. [72]
Nonexperimental Replication	Leone and Schultz [73]
Nonindependent Replication	Monroe [74]
Operational Replication	Lykken [69]; Kelly et al. [72]
Partial Replication	Hendrick [70]; Monroe [74]
Reproducibility of an experiment under a fixed theoretical interpretation	Radder [75]
Retest Replication	La Sorte [33]
Scientific Replication	Hunter [31]
Sequential Replication	Monroe [74]
Systematic Replication	Kantowitz et al. [65]; Finifter [71]
Types E..H	Bahr et al. [62]
Type II	Easley et al. [61]
Varied Replication	Van IJzendoorn [63]
Virtual Replication	Finifter [71]

Table 8. Some identified elements that can vary in the replication

Variable element	Replication Type	Author(s)
Measurement instruments	Differentiated Replication	Lindsay and Ehrenberg [68]
Measures	Operational Replication	Kelly et al. [72]
Method	Conceptual Replication	Schmidt [35]
Place	Types B,F,J,N,D,H,L,P	Bahr et al. [62]
Populations	Empirical Generalization	Tsang and Kwan [66]
Research Design	Retest Replication	La Sorte [33]
Researcher	Independent Replication	Monroe [74]
Sample	Virtual Replication	Finifter [71]

Although the overall objective of a replication is to check an experimental result, we find that different replication types have specific aims or purposes. For example, according to Lykken [69], the purpose of *operational replication* is to check that the experimental recipe outputs the same results with another researcher. However, Finifter's *systematic replication* [71] aims to output new findings using different methods to the baseline experiment.

Each specific aim of a replication type denotes an aspect of the experiment that needs to be verified. The more experimental aspects or elements are verified, the greater the confidence that the observed effect is not artifactual. An effect observed in an experiment may not be observed at sites other than where it was replicated, by other researchers, using other materials or methods or under other conditions. Different replication types should be run to check that the different experiment elements do not bias the observed findings and that the experiment results are real.

Consequently, there are several degrees of similarity between a replication and the baseline experiment. The changes serve different replication purposes. Although the general purpose of a replication is to check a previously observed finding, each replication type has special goals depending on what specific element of the experiment is to be checked.

7 Types of Replications in SE

We did not find any specific research aiming to build a typology or classification of replications in the field of ESE. We did locate, however, three works in our discipline that classified replications as part of the research conducted.

The first piece of research is a master's thesis [81] that set out to study the use of replication of controlled experiments in ESE. Almqvist [81] surveys 44 articles describing 51 controlled experiments and 31 replications. He runs a systematic review as a method for identifying relevant articles. In Chapter 4 of the thesis, Almqvist [81] defines several categories for organizing the identified experiments. In one of the categories, he develops a classification for categorizing the identified replications. Almqvist takes the replication types described by Lindsay and Ehrenberg [68] as a reference and adds *internal* and *external replication*. On this basis, he defines the following four types of replications:

1. Similar-external replications.
2. Improved-internal replications.
3. Similar-internal replications.
4. Differentiated-external replications.

The second classification is found in an article by Basili et al. [2], presenting a framework for organizing sets of related studies. This article describes the different aspects of the framework being one of these aspects a classification of replications composed of three major categories, where two of these categories define several types of replications. Basili et al. [2] illustrate the classification with examples of different replications that they have run. The classification is composed of a total of six replication types:

1. Strict replications.
2. Replications that vary the manner in wich the experiment is run.
3. Replications that vary variables intrinsic to the object of study.
4. Replications that vary variables instrinsic to the focus of the evaluation.

5. Replications that vary context variables in the environment in wich the solution is evaluated.
6. Replications that extend the theory.

The third classification is found in a research conducted by Krein and Knutson [82]. The paper presents a framework for organizing research methods in SE. Krein and Knutson [82] define a replication taxonomy with four types of replications:

1. Strict replication. Which is meant to replicate a prior study as precisely as possible.
2. Differentiated replication. Which intentionally alters aspects of the prior study in order to test the limits of that study's conclusions.
3. Dependent replication. Which is a study that is specifically designed with reference to one or more previous studies, and is, therefore, intended to be a replication study.
4. Independent replication. Which addresses the same questions and/or hypotheses of a previous study, but is conducted without knowledge of, or deference to, that prior study either because the researchers are unaware of the prior work, or because they want to avoid bias.

Other ESE works mention replication types but do not refer to any classification. For example, Brooks et al. [83] and Mendonça et al. [84] mention differences between *internal* and *external replication*. Shull et al. [26] discuss some types of replications (*exact*, *independent*, *dependent* and *conceptual replications*) to describe the role that they play in ESE. Finally, Lung et al. [85] mention two types of replication (*literal* and *theoretical replication*) to explain the type of replication that they ran, and Mandić et al. [86] discuss two types of replications, namely, *exact* or *partial replications*, and replications designed to improve the goal of the original experiment.

8 Purposes of Replication in ESE

The elements of an experiment to be replicated vary depending on the purpose of the replication. We have identified five elements that can vary in a replication:

1. *Experimenters*. The experimenters in a replication can be the same people as participated in the baseline experiment, different experimenters or a mixture of both, though some cooperation between the baseline experiment researchers and the replicators.
2. *Site*. The replication can be run at the same site as the baseline experiment or at another place.
3. *Experimental Protocol*. This term refers to the experimental design, instruments, materials, experimental objects, forms and procedures used to run an experiment. The experimental protocol is how these elements are set up for use by the experimenter to observe the effects of the treatments. Different elements of the protocol can be changed in a replication.

4. *Construct Operationalizations.* Operationalizations describe the act of translating a construct into its manifestation. In a controlled experiment we have cause and effect operationalizations. The cause operationalizations represent the primary treatments to be evaluated in the experiment (independent variables) whereas the effect operationalizations represent the response variables (dependent variables) used to measure the effects of the treatments. Both types of operationalization contain elements that can be varied in a replication.
5. *Population Properties.* In SE experiments there are at least two populations that are worth generalizing: the subjects and the experimental objects with which subjects work or interact during the experiment. The generalization takes place when the replication changes the properties of the subject or the experimental objects.

Based on the elements that may vary in a replication, we identify the following purposes of a replication in ESE:

1. *Control for Sampling Error.* If the basic elements of the baseline experiment structure are kept unchanged, the purpose of the replication is to verify that the results output by that experiment are not chance outcomes. This function is useful for verifying that the effect identified in the baseline experiment is not due to a Type-I error.
2. *Control for Experimenters.* If different experimenters run the replication, then it aims is to verify that the experimenters do not influence the results.
3. *Control for Site.* If the replication is run at another site, then it aims is to verify that the results are independent of the site where the experiment is run.
4. *Control for Artifactual Results.* If the experimental protocol is changed, the purpose of the replication is to verify that the observed results are not artifactual, that is, they reflect reality and are not a product of the experimental protocol setup.
5. *Determine Limits for Operationalizations.* If the operationalizations are changed a replication aims to determine the range of variation of the primary treatments (independent variables) and the measures (dependent variables) used to gauge the effects of the treatments.
6. *Determine Limits in the Population Properties.* If the population properties are changed, the purpose of the replication is to determine the types of experimental subject or objects to which the results of the replication hold.

9 Conclusions

Replication plays an important role in scientific progress where facts are at least as important as ideas [31]. Experiments have to be replicated to identify evidences. If we want to build up a SE body of knowledge based on empirical evidence, different types of replications have to be run. In this chapter we have

studied the concept of replication as it is used in other scientific disciplines with the aim of getting a better understanding of this mechanism.

Although we identified several replication typologies, replication types are not standardised at either the intra or interdisciplinary level. Some authors use the same replication name, although they each define the replication differently. Also authors use different replication names to refer to equivalent replications types.

Several of the different replication types that we have found describe changes of the structure of the experiment to be replicated. That is, replication can have different levels of similarity to the baseline experiment. The changes to the experiment in a replication are linked with the verification purposes. Although the aim of a replication is to verify the experimental outcomes, a replication has specific purposes depending on which elements in the experiment are varied.

All different replication purposes have to be reached and satisfied in order for an experiment result to be considered verified. A systematic approach where different types of replications are planified can help experimenters to advance step by step in the verification path.

Discovering new conditions influencing the results of the experiments (and thus software development) is an important co-lateral effect of replications. With a better understanding of these conditions, we will be able to assemble the small segments learnt in systematically varied replications to put together a piece of knowledge.

References

1. Tichy, W.: Should Computer Scientists Experiment more? Computer 31(5), 32–40 (1998)
2. Basili, V., Shull, F., Lanubile, F.: Building Knowledge through Families of Experiments. IEEE Transactions on Software Engineering 25(4), 456–473 (1999)
3. DeMarco, T.: Software Engineering: An Idea Whose Time has Come and Gone? IEEE Software 26(4), 95–96 (2009)
4. Meyer, B.: Credible Objective Answers to Fundamental Software Engineering Questions. LASER Summer School on Software Engineering (2010)
5. Meyer, B.: Empirical Research: Questions from Software Engineering. In: 4th International Symposium on Empirical Software Engineering and Measurement (ESEM 2010) (2010)
6. Brinberg, D., McGrath, J.E.: Validity and the Research Process, p. 176. Sage Publications, Inc. (June 1985)
7. Hayes, W.: Research Synthesis in Software Engineering: A Case for Meta-Analysis. In: METRICS 1999: Proceedings of the 6th International Symposium on Software Metrics, p. 143. IEEE Computer Society (1999)
8. Miller, J.: Can Results from Software Engineering Experiments be Safely Combined? In: METRICS 1999: Proceedings of the 6th International Symposium on Software Metrics, p. 152. IEEE Computer Society (1999)
9. Miller, J.: Applying Meta-analytical Procedures to Software Engineering Experiments. J. Syst. Softw. 54(1), 29–39 (2000)
10. Miller, J.: Replicating Software Engineering Experiments: A poisoned Chalice or the Holy Grail. Information and Software Technology 47(4), 233–244 (2005)

11. Hannay, J., Dybå, T., Arisholm, E., Sjøberg, D.: The Effectiveness of Pair Programming: A Meta-analysis. Information and Software Technology, Special Section: Software Engineering for Secure Systems 51(7), 1110–1122 (2009)
12. Jørgensen, M.: A Review of Studies on Expert Estimation of Software Development Effort. Journal of Systems and Software 70(1-2), 37–60 (2004)
13. Pickard, L., Kitchenham, B., Jones, P.: Combining Empirical Results in Software Engineering. Information and Software Technology 40(14), 811–821 (1998)
14. Shull, F., Basili, V., Carver, J., Maldonado, J., Travassos, G., Mendonça, M., Fabbri, S.: Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem. In: SESE 2002: Proceedings of the 2002 International Symposium on Empirical Software Engineering, p. 7. IEEE Computer Society (2002)
15. Juristo, N., Moreno, A., Vegas, S.: Reviewing 25 Years of Testing Technique Experiments. Empirical Softw. Engg. 9(1-2), 7–44 (2004)
16. Basili, V., Selby, R.: Comparing the Effectiveness of Software Testing Strategies. IEEE Trans. Softw. Eng. 13(12), 1278–1296 (1987)
17. Porter, A., Votta, L., Basili, V.: Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment. IEEE Trans. Softw. Eng. 21(6), 563–575 (1995)
18. Fusaro, P., Lanubile, F., Visaggio, G.: A Replicated Experiment to Assess Requirements InspectionTechniques. Empirical Softw. Engg. 2(1), 39–57 (1997)
19. Miller, J., Wood, M., Roper, M.: Further Experiences with Scenarios and Checklists. Empirical Software Engineering 3(1), 37–64 (1998)
20. Sandahl, K., Blomkvist, O., Karlsson, J., Krysander, C., Lindvall, M., Ohlsson, N.: An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections. Empirical Software Engineering 3(4), 327–354 (1998)
21. Porter, A., Votta, L.: Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects. Empirical Software Engineering 3(4), 355–379 (1998)
22. Wood, M., Roper, M., Brooks, A., Miller, J.: Comparing and Combining Software Defect Detection Techniques: A Replicated Empirical Study. SIGSOFT Softw. Eng. Notes 22(6), 262–277 (1997)
23. Juristo, N., Vegas, S.: Functional Testing, Structural Testing, and Code Reading: What Fault Type Do They Each Detect? ESERNET, 208–232 (2003)
24. Vegas, S., Juristo, N., Moreno, A., Solari, M., Letelier, P.: Analysis of the Influence of Communication between Researchers on Experiment Replication. In: ISESE 2006: Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, pp. 28–37. ACM (2006)
25. Juristo, N., Vegas, S.: Using Differences among Replications of Software Engineering Experiments to Gain Knowledge. In: ESEM 2009: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 356–366. IEEE Computer Society (2009)
26. Shull, F., Carver, J., Vegas, S., Juristo, N.: The Role of Replications in Empirical Software Engineering. Empirical Softw. Engg. 13(2), 211–218 (2008)
27. Kitchenham, B.: The Role of Replications in Empirical Software Engineering – A Word of Warning. Empirical Softw. Engg. 13(2), 219–221 (2008)
28. Miller, J.: Triangulation as a Basis for Knowledge Discovery in Software Engineering. Empirical Softw. Engg. 13(2), 223–228 (2008)
29. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. L. Erlbaum Associates (1988)

30. Dybå, T., Kampenes, V., Sjøberg, D.: A Systematic Review of Statistical Power in Software Engineering Experiments. Information and Software Technology 48(8), 745–755 (2006)
31. Hunter, J.: The Desperate Need for Replications. Journal of Consumer Research 28(1), 149–158 (2001)
32. Kampenes, V., Dybå, T., Hannay, J., Sjøberg, D.: A Systematic Review of Effect Size in Software Engineering Experiments. Information and Software Technology 49(11-12), 1073–1086 (2007)
33. La Sorte, M.A.: Replication as a Verification Technique in Survey Research: A Paradigm. The Sociological Quarterly 13(2), 218–227 (1972)
34. Singh, K., Ang, S.H., Leong, S.M.: Increasing Replication for Knowledge Accumulation in Strategy Research. Journal of Management 29(4), 533–549 (2003)
35. Schmidt, S.: Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. Review of General Psychology 13(2), 90–100 (2009)
36. Moonesinghe, R., Khoury, M.J., Janssens, A.C.: Most Published Research Findings Are False – But a Little Replication Goes a Long Way. PLoS Med. 4(2), 218–221 (2007)
37. Pfleeger, S.L.: Experimental Design and Analysis in Software Engineering: Part 2: how to set up and experiment. SIGSOFT Softw. Eng. Notes 20(1), 22–26 (1995)
38. Polit, D.F., Hungler, B.P.: Nursing Research: Principles and Methods, p. 816. Lippincott Williams & Wilkins (1998)
39. Berthon, P., Pitt, L., Ewing, M., Carr, C.L.: Potential Research Space in MIS: A Framework for Envisioning and Evaluating Research Replication, Extension, and Generation. Info. Sys. Research 13, 416–427 (2002)
40. Popper, K.: The Logic of Scientific Discovery. Hutchinson & Co. (1959)
41. Hempel, C.G.: Philosophy of Natural Science. Prentice-Hall (1962)
42. Campbell, D.T., Stanley, J.C.: Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin Company (June 1963)
43. Collins, H.M.: Changing Order: Replication and Induction in Scientific Practice. Sage Publications (1985)
44. Broad, W., Wade, N.: Betrayers Of The Truth, Fraud and Deceit in the Halls of Science. Simon & Schuster, Inc. (1982)
45. Fahs, P.S., Morgan, L.L., Kalman, M.: A Call for Replication. Journal of Nursing Scholarship 35(1), 67–72 (2003)
46. Glass, P., Avery, G.B., Subramanian, K.N.S., Keys, M.P., Sostek, A.M., Friendly, D.S.: Effect of Bright Light in the Hospital Nursery on the Incidence of Retinopathy of Prematurity. New England Journal of Medicine 313(7), 401–404 (1985)
47. Ackerman, B., Sherwonit, E., Williams, J.: Reduced Incidental Light Exposure: Effect on the Development of Retinopathy of Prematurity in Low Birth Weight Infants. Pediatrics 83(6), 958–962 (1989)
48. Reynolds, J.D., Hardy, R.J., Kennedy, K.A., Spencer, R., van Heuven, W., Fielder, A.R.: Lack of Efficacy of Light Reduction in Preventing Retinopathy of Prematurity. New England Journal of Medicine 338(22), 1572–1576 (1998)
49. Seiberth, V., Linderkamp, O., Knorz, M.C., Liesenhoff, H.: A Controlled Clinical Trial of Light and Retinopathy of Prematurity. Am. J. Ophthalmol. 118(4), 492–495 (1994)
50. Restivo, S.: Science, Technology, and Society: An Encyclopedia, p. 728. Oxford University Press (May 2005)
51. Collins, H.: The Experimenter's Regress as Philosophical Sociology. Studies in History and Philosophy of Science Part A 33, 149–156(8) (2002)

52. Hume, D.: An Enquiry Concerning Human Understanding (1749)
53. Hempel, C.G.: Studies in the Logic of Confirmation (I.). Mind 54(213), 1–26 (1945)
54. Good, I.: The White Shoe Is A Red Herring. British Journal for the Philosophy of Science 17(4), 322 (1967)
55. Goodman, N.: Fact, Fiction, and Forecast. Harvard University Press (1955)
56. Bayes, T.: An Essay towards solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London (1763)
57. Fisher, R.A.: The Design of Experiments. Oliver & Boyd (1935)
58. Neyman, J.: First Course in Probability and Statistics. Henry Holt (1950)
59. Rivadula, A.: Inducción, Deducción y Decisión en las Teorías Estadísticas de la Inferencia Científica. Revista de Filosofía 9, 3–14 (1993)
60. Singh, G.: A Shift from Significance Test to Hypothesis Test trough Power Analysis in Medical Research. Journal of Postgraduate Medicine 52(2), 148–150 (2006)
61. Easley, R., Madden, C., Dunn, M.: Conducting Marketing Science: The Role of Replication in the Research Process. Journal of Business Research 48(1), 83–92 (2000)
62. Bahr, H.M., Caplow, T., Chadwick, B.A.: Middletown III: Problems of Replication, Longitudinal Measurement, and Triangulation. Annu. Rev. Sociol 9(1), 243–264 (1983)
63. Van IJzendoorn, M.H.: A Process Model of Replication Studies: On the Relation between Different Types of Replication. Leiden University Library (1994)
64. Evanschitzky, H., Armstrong, J.S.: Replications of Forecasting Research. International Journal of Forecasting 26(1), 4–8 (2010)
65. Kantowitz, B.H., Roediger III, H.L., Elmes, D.G.: Experimental Psychology, p. 592. Wadsworth Publishing (1984)
66. Tsang, E., Kwan, K.-M.: Replication and Theory Development in Organizational Science: A Critical Realist Perspective. The Academy of Management Review 24(4), 759–780 (1999)
67. Mittelstaedt, R., Zorn, T.: Econometric Replication: Lessons from the Experimental Sciences. Quarterly Journal of Business & Economics 23(1) (1984)
68. Lindsay, R.M., Ehrenberg, A.S.C.: The Design of Replicated Studies. The American Statistician 47(3), 217–228 (1993)
69. Lykken, D.T.: Statistical Significance in Psychological Research. Psychol. Bull. 70(3), 151–159 (1968)
70. Hendrick, C.: Replications, Strict Replications, and Conceptual Replications: Are They Important?, pp. 41–49. Sage, Newbury Park (1990)
71. Finifter, B.: The Generation of Confidence: Evaluating Research Findings by Random Subsample Replication. Sociological Methodology 4, 112–175 (1972)
72. Kelly, C., Chase, L., Tucker, R.: Replication in Experimental Communication Research: an Analysis. Human Communication Research 5(4), 338–342 (1979)
73. Leone, R., Schultz, R.: A Study of Marketing Generalizations. The Journal of Marketing 44(1), 10–18 (1980)
74. Monroe, K.B.: Front Matter. The Journal of Consumer Research 19(1) pp. i–iv (1992)
75. Radder, H.: Experimental Reproducibility and the Experimenters' Regress. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1, 63–73 (1992)
76. Wikipedia: Reproducibility — Wikipedia, The Free Encyclopedia (2009)
77. Cartwright, N.: Replicability, Reproducibility, and Robustness: Comments on Harry Collins. History of Political Economy 23(1), 143–155 (1991)

78. Radder, H.: In and About the World: Philosophical Studies of Science and Technology, p. 225. State University of New York Press, Albany (1996)
79. Easterbrook, S., Singer, J., Storey, M., Damian, D.: Selecting Empirical Methods for Software Engineering Research. In: Guide to Advanced Empirical Software Engineering, pp. 285–311. Springer, Heidelberg (2008)
80. Park, C.L.: What Is The Value of Replicating other Studies? Research Evaluation 13, 189–195(7) (2004)
81. Almqvist, J.P.F.: Replication of Controlled Experiments in Empirical Software Engineering – A Survey (2006)
82. Krein, J.L., Knutson, C.D.: A Case for Replication: Synthesizing Research Methodologies in Software Engineering. In: 1st International Workshop on Replication in Empirical Software Engineering Research, RESER 2010 (2010)
83. Brooks, A., Daly, J., Miller, J., Roper, M., Wood, M.: Replication of experimental results in software engineering. Number ISERN–96-10 (1996)
84. Mendonça, M., Maldonado, J., de Oliveira, M., Carver, J., Fabbri, S., Shull, F., Travassos, G., Höhn, E., Basili, V.: A Framework for Software Engineering Experimental Replications. In: ICECCS 2008: Proceedings of the 13th IEEE International Conference on Engineering of Complex Computer Systems, pp. 203–212. IEEE Computer Society (2008)
85. Lung, J., Aranda, J., Easterbrook, S., Wilson, G.: On the Difficulty of Replicating Human Subjects Studies in Software Engineering. In: ICSE 2008: Proceedings of the 30th International Conference on Software Engineering, pp. 191–200. ACM (2008)
86. Mandić, V., Markkula, J., Oivo, M.: Towards Multi-Method Research Approach in Empirical Software Engineering. In: Bomarius, F., Oivo, M., Jaring, P., Abrahamsson, P. (eds.) PROFES 2009. LNBIP, vol. 32, pp. 96–110. Springer, Heidelberg (2009)

A Descriptions of the Replications Typologies

A.1 Bahr et al. [62]

Types A..P. This classification categorizes replications according to four dichotomic properties (equal or different) of a replication. These properties are: time, place, subjects and methods. Based on combinations of these properties, Bahr et al. define 16 replication types.

A.2 Easley et al. [61]

Type 0 (Precise Duplication). This replication is defined as a precise duplication of a prior study. Therefore, Type 0 (precise duplication) studies are those studies

in which every nuance of the experimental setting is precisely reproduced; as such, the cause-effect relationship is finite. The ability to conduct a Type 0 replication is limited to experimenters in only some of the natural sciences. As others have stated, it is an impossibility to conduct a Type 0 replication in a social science context because uncontrolled extraneous factors have the potential to interact with the various components in an experimental setting. For example, human subjects cannot be precisely duplicated. A social scientist is limited only to matching subjects as closely as possible.

Type I (Duplication). A type I replication is a faithful duplication of a prior study and, as such, is considered the "purest" form of replication research in the social sciences. It should be mentioned at this point that a Type I replication is the one most closely associated with the term "replication" in the minds of most researches. More over, this is also the type of replication research most criticized for not being creative. This is somewhat ironic, given the apparent receptivity of reviewers to cross-cultural research that, in many cases, is usually the study of the generalizability of findings from a single country or culture to others and, thus, is simply a Type I replication.

Type II (Similar). A type II replication is a close replication is a close replication of a prior study, and a Type III replication is a deliberate modification of a prior study. Type II replications are the most common form of replication research in marketing settings and are useful in testing phenomena in multiple contexts. If effects are shown in a variety of testing contexts, the case for the findings is strengthened. This has been called the process of triangulation.

Type III (Modification). This replication is a deliberate modification of a prior study. In a Type III replication, the threat of extraneous factors inherent to the nature of human subjects, unless explicitly accounted for in theory testing, is not a factor of concern with regard to replicability.

A.3 Evanschitzky and Armstrong [64]

Real Replications. This replication is a duplication of a previously published empirical study that is concerned with assessing whether similar findings can be obtained upon repeating the study. This definition covers what are variously referred to as "exact", "straight" or "direct" replications. Such works duplicate as closely as possible the research design used in the original study by employing the same variable definitions, settings, measurement instruments, analytical techniques, and so on.

Model Comparisons. This replication is an application of a previously published statistical analysis that is concerned with assessing whether a superior goodness-of-fit can be obtained, comparing the original statistical model with at least one other statistical model.

Data Re-analyses. This replication can be defined as an application of previously published data that is concerned with assessing whether similar findings can be obtained using a different methodology with the same data or a sub-sample of the data.

A.4 Finifter [71]

Virtual Replication. The intention is to repeat an original study not identically but "for all practical purposes" to see whether its results hold up against chance and artifact. Virtual replications are also frequently conducted to find out how dependent a result is on the specific research conditions and procedures used in an original study. To answer this question, one or more of the initial methodological conditions is intentionally altered. For example, a survey or experiment might be repeated except for a change in measuring devices, in the samples used, or in research personnel. If the initial result reappears despite changes, faith in the original finding mounts.

Systematic Replication. The emphasis in systematic replication is not on reproducing either the methods or the substance of a previous study. Instead, the objective is to produce new findings (using whatever methods) which are expected by logical implication to follow from the original study being replicated. When such an implication is actually confirmed by systematic replication, confidence is enhanced not only in the initial finding that prompted the replication but also both in the derived finding and in whatever theoretical superstructure was used to generate the confirmed inference.

Pseudoreplication. It can be defined according to three main operational variations: the repetition of a study on certain subsets of an available total body of real data; the repetition of areal data study on artificial data sets which are intended to simulate the real data; and the repeated generation of completely artificial data sets according to an experimental prescription.

A.5 Hendrick [70]

Strict Replication. An exact, or strict, replication is one in which independent variables (treatments) are duplicated as exactly as possible. That is, the physical procedures are reinstituted as closely as possible. It is implicitly assumed that contextual variables are either the same as in the original experiment, or are irrelevant.

Partial Replication. A partial replication is some change (deletion or addition) in part of the procedural variables, while other parts are duplicated as in the original experiment. Usually some aspect of the procedures is considered "unessential", or some small addition is made to expedite data collection.

Conceptual Replication. A conceptual replication is an attempt to convey the same crucial structure of information in the independent variables to subjects, but by a radical transformation of the procedural variables. In addition, specific task variables are often necessarily changed as well.

A.6 Hunter [31]

Statistical Replication. For statistical replications as perfectly replicated studies:

1. All studies measure the independent variable in exactly the same way.
2. All studies measure the dependent variable in exactly the same way.
3. All studies use exactly the same procedure.
4. All studies draw samples from the same population.

Scientific Replication. For scientific replications for simple causal studies:

1. All studies measure the same independent variable X.
2. All studies measure the same dependent variable Y.
3. All studies use essentially the same procedure.
4. All studies should sample from populations that are equivalent in terms of the study question and hence the study outcome. The difference is that statistical replications assume that the word "same" means identical, while scientists interpret the word "same" to mean equivalent.

Conceptual Replication. This replication verifies one of the hypotheses that were not tested in the original study. The researcher of the original study defines control groups to test the most obvious alternative hypotheses against administrative details that are thought to be irrelevant. Any treatment, intervention or manipulation is a set of administrative procedures, which are mostly intrinsic to the active ingredient of the treatment. These replications examine whether the administrative procedures influence the treatments as reflected in the dependent variable.

A.7 Van IJzendoorn [63]

Complete Secondary Analysis. It is a kind of replication in which all parameters except the researcher and the method of data analysis are kept constant. Secondary analysis also is one of the most inexpensive and efficient types of replication, because it is based on existing data sets. One of the main barriers to secondary replication is, however, the accessibility of the original data sets. The complete secondary analysis may include recoding of the original raw data. In this replication, there are two phases of processing the raw data involved: the coding and analyzing of the data.

Restricted Secondary Analysis. In this type, the coding system is not changed but only the methods of analyzing the data, to see whether the original results survive statistical criticism or the application of refined methods of statistical analysis.

Exact Replication. A replication will be called "exact" if it is essentially similar to the original study. This replication is applied to (dis)confirm the doubts, and to check the assumptions of the varied replications. Many scientists feel that exact replications may be carried out, but usually are irrelevant for scientific progress.

Varied Replication. Replications should be carried out in which new data under different conditions are being collected. From the start, the original study will be "trusted" so much that rather significant variations in the design will be

applied. Larger variations may lead to more interesting discoveries in addition to the original study, but they will be followed by smaller variations if more global replications fail to produce new "facts". If even modest variations fail to reproduce the results, a more or less exact replication is needed.

A.8 Kantowitz et al. [65]

Direct Replication. This is the attempt to repeat the experiment as closely as is practical, with as few changes as possible in the original method.

Systematic Replication. The experimenter attempts to vary factors believed to be irrelevant to the experimental outcome. If the phenomenon is not illusory, it will survive these changes. If the effect disappears, then the researcher has discovered important boundary conditions on the phenomenon being studied.

Conceptual Replication. One attempts to replicate a phenomenon, but in a way radically different from the original experiment.

A.9 Kelly et al. [72]

Literal Replication. The earlier findings may be reexamined using the same manipulations (independent variables, experimental procedures, etc.) and measures (dependent variables, methods of data analysis, etc.).

Operational Replication. If the experimenter wishes to vary criterion measures, the experiment would be termed an operational replication. In this instance, the dependent variable would represent a different operationalization of the construct; the essential conceptual meaning would remain unchanged.

Instrumental Replication. This replication is carried out when the dependent measures are replicated and the experimental manipulations are varied. Variations in the implementation of experimental procedures which do not go beyond the originally established relationship would be included in this category.

Constructive Replication. A constructive replication attempt may be identified when both manipulations and measures are varied. This replication involves the attempt to achieve equivalent results using an entirely original methods recipe.

A.10 La Sorte [33]

Retest Replication. In its general form retest replication is a repeat of an original study with few if any significant changes in the research design. The retest has two major purposes: 1) it acts as a reliability check of the original study, and 2) inconsistencies and errors in procedure and analysis can be uncovered in the repeat. Although the retest increases one's confidence that the findings are not artifactual, it does not eliminate the possibility of error in process, especially when the same investigator conducts both studies.

Internal Replication. The differences between the retest and internal replication are mainly procedural. Instead of seeking confirmation of an original study,

the internal replication is built into the original study design. So the data, part of which are used for the replication, are gathered simultaneously by the same investigator using a common set of research operations. One finds variations in the procedures for selecting the samples. Two of these procedures are: 1) drawing two or more independent samples, and 2) taking a single sample and later dividing it into subsamples for purposes of analysis and comparison. The internal replication provides an additional data supply which acts to cross-check the reliability of the observed relationships. Thus it is methodologically superior to the single study where the hypothesis is tested only once by one body of data.

Independent Replication. Independent replication is the basic procedure for verifying an empirical generalization. It does this by introducing significant modifications into the original research design in order to answer questions about the empirical generalization that go beyond those of reliability and confirmation. The essential modifications include independent samples drawn from related or different universes by different investigators. These replications differ in design and purpose. They can, however, be broadly categorized into three problem areas. First, is the empirical generalization valid? Second, does further investigation extend it to other social situations or subgroups outside the scope of the original study? Or, third, is the empirical generalization limited by the conditions of particular social situations or specific subgroups?

Theoretical Replication. It involves the inductive process of examining the feasibility of fitting empirical findings into a general theoretical framework. These replications seek to verify theoretical generalizations. In these replications, empirical variables, which have concrete anchoring points are abstracted and conceptualized to a higher theoretical plane, it is necessary to sample a variety of groups using different indicators of the same concepts.

A.11 Leone and Schultz [73]

Experimental Replication. The same experiment is conduced more than once, although there can be (especially with social systems) no perfect replications. It involves the same method and the same situation.

Nonexperimental Replication. The same method is applied to different situations.

Corroboration. It involves different method and same situation, or different method and different situation.

A.12 Lindsay and Ehrenberg [68]

Close Replication. This replication attempts to keep almost all the know conditions of the study much the same or at least very similar (for example, the population or populations in question, the sampling procedure, the measuring techniques, the background conditions, and the methods of analysis). A close replication is particularly suitable early in a program of research to establish quickly and relatively easily and cheaply whether a new result can be repeated at all.

Differentiated Replication. It involves deliberate, or at least known, variations in fairly major aspects of the conditions of the study. The aim is to extend the range of conditions under which the result still holds. Exploring a result with deliberate variations in the conditions of observation is the essence of generalization. According to the authors, there are three reasons for running a differentiated replication:

1. Use different methods (different measuring instruments, analysis procedures, experimental setups, and/or investigators) to reach the same result (triangulation),
2. Extended the scope of the results,
3. Define the conditions under which the generalization no longer holds.

A.13 Lykken [69]

Literal Replication. This involves exact duplication of the first investigator's sampling procedure, experimental conditions, measuring techniques, and methods of analysis.

Operational Replication. One strives to duplicate exactly just the sampling and experimental procedures given in the first author's report. The purpose of operational replication is to test whether the investigator's "experimental recipe" the conditions and procedures he considered salient enough to be listed in the "Methods" section of his report will in other hands produce the results that he obtained.

Constructive Replication. One deliberately avoids imitation of the first author's methods. To obtain an ideal constructive replication, one would provide a competent investigator with nothing more than a clear statement of the empirical "fact" which the first author would claim to have established.

A.14 Mittelstaedt and Zorn [67]

Type I. The replicating researcher uses the same data sources, models, proxy variables and statistical methods as the original researcher.

Type II. The replicating researcher uses the same data sources, but employs different models, proxy variables and/or statistical methods.

Type III. The replicating researcher uses the same models, proxy variables and statistical methods, but applies them to different data than those used by the original researcher.

Type IV. In this replication, different models, proxy variables and statistical methods are applied to different data.

A.15 Monroe [74]

Simultaneous Replication. Does the same researcher in the same study investigate consumer reactions to more than one product, or to more than one advertisement?

Sequential Replication. Does the researcher or another researcher repeat the study using the same or different stimuli at another point in time?

Nonindependent Replication. The replication is conducted by the same researcher.

Independent Replication. The replication is conducted by different researcher.

Assumed Replication. For example, a researcher using both males and females simultaneously in a study and finding no gender covariate effect assumes replication across gender.

Demonstrated Replication. What is preferable is separate gender conditions wherein the effect has or has not been obtained separately for males an females, that is, demonstrated.

Strict Replication. The replication is a faithful reproduction of the original study.

Partial Replication. The replication is a faithful reproduction of some aspects of the original study.

Conceptual Replication. The replication uses a similar conceptual structure but incorporates changes in procedures and independent variables.

A.16 Radder [75]

Reproducibility of the material realization of an experiment. In this type of reproduction, the replicator correctly performs all the experimental actions following instructions given by the experimenter who ran the previous experiment. This reproduction is based on a division of labour, where other previously instructed people can run the replication without being acquainted with the theory underlying the experiment. As in this reproduction it is possible to follow the same procedure to verify the outcome without detailed knowledge of the theory, there may be differences in the theoretical interpretations of the experiment.

Reproducibility of an experiment under a fixed theoretical interpretation. This reproduction implies that the conditions of the previous experiment can be intentionally altered in the replications, provided that the variations are irrelevant to the theoretical interpretation of the experiment.

Reproducibility of the results of an experiment. This type of reproduction refers to when it is possible to achieve the same result as a previous experiment using different methods. This category excludes a reproduction of the same material operationalization.

A.17 Schmidt [35]

Direct Replication. This involves repeating the procedure of a previous experiment. In this replication, the context variables, the dependent variable or subject selection are open to modification.

Conceptual Replication. This is the use of different methods to retest the hypothesis or result of a previous experiment.

A.18 Tsang and Kwan [66]

Checking of Analysis. In this type of replication, the researcher employs exactly the same procedures used in a past study to analyze the latter's data set. Its purpose is to check whether investigators of the original study have committed any errors in the process of analyzing the data.

Reanalysis of Data. Unlike the checking of analysis, in this type of replication, the researcher uses different procedures to reanalyze the data of a previous study. The aim is to assess whether and how the results are affected by problems of definition, as well as by the particular techniques of analysis. Quite often the replication involves using more powerful statistical thecniques that were not available when the original study was conducted.

Exact Replication. This is the case where a previous study is repeated on the same population by using basically the same procedures. The objective is to keep the contingent conditions as similar as possible to those of the previous study. The researcher usually uses a different sample of the subjects. The main purpose is to assess whether the findings of a past study are reproducible.

Conceptual Extension. A conceptual extension involves employing procedures different from those of the original study and drawing a sample from the same population. The differences may lie in the way of measuring constructs, structuring the relationships among constructs, analyzing data, and so forth. In spite of these differences, the replication is based on the same theory as the original study. The findings may lead to a revision of the theory.

Empirical Generalization. In this replication, a previous study is repeated on different populations. The researcher runs an empirical generalization to test the extent to which the study results can be generalized to other populations. It follows the original experimental procedures as closely as possible.

Generalization and Extension. The researcher employs different research procedures and draws a sample from a different population of subjects. The more imprecise the replication, the greater the benefit to the external validity of the original finding, if its results support the finding. However, if the result fail to support the original finding, it is difficult to tell whether that lack of support stems from the instability of the finding or from the imprecision of the replication.