VOL. 3, NO. 9 SEP, 2012 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

Design Science Evaluation – Example of Experimental Design

[1] Lukasz Ostrowski, [2] Markus Helfert
[1, 2] School of Computing, Dublin City University
[1] lostrowski@computing.dcu.ie, [2] markus.helfert@computing.dcu.ie

ABSTRACT

Evaluation plays a major part in Design Science Research; however, researchers provide very little examples of how one could actually conduct this part at the operational research level. To address this need, we present an example of utility evaluation of design science artifact using an experimental design. We investigate whether an artifact as a treatment of for process development in design science research methodology improved the representational information quality of design science artifacts. The control condition is that each practitioner was presented with two artifacts in a basic two-condition repeated measures design. The improvement was measured after examining each artifact using paper-questionnaire. The paper presents DS researchers to numerous benefits that a simple experiment can provide.

Keywords: *Design Science Evaluation, Experimental Design, Wilcox on test*

1. INTRODUCTION

Evaluation delivers evidence that a solution developed in design science research (DSR) achieves the purpose for which it was designed. Without evaluation, outcomes are unconfirmed declarations that the artifacts meet their purpose (i.e. be useful for solving a problem or making some improvement). Design science artifacts "are assessed against criteria of value or utility – does it work?" [1]. The essential aim is to rigorously demonstrate the utility of the artifact being evaluated. Rigor in DSR should be approached from two directions. One is to establish if the artifact causes an observed improvement, its efficacy. The second direction is to establish if the artifact works in a real situation, its effectiveness. 2].

Evaluating the utility of design science artifacts can also be perceived through the information system design theories [3, 4] or design principles [5] , which formalizes the knowledge of the utility of design science artifacts. We confirm or disprove the design theory by evaluating design science artifacts [6] . A new solution should provide greater relative utility than existing artifacts that can be used to achieve the same purpose [7].

Utility of artifacts is a complex deliverable. It may depend on many different attributes of the artifact or desired outcomes of the use of the artifact. Researchers state that "artifacts can be evaluated in terms of functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization, and other relevant quality attributes" [8] . Hence, each evaluation is quite specific to the artifact, its purpose, and the purpose of the evaluation [6]. We can distinguish two types of artifacts, product and process [3] . The former represents tools, diagrams or software that people use to solve a problem. The latter is in a form of a method or procedure that guides someone what to do to solve a problem, thus a person must interact to provide utility of the artifact. All of these properties of the artifact in some way contribute to the utility of the design science artifacts and act as criteria that are candidates for evaluation in determining the overall utility.

Researchers identified a number of methods that can be used for evaluation of design science artifacts. Hevner [8] proposed five classes of evaluation methods: (1) Observational methods include case study and field study. (2) Analytical methods include static analysis, architecture analysis, optimization, and dynamic analysis. (3) Experimental methods include controlled experiment and simulation. (4) Testing methods include functional testing and structural testing. (5) Descriptive methods include informed argument and scenarios. Peffers et al [9] divide evaluation into two activities, demonstration and evaluation. The former demonstrates that the artifact feasibly works to achieve its purpose in at least one context. The latter considers how well the artifact supports a solution to a problem.

Venable [7] divides evaluation into artificial and naturalistic. Artificial evaluation includes laboratory experiments, field experiments, simulations, criteria-based analysis, theoretical arguments, and mathematical proofs. It evaluates a solution in a contrived and non-realistic way. Naturalistic evaluation explores the performance of a solution in its real environment. By performing evaluation in a real environment (real people, real systems, and real settings

[10]), naturalistic evaluation embraces all of the complexities of human practice in real organizations. This approach is always empirical, and includes methods such as case studies, field studies, surveys, and action research [6] . The dominance of the naturalistic paradigm brings to naturalistic DSR evaluation the benefits of stronger internal validity [11] . However, these authors provide no guidance for choosing between methods, and there is little guidance in the DSR literature about the choice of strategies and methods for evaluation in DSR. The most cited guide selection of evaluation strategies for a DSR project is 2-by-2 framework [12] , extended by Venable [6] . They identify that evaluation design needs to decide what, how and when will be evaluated. However, this framework is not a framework for evaluating DSR projects as a whole, but it aids DSR researchers in the design of the evaluation component of their DSR [12] .

However, beyond providing the framework and an idea of what needs to be designed in the DSR component of research, researchers provide very little guidance in how a researcher should or could actually conduct the evaluation in DSR. This state of affairs in DSR constitutes what we can call "a gap of practical evaluation examples".

The purpose of this paper is to respond to this evaluation gap by giving such an example with clear guidance for how one could design and conduct evaluation with DSR. This example presents experimental class evaluation [8], based on field experiment structure [13], of the reference model [14] which is the process type artifact [3] in the most possibly naturalistic settings [7]. The reference model artifact provides guidance on process oriented design science research. Its utility is evaluated in terms of how well information provided by the artifacts developed with the reference model fits for use - the quality of information representation to information consumers [15] . We understand information quality as the fitness for use of the information provided to stakeholders. We wanted to know whether the reference model as a usage (treatment) of for process development in design science research methodology improved the representational information quality of the research artifacts (i.e. an outcome of design science research). In our case the research artifact was always a process. Each developed artifact aimed to fulfill exactly the same research objectives; to produce an IT service process for a public organization. We had predicted that having used the reference model to develop a design science process oriented artifact would lead to greater representational information quality than following this methodology without the model.

The experiment is organized as follows. The next section discusses rationale behind an experiment, selected variables, and validity of measuring. Based on that discussion, the subsequent sections present the experimental design and execution. Next, we justify the test for the obtained data and interpretation of results.

2. THE RATIONALE BEHIND THE EXPERIMENT

Popper [16] suggested that truth of a scientific statement or theory could be tested only by comparing two hypotheses that differ in a single respect. Mill [17] indicated that by comparing two situations that differ only in the presence of the causal variable, causality could be isolated. Both Mill and Popper pointed out to the fundamental importance of controlling all factors other than the one that is of interest to the scientist. Experiments are conducted following this rationale. Mill [17] proposed that causal factors could be isolated only by comparing two conditions: one in which supposed cause is present, and one in which supposed cause is absent. The variable that we typically manipulate is the one we have proposed as a cause and in the simplest situation we manipulate it by changing whether the cause is present or absent. This manipulation is called levels of the variable. For example, two levels mean that the supposed cause can be present or absent. The variable that is manipulated is named the independent variable (it depends on the experimenter), the one that is not manipulated by the experimenter, is called the dependent variable (the outcome of the experiment). In our experiment, there was one independent variable, the way in which an artifact was developed. It had two levels: an artifact developed with the reference model based on design science methodology or an artifact developed only with the design science methodology. The outcome of the experiment was the total scores of participants rating of the representational information quality of artifacts (i.e. how good the artifacts represent information to their stakeholders).

2.1 Experiment Validity

One important issue when deciding how to measure the dependent variable is validity. Validity refers to the fact that we measure what we think we are measuring [13] . When researchers go for a self-report measure (e.g. questionnaire) of the experimental outcome they should consider content validity. It refers to the items in the questionnaire, which must relate to

the construct being measured. This is achieved if items are representative, not deliberately similar to other items, and questions cover the full range of the construct. In our experiment, to achieve the content validity, we built the questionnaire on the representational information quality dimensions [15]: concise representation, consistency, ease of understanding, and interpretability. Questions for each dimension were constructed based on their identified attributes [18]. In terms of measurement, we used an 11-point Likert type scale. The number 10 was labeled as "Extremely good", while 0 as "Not at all", and 5 as "Average". Most questions in the questionnaire were formulated as "how <Attributes of the Item> is the artifact?" For example, "How easy is the artifact to understand?" The data then consist of each participant providing a score (rating) of how they found the artifact in terms of the quality of represented information.

Quality cannot be taken for granted or assumed. Instead, quality is a subjective term for which each person has his own definition [19]. We can be reasonable confident that a score of 8 refers to better representation than a score of 7 and that a score of 9 almost certainly represents information better than a score of 8. However, we cannot conclude by how much guidance having the score 9 is better compared to other guidance having the score 8 or a 7. A score of 8 might represent an enormous difference over a score of 7, whereas a score of 9 might represent only a minor gain over a score of 8 – or vice versa. In addition, it probably isn't realistic to assume that if one researcher rates one attribute (e.g. ease to find key points) of the guidance as 3 then the attribute of the guidance is, in reality, half as good as the same attribute of a different guidance which was rated as 6. We might question whether two guidance which both were rated as 7 are likely to be equally good. Hence, we treated this data (ratings) as ordinal data (i.e. an arbitrary numeric scale where the exact numeric quantity of a particular value has no significance beyond its ability to establish a ranking over a set of data points [20]). This assumption is important for further data analysis in the below results section.

There are different ways in which validity of results can be assured. If obtained data are due only to the manipulation, then there is no lack in internal validity. Selecting an appropriate experimental design gives reasonable confidence for internal validity. We used repeated measure design (see the experimental design section below). If findings are not only valid for the specific situation within which were obtained, then there is no lack in external validity. To achieve this, we would need to run more experiments in different environments to be on reasonable safe ground of our prediction. Hence, at this stage our results will show lack of external validity. The findings might be true only for this particular scenario.

2.2 Experiment Reliability

Validity is a necessary but not sufficient condition of a questionnaire. A second consideration is reliability. Reliability is the ability of the measure to produce the same results under the same conditions [13]. To be reliable the questionnaire must first be valid. One of the ways to assess scale reliability of questionnaire is to test the same group of participants twice: if the questionnaire is reliable we would expect each participant's scores to be the same at both points in time. So, scores on the questionnaire should correlate quite well. However, in the real experimental environment, if we did test the same participants twice then we would expect some practice effects and confounding effects, people might remember their responses from last time, and testing twice is also time-consuming. There are statistical methods to overcome this problem. Cranach [21] suggested splitting the data in half in every conceivable way and computing the correlation coefficient for each split. The average of these values is known as Cranach's alpha, which is the most common measure of scale reliability. An acceptable value for Cranach's alpha is a value greater than 0.7 [22] ; values substantially lower indicate an unreliable scale. Test of reliability applies upon data of the experiment are collected. In our experiment, answers from 100 respondents were found. We used SPSS software to calculate Cranach's alpha to determine to what degree our questionnaire were successful in constructing questions that measure a participant's opinion. Results in Table 4 indicate a reliably acceptable scale of our questionnaire.

Table 1: Reliability Analysis

Cranach's Alpha	Cranach's Alpha Based on Standardized Items	N of Items
.758	.735	11

In the following section we describe the experimental design, its execution and data collection.

3. EXPERIMENTAL DESIGN

We looked for a difference between the artifacts in terms of improvement of the perceived

representational information quality. We asked practitioners, hereinafter called participants, to examine two artifacts, first developed with the reference model and second without it following design science research. Then, we asked participants to respond to the questionnaire.

In this experiment we used a basic two condition repeated measures design Figure 1 [13]. Under this design each practitioner was randomly assigned to the order in which the artifacts were examined. The improvement was measured after examining each artifact. To maximize our chances of finding a difference we used a sample of 50 participants. We got each participant to take part in both conditions (they examined both artifacts). The order in which artifacts were assessed was counterbalanced (see Figure 1), and there was a delay of 20 minutes between examining the artifacts.

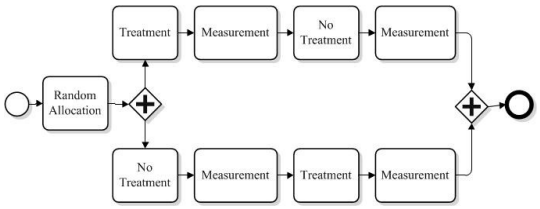

Figure 1: A basic- two condition repeated measures design

Treatment represents the artifact developed with the reference model. No treatment refers to the artifact used without implication of the reference model in design science research. Measurement is the phase when the questionnaire was provided. We had predicted that having used the reference model to develop a design science process oriented artifact would lead to greater information quality than following this methodology without the model. Developing an artifact without the reference model refers to the fact that researchers were free to choose methods while carrying out the research. Development of an artifact with the reference model imposed methods on researchers.

Out case was wholly repeated measures design (within subjects design). We used the same participants in every condition – so they produce one result for every condition of the experiment. We chose this design for two reasons. One was the fact that it was more economical to run in terms of time and effort. We used the same participants twice. The second was sensitivity. We were keen to find the differences in our results which had been produced by our experimental design. These differences would become clear only after reducing all the random 'noise' produced in our data by the fact that participants differed from each other. In this repeated-measures design, there are only a few sources of random variation to vague the effects of our manipulation of the independent variable. Usually, researchers need to deal with differences between in the experimental conditions, random differences between individuals within a group, and random differences between individuals in one group and individuals in another group. In repeated-measure design the last aspect is eliminated. Hence, all the efforts are put on the individual variation in participant's response to the experimental manipulation.

However, there are some pitfalls of this design. Although our manipulation had no effect on participants' behavior, they could still give slightly more or less different responses in our different experimental conditions. If we observed merely a random fluctuation in their performance, this should not be a problem because this behavior should cancel out across conditions. However, systematic variations in performance may cause some issues. Participants could get bored with time and better practice at examining the artifacts, for example. These systematic effects may interact with the manipulations of the independent variable and reduce interpretability of the results. This is called a 'carry over' effect from one condition to another (Field & Hole 2003). For example, if each participant took part in each condition, in the same order, and we found satisfactory differences in our manipulation, we would not be able to tell if the effect was due to manipulation or due to practice in most cases. To avoid the 'carry over effect', we can counterbalance the order, half the participants get the conditions in order A then B and the other half get the order B then A.

Participants of this experiment were employees (practitioners) of a public organization. The organization provided IT services for navy's various departments. The practitioners in the numbers of fifty were between 29-58 years of age (M 43, SD 3.4). The gender was split in 37 males, and 13 females. All were free from any obvious physical or sensory impairment. Their work experience in the organization was between 0.5 to 12 years (M 5, SD 1.3). They role mainly were engineers from fields of electronics, design, architecture, and computing. Participants took part in the experiment willingly, and therefore we believed their responses to the questionnaire were genuine.

3.1 Procedure

For this experiment we needed two artifacts, one developed with the reference model and the other without it. Each artifact was developed accordingly to the same research objectives, and should represent a process of an IT service request that this public organization uses. The artifacts were developed by teams of 8 students within 2 months. The actual development of these artifacts is out of scope of this paper. However, the development was also conducted in rigour of an experiment; hence we assumed that the only variance was due to the presence or absence of the reference model.

The examination of those artifacts was conducted between 10 a.m. and 1 p.m. in the conference room of the public organization. Within first 20 minutes we allocated participants to each condition. We used a random number generator of a computer. As each participant arrived, we followed a rule such as: if the next random number is even, the participant goes to the conditions with order A then B (Figure 1); if it is odd, the participant goes to the conditions with order B then A. This way, we avoided running participants in ways which are likely to produce systematic differences between groups of the orders of conditions. For example, by assigning all the participants who turned up on time to one condition, and the other all participants who came late to another condition, we might have a group of people who pay more attention to details (like being on time) and those who don't in the another group.

Once everyone was assigned to a condition, we provided the artifacts accordingly. 5 minutes were given to explain under which angle the participants should examine the artifact were followed by another 5 minutes to explain our questionnaire in a paper form. Afterwards, the thorough examination was allowed for 30 minutes. Then, we provided the questionnaire, which was not available during the examination, and allowed 10 minutes to provide answers. Participants still had access to the relevant artifact. After this stage, there was 10 minutes break during which we collected the questionnaire and swapped artifacts accordingly. The following examination phase looked similar: 30 minutes examination, and 10 minutes for answering questions. There was explanation, since we had the same participants and the time-delay was minor. Upon collecting all questionnaires the experiment was over, and we went to digitize the data for analysis.

In the following section we discuss the applicable test for the obtained data, and present main calculations with interpretation of results.

4. RESULTS

One distinction needs to be made before we could look into available tests for our data. We need to check whether our data meet requirements of parametric test or we should look for a test among non-parametric tests. The parametric tests assume that our data is normally distributed. That would be a roughly bell-shaped frequency distribution of scores in each group (a group with or without the reference model, around the mean of a group. The data must also show homogeneity of variance. We would be looking for spreads of scores in each group that weren't wildly dissimilar from each other. Finally the data measurements must be on an interval or ratio scale [13]. As we discussed (in experimental validity section) our data measurements was at the ordinal level; therefore we could not use a parametric- test. We had to find a suitable non-parametric test.

Using the chart in Figure 2 together with our discussion, we can now decide which test we should run on our data. Beginning from the top Figure 2, we knew that our data consisted of scores; so that ruled out Chi-Squared as an option. Chi-Squared applies when you have nominal (categorical) data with each person contributing only once to each category. We had an experiment design, so that ruled our using correlations. Correlation is when you look for relationships between variables without manipulating them, as you do in an experiment. We had one independent variable, and we had a repeated-measures design. There were two conditions: an artifact developed with the reference model and without it. We had now narrowed our choice of test down to either a repeated-measures t-test, or its non-parametric equivalent, Wilcox on test. Since our data did not satisfy the requirements for a parametric test, we used the Wilcox on test on our data.

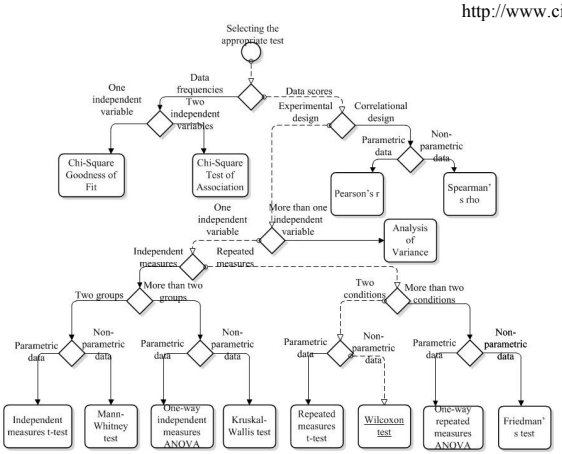

Figure 2: Selection of the appropriate test for the data

4.1 The Wilcox on Signed-Rank Test

The Wilcox on signed-rank test is used for testing differences between groups when there are two conditions and the same participants have been used in both conditions. In our experiment each participant examined both artifacts (developed with or without the reference model). We measured the total scores of how well information provided by artifacts fits for use after developing them with or without use of the reference model. It was hypothesized that using the reference model for process oriented artifacts in design science research would improve their quality of information provided.

We used IBM SPSS software to run the Wilcox on test. Upon collecting the data, we noticed the data were non-normal in the artifact developed without the reference model condition. In fact the Kolmogorov-Smirnov test (Table 1) was significant for the artifact without the reference model condition ($D(50) = .111$, $p < .05$), indicated that a non-parametric test was appropriate.

Table 1: Test of Normality

Condition		Kolmogorov-Smirnov		
		Statistic	df	Sig.
Total Score of Information Quality	with the reference model	.111	50	.170
	without the reference model	.131	50	.032

Wilcox on test belongs to the group of non-parametric tests. This group makes less strict assumptions about the non-normal distribution of data being analyzed. The way they get around the problem of the distribution of the data is by not using the raw scores. Instead, the data are ranked. For example, the numerical data 34, 53, 22, 79 were observed, the ranks of these data items would be 2, 3, 1 and 4 respectively. The analysis is then carried out on the ranked data. However, by ranking the data we lose some information about the magnitude of difference between scores and because of this non-parametric tests are less powerful than parametric equivalents.

The Wilcox on test first looks for the difference between each pair of scores in our two conditions, and then ranks these differences for further examination. It compares the ranks of each participant regarding the artifact developed with and without the reference model. The differences between ranks can be positive (the rank in condition two is bigger than the rank in condition one), negative (the rank in condition two is smaller than the rank in condition one) or tied (the ranks in the two conditions are identical). Table 2 shows a summary of these ranked data. It tells us the number of negative ranks (i.e. participants scoring better the artifact developed without using the reference model rather than with it) and the number of positive ranks (i.e. better scores of the artifact developed with the reference model than without it). The footnotes (a, b, c) under the table help determine to what condition the positive and negative ranks refer. The Table 2 shows that 6 of the 50 participants found better information quality of the artifact developed without the reference model, whereas 43 of the 50 participants favored the artifact developed with the reference model. There was 1 tied rank (i.e. a participant who equally assessed both artifacts. The table also shows the average number and the sum of negative and positive ranks.

Table 2: Wilcox on Signed Ranks Test

	N	Mean Rank	Sum of Ranks
Negative Ranks	6[a]	3.50	21.00
Positive Ranks	43[b]	28.00	1204.00
Ties	1[c]		
Total	50		

a. developed with the reference model < developed without the reference model

b. developed with the reference model > developed without the reference model

c. developed with the reference model = developed without the reference model

Wilcox on test can be converted to a z-score, which indicates how many standard deviations an observation or datum is above or below the mean. In other words it allows calculating the exact significance values based on the normal distribution. Table 3 tells us that the test statistic is based on the negative ranks, that the z-score is -5.886 (negative means that it is below the group mean) and that this value is significant at p= .0003 (i.e. very high significance indicates that those scores very unlikely happened by chance).

Table 3: Wilcox on test converted to z-score

	developed with the reference model - developed without the reference model
Z	-5.886[a]
Asymp. Sig. (1-tailed)	.0003

a. Based on negative ranks.

Most participants fall into the category with positive ranks. We can tell that because the mean rank is higher for the positive ranks. So this means that most people fell into the category of scoring better for the artifact developed with the reference model. There were significantly more people who had positive ranks than had negative ranks. Therefore, we can conclude that significantly information provided by the artifact developed with the reference model is of better representational information quality. This is in the direction to our hypothesis, so we used the 1-tailed significance value (.0003).

4.2 Data Display

A good way to display non-parametric data is by using a box plot diagram. Non-parametric tests are not testing differences between means; they are testing differences between ranks. As we already discussed, we are dealing with ordinal data, and therefore comparing means is not good representation of our data. Hence, box plot shows the median (the middle score), and so better represents what the non-parametric test is looking at. Figure 4 shows our data on such a diagram. The shaded box represents the range between which 50% of the data fall. The horizontal bar within the shaded box is the median. The 'I' shape shows the limits within which most of all of the data fall. The lower bar is the lowest score and the upper bar is the highest score in the data. However, if there is an outlier (a score very different from the rest – there is none in our data), then it will fall outside of the bars and the bars represent all of the data that fall within +/- 3 standard deviations of the mean.

Figure 3 illustrates that after using the reference model to develop an artifact the median number of total score of information quality was higher than without the model being involved in the artifact development. The fact that the median is higher with the reference model confirms the direction of our conclusions (i.e. information provided by the artifact developed with the reference model is of better representational information quality)

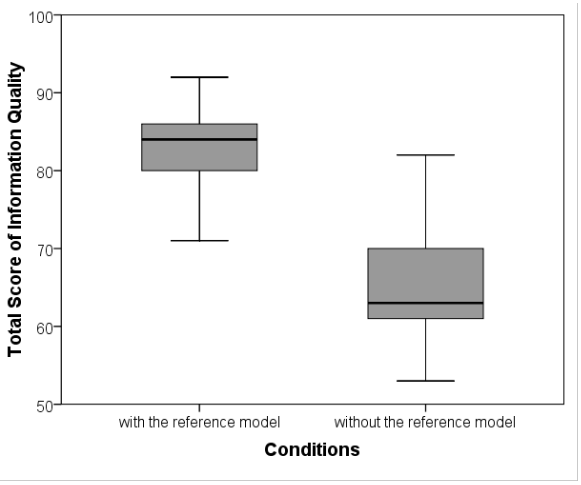

Figure 3: Box plot for the information quality of artifacts developed with or without the reference model

4.3 The Effect Size

The fact that our test statistic was significant (p< 0.5), didn't mean that the effect it measured was meaningful or important. The solution to this was to measure the size of the effect that we were testing. Measuring the size of an effect either by experimental manipulation or observation of the strength of relationship

between variables is knows as an effect size [13]. It is an objective and standardized measure of the importance of observed effect. There are many measures of effect size, but the most common one is Pearson's correlation coefficient [23]. To measure the size of experimental effects we look at the proportion of total variance in the data that can be explained by the experiment, which is equal to r^2 (coefficient of determination). Since this is a proportion, it must have a value between 0 and 1. 0 means that the experiment explains none of the variance at all. 1 means that the experiment can explain all of the variance. It can also have minus values (but not below -1), however, in experimental manipulation the sign of r merely reflects the way in which the experimenter coded their groups [24]. Generally, the bigger the value is the bigger experimental effect. If we take the square root of this proportion, we get the Pearson correlation coefficient, r, which is also constrained to lie between 0 (no effect) and 1 (a perfect effect). It provides an objective measure of the importance of the experimental effect. There is no difference in what experiment has been done, what and how variable has been measured we know that a correlation of 0 means the experiment had no effect and a value of 1 means that the experiment completely explains the variance in the data). There are some widely accepted suggestions about what constitutes a large or small effect [25]:

- r=0.10 (small effect): in this case the effect explains 1% of the total variance.
- r= 0.30 (medium effect): the effect accounts for 9% of the total variance.
- r=0.50 (large effect): the effect accounts for 25% of the variance.

We can use these guidelines to assess the importance of our experimental effects. Since we converted our test statistic into a z-score, we can easier calculate the effect size. The equation to convert a z-score into the effect size estimate, r is as follows [26]:

$$r = \frac{Z}{\sqrt{N}}$$

in which Z is the z-score that we have from Table 3 , and N is the size of the study (i.e. number of observations. The effect size is therefore:

$$r = \frac{-5.886}{\sqrt{100}}$$

r =-0.5886

This represents a large effect of our experiment (it is close to Cohen's benchmark of 0.5), which tells us that the effect of whether the artifact developed with or without the reference model was examined was a substantive effect. The effect accounts for 35% of total variance.

4.4 Interpretation of the Results

The number of total scores of information quality after examining the artifact developed with the reference model (Mdn=84) was significantly higher than after examining the artifact developed without the reference model (Mdn=63, T= 21.00, p< .05, r = - 0.5886). We can say that using the reference model we can explain 35% of the total variability in total scores of the representational information quality of artifacts.

5. CONCLUSION

Evaluation plays a major part in Design Science Research; however, researchers provide very little examples of how one could actually conduct this part. To address this need, we have presented an example of utility evaluation of design science artifact using an experimental design.

We wanted to know whether the reference model [14] as a treatment of for process development in design science research methodology improved the quality of the design science artifacts. The control condition was that each practitioner was presented with two artifacts in a basic two-condition repeated measures design.

First we presented the rationale behind the experimental design. We showed that the aim was to produce results which are valid (they actually show what we intended them to show), reliable (produce the same results under the same conditions), and generalizable (findings should have wider application). We looked into internal and external validity of the experiment. We showed how to approach the measure of scale reliability for questionnaires. Although we covered the first two faces, we would need to conduct the experiment in more and different settings to make our findings generalizable.

Next, selection of the appropriate test for the obtained data was showed. We justified the ordinal type of our data and the Wilcox on test that suited to the experimental conditions. Finally, we described box plot diagram that was appropriate for the ordinal data, and the

concluded with interpretation of the findings and their effect size.

In reporting our experiment we tried to present the method and procedure in good detail in order to allow for potential reproduction, and confirmation of findings. It can be argued that this example is only useful to very limited instances of design science artifacts, and more general view should be taken. Although, we agree on the broader perspective, our intention was to move to the operational level of research and give researchers an example of how one of the evaluation strategies can be really applied.

We aim to describe and test other evaluation methods in design science research in our further work. We will investigate closely selected methods in numerous design research projects, including our own and student projects. Nonetheless, further research is needed to gain more experience in evaluation of design science artifacts, their utility, and discovering new design science evaluation methods.

ACKNOWLEDGMENTS

This work was supported by the Irish Research Council for Science, Engineering and Technology (IRCSET) under the Enterprise Partnership Scheme.

REFERENCES

[1].March, S. T., Smith, G. F.: Design and Natural Science Research on Information Technology. Decision Support System, 251-266 (1995)

[2].Checkland, P., Scholes, J.: Soft Systems Methodology in Practice 1st edn. J.Wiley, Chichester (1990)

[3].Gregor, S., Jones, D.: The Anatomy of a Design Theory. Journal of Assoc. Information Systems 8, 312-335 (2007)

[4].Pries-Heje, J., Baskerville, R.: The Design Theory Nexus. MIS Quarterly, 731-755 (2008)

[5].Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., Lindgren, K.: Action Design Research. MIS Quartely 35(1), 37-56 (2011)

[6].Venable, J., Pries-Heje, J., Baskerville, R.: A Comprehensive Framework for Evaluation in Design Science Research. In : DESRIST 2012, Las Vegas, pp.423-438 (2012)

[7].Venable, J.: A Framework for Design Science Research Activities. In Khosrow-Pour, M., ed. : The 2006 Information Resource Management Association Conference, Washington DC: Idea Group Publishing, pp.184-187 (2006)

[8].Hevner, A. R., March, S. T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly 28, 75-106 (2004)

[9].Peffers, K., Tuunanen, T., Rothenberger, M.: A Design Science Research Methodology. Journal of Management Information Systems 24(3), 45-77 (2007)

[10].Sun, Y., Kantor, P.: Cross-Evaluation: A New Model for Information System Evaluation. Journal of the American Society for Information Science and Technology 57(5), 614-62 (2006)

[11].Gummesson, E.: Qualitative Methods in Management Research 1st edn. Chart-well-Bratt, Lund, Sweden (1988)

[12].Pries-Heje, J., Baskerville, R., Venable, J.: Strategies for Design Science Research Evaluation. In : 16th European Conference on Information Systems, pp.255-266 (2008)

[13].Field, A., Hole, G.: How to Design and Report Experiment 1st edn. Sage Publication, London (2003)

[14].Ostrowski, L., Helfert, M.: Reference Model in Design Science Research to Gather and Model Information. In : 18th Americas Conference on Information Systems, Seattle (2012)

[15].Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4), 5-34 (1996)

[16].Popper, K.: The poverty of historicism 1st edn. Routledge, London (1957)

[17].Mill, J. S.: A system of logic: ratiocinative and inductive 1st edn. Longmans, Green, London (1865)

[18].Ge, M.: Information Quality Assessment and Effects on Inventory Decision-Making., Dublin (2009)

[19].Fishman, N.: Viral Data in SOA: An Enterprise Pandemic 1st edn. IBM Press (2009)

[20].Sheskin, D.: Handbook of Parametric and Nonparametric Statistical Procedures 4th edn. Chapman & Hall/CRC, Boca Raton (2007)

[21].Cronbach, L. J.: Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16(3), 297-334 (1951)

[22].Kline, P.: The Handbook of Psychological Testing 1st edn. Routledge, London (1999)

[23].Ellis, P. D.: The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. 1st edn. United Kingdom: Cambridge University Press, Cambridge (2010)

[24].Field, A. P.: Discovering Statistics using SPSS for Windows: advanced techniques for the beginner 1st edn. Sage, London (2000)

[25].Cohen, J.: Statistical Power Analysis for the Behavioural Sciences 1st edn. New York: Academic Press, New York (1988)

[26].Rosenthal, R., Rosnow, R.: Essentials of Behavioral Research, Methods and Data Analysis 1st edn. McGraw-Hill, San Francisco (1991)