Web Usability Measurement: Comparing Logic Scoring Preference to Subjective Assessment

Michael Chun Long Yip and Emilia Mendes

Computer Science Department, The University of Auckland, Auckland, New Zealand
myip005@ec.auckland.ac.nz
emilia@cs.auckland.ac.nz

Abstract. This paper investigates one of the existing methods for measuring usability – Logic Scoring Preference (LSP), and discusses the results of two formal experiments carried out to assess the extent to which LSP embodies the subjective perception of users in regards to Web usability. The two experiments used Computer Science students as experimental subjects. Our results suggest that scores obtained via LSP are significantly different from scores obtained via subjective opinion. In addition, we obtained contradictory results when investigating the consistency of LSP scores across subjects.

1 Introduction

There are many reasons for why usability should be considered in a software development process [4]:

- Ensuring that the product best suits its target users will make it the product of choice among competitors.
- Having a superior product can justify a slightly higher price, since people would not mind paying more for a product that they trust.
- More money can be made through the ability to sell a product that is easier to use
- Even if the end users are not customers, but employees, a more usable product increases productivity among workers.
- A more intuitive product would also mean that less time is spent learning how to perform a new task. Better productivity means more work is done, and therefore usability saves time, which in turn saves money.
- A formal usability test provides evidence that the product is not defective and lives up to expectations. This can be important for lawful purposes.
- Besides monetary gains, a more usable product contributes to better quality resulting in a better relationship between developers and consumers, which in turn ensures patronage.
- In addition, a usable product gives comfort to the user, making them less stressed and allowing them to enjoy using the software, even if the product is not meant for entertainment purposes.

There are several methods proposed in the literature that can be employed for assessing usability [4]. One such method, which is the focus of this research, is Feature Analysis. This method encompasses the evaluation of an application by considering key features, their importance and their effect on usability. This is generally accomplished using some score calculation. Such method is useful not only for measuring

D. Lowe and M. Gaedke (Eds.): ICWE 2005, LNCS 3579, pp. 53–62, 2005.

usability and for comparison with other systems, but also to provide detailed results indicating which areas or features need further improvement.

Feature analysis within a Web engineering context was first used by Olsina and Rossi [6]. They propose a Website quality evaluation method (WebQEM), which uses a feature analysis technique to calculate a score that measures the quality of a Website. The feature analysis technique employed is called Logic Scoring Preference (LSP), and will be detailed in Section 2. WebQEM bases its feature list on the ISO 9126 quality model [5], with the highest-level features as Usability, Functionality, Reliability, and Efficiency.

Unlike Olsina and Rossi [6], this research focuses only on a subset of Web quality measurement, that is, Web usability measurement. We conducted two formal experiments to investigate the following:

- To what extent the LSP method captures the subjective views of users regarding Web usability.
- To what extent the usability scores obtained using LSP are consistent across subjects with similar experience using the Web, for the same Website.

The results obtained from both experiments did find a significant difference between the scores obtained using LSP and those based on subjective opinion. However, we obtained contradictory results when investigating the consistency of LSP scores across subjects.

The remainder of this paper is organised as follows: Sect. 2 introduces the LSP method. Sect. 3 presents our research method and the hypotheses we investigated. The data analysis is described in Sect. 4, followed by a summary and discussion of the results in Sect. 5. Finally, our conclusions and comments on future work are presented in Sect. 6.

2 Logic Scoring Preference

The Logic Scoring Preference method, or LSP, was proposed in 1996 by Dujmovic, who used it to evaluate and select complex hardware and software systems. The purpose of LSP is to evaluate features quantitatively (by means of logic scoring) for the comparison of different entities (e.g. software systems, applications) [1],[2],[3].

In LSP, the features are decomposed into aggregation blocks. This decomposition continues within each block until all the lowest level features are directly measurable. This is illustrated in Fig. 1:

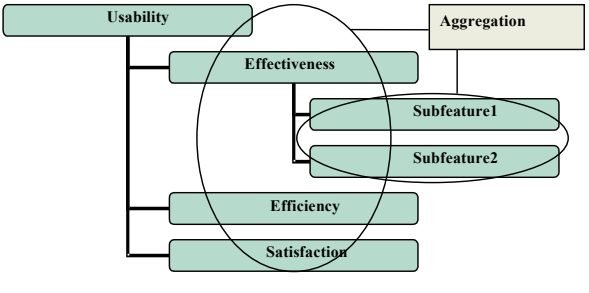

Fig. 1. Aggregation Blocks

Thus, a tree of decomposed features at one level will have a number of aggregation blocks, each resulting in a higher-level feature going up the tree right through to the highest-level features (see Fig. 1).

Next, for each feature, an elementary criterion is defined. For this, the elementary preference E_i needs to be determined by calculating a percentage from the feature score X_i. This relationship is represented in the following equation:

$$E_i = G_i(X_i) \tag{1}$$

Where E is the elementary preference.

G is the function for calculating E.

X is the score of a feature.

i is the number of a particular feature.

One way to evaluate the elementary criterion is by use of a preference scale. In this scale, a cut-off point needs to be defined on either side of the scale. For example a scale related to response time may use as cut-off points a response time of 1 second or less for a score of 100%, and 6 seconds or over for a score of 0% [3].

The elementary preferences for each measurable feature in one aggregation block are used to calculate the preference score of the higher feature. This in turn is used with the preferences scores of an even higher feature, continuing right up until a global preference is reached. The global preference is defined as:

$$E = L(E_1, \ldots, E_n) \tag{2}$$

Where E is the global preference.

L is the function for evaluating E.

E_n is the elementary preference of feature n.

n is the number of features in the aggregation block.

The function L yields an output preference e_0, for the global preference E, or any subfeature E_i. Its formula is:

$$e_0 = (W_1E_1^r + \ldots + W_kE_k^r)^{1/r},\ W_1 + \ldots + W_k = 1 \tag{3}$$

Where e_0 is the output preference.

W is the weight of the particular feature.

E is the elementary preference of a feature.

k is the number of features in the aggregation block.

r is a conjunctive/disjunctive coefficient of the aggregation block.

For each E_i a weight W is defined for the corresponding feature. The weight is a fraction of 1 and signifies the importance of a particular feature within the aggregation block.

To illustrate the use of LSP we provide an example [3]:

Assuming an aggregation block consisting of 3 inputs, *x*, *y*, and *z*. Their weights are 0.5, 0.3, and 0.2 respectively, and their elementary scores are 0.7, 0.9, and 0.6 respectively.

The chosen conjunction value for this aggregation block is C-+, with which the *r*-value is -0.208. For details on how the *r*-values are calculated please refer to [2],[3].

The equation is therefore:

$$(0.5\times0.7^{-0.208} + 0.3\times0.9^{-0.208} + 0.2\times0.6^{-0.208})^{(-1/0.208)} = 0.730255 \tag{4}$$

The parent feature of this aggregation block now has the score of 0.730255, which will be used for evaluating the score of the aggregation block in which the parent feature belongs to.

The *r* coefficient represents the degree of simultaneity for a group of features within an aggregation block. This is described in terms of conjunction and disjunction. Conjunction refers to how desirable it is that the features within an aggregation block should exist together, while disjunction is the antonym of conjunction. The formula for e_0 represents certain types of mathematical means when certain values for r are used. For example, when *r*=1, the formula is the same as that of a regular arithmetic mean, and when *r*=2 the formula yields the square mean. The other two means mentioned are the geometric and harmonic means. Dujmovic presents 20 such functions [3] and an abridged list of 9 generalised functions [2].

3 Research Method and Hypotheses

The overall research question in our study was to assess the usefulness of using LSP for Web usability measurement. To address this question we have compared LSP scores to scores obtained from subject opinion, and also looked specifically at how similar LSP scores were from one another, given the same Website and users with similar experience.

We refined our research question in two null hypotheses, which are as follows:

H_{A0} – The usability scores obtained using LSP for the same Website are similar across subjects with similar experience in using the Web.

H_{B0} – Usability scores obtained using LSP are not significantly different from usability scores obtained via subjective opinion for the same Website, for subjects with similar experience in using the Web.

Our alternative hypotheses, i.e., what we expected to occur, were then stated as:

H_{A1} – The usability scores obtained using LSP for the same Website are not consistent across subjects with similar experience in using the Web.

H_{B1} – Usability scores obtained using LSP are significantly different from usability scores obtained via subjective opinion for the same Website, for subjects with similar experience in using the Web.

The dependent variable in both experiments was the final score given to a Website. The independent variables were: subjects' experience in usability assessment and usability measurement technique (LSP or subjective opinion).

There were also several confounding factors that we had to take into account, some of which we were able to control, which were as follows:

- Subjects' understanding of the method
- Subjects' computing skills level
- Subjects' previous experience using the Web
- Subjects' understanding of English
- Type of Website (e.g. e-commerce, academic)
- Server load
- Internet speed
- Environment, Location
- Time, instance, date of evaluation
- Computers used (e.g. processor speed, display unit, input methods)

A confounding factor is a variable that can hide a genuine association or incorrectly suggest the existence of an association between variables. If not taken into account, confounding factors can bias the results of a study.

Except for 'type of website', we were able to control on both experiments all the confounding factors we identified. Table 1 provides details on the methods used to control the confounding variables.

We had planned to use a single type of Website in both experiments to control one of the confounding factors (Type of Website). Our choice was to use a Website of a New Zealand tertiary Institution (Otago University). Unfortunately, due to technical problems beyond our control, we had to use the University of Auckland's Website on our first experiment, since this was the only website we had access to since we were restricted to access only our intranet. Further discussion on this issue is provided in Section 5.

In terms of both experiments' design, we used a one-factor, two-treatment design. The factor was represented by subjects' previous experience using the Web. The two treatments were the two usability assessment techniques: LSP versus subjective opinion. It was not possible to have a control object in any of our experiments since we did not have a real 'placebo' treatment, similar to what is used in medical experiments. A control represents 'absense of', however even a subjective opinion still affects the outcome, which is the final website score. Our experimental objects were the Websites assessed, and the experimental subjects were the students who volunteered to participate.

For each of the experiments data was gathered using two questionnaires, one for LSP and another for subjective assessment. The LSP questionnaire was organised in three parts, as follows:

- Part I asked subjects about the relationship between features. These features are the same as the usability features suggested in [6]. Information received from the first part includes features' weights, and their simultaneity between groups.
- Part II asked subjects to identify upper and lower thresholds for each feature. These would identify cut-off points for each feature, which represent acceptable and unacceptable values.
- Part III asked subjects to evaluate a given website based on the measurable features from the first two parts, using the scales that they have defined in part two.

The subjective assessment questionnaire asked subjects to rate a given website's usability using a 100-point scale (0% means completely useless; 100% means absolute best).

Both questionnaires were implemented as Web forms and the data was stored on a relational database. This was done to facilitate data analysis. Two pilot studies were carried out beforehand to validate these questionnaires and to make sure subjects would use no more than 20 to 30 minutes to assess the website(s) and fill-out the questionnaires.

Regarding the size of our samples, we had 10 subjects in our first experiment and 12 in the second. We emailed out invitations to our third-year and postgraduate computer science students and 22 subjects in total volunteered to participate. We are aware that our samples were self-selected rather than random, however this was the only way to obtain participants to both experiments.

Table 1. Confounding factors which were controlled on both experiments

Confounding Factors	Method
Subjects' understanding of the method	Questionnaires that did not require previous knowledge of either LSP or subjective assessment.
Subjects' computing skill levels	Sample included only third-year and postgraduate computer science students.
Subjects' previous experience using the Web	Sample included only third-year and postgraduate computer science students.
Subjects' understanding of English	Previous to the experiments subjects had to rate themselves on their understanding of English. All rated themselves high, later confirmed by one of the authors.
Server load	A single server hosting the questionnaire, single servers hosting the websites.
Internet speed	Internet speed was the same for all computers.
Environment, Location	A single laboratory was used
Time, instance, date of evaluation	All subjects participated in the experiment at the same time and place.
Computers used (e.g. processor speed, display unit, input methods)	All computers had the same configuration and speed.

Both experiments were conducted using the same laboratory, however within a few weeks from each other. One of the authors managed the execution of both experiments.

4 Data Analysis

All the statistical results were obtained using SPSS v.10.1. Statistical tests were selected based on the type and distribution of the data. Our dependent variable was measured on a ratio scale however to decide on which test to use we also had to determine if the distribution of scores was normally or non-normally distributed. We employed the Kolmogorov-Smirnov nonparametric test to test for normality. All significance levels were set at 0.05. A significance level is used as a cut-off point to determine if a null hypothesis should be rejected or not. Generally significance levels are set at 0.1, 0.05 and 0.01.

Other statistical tests used were the two-independent samples t-test (2-TT) and the Mann-Whitney test for independent samples (2-MW). Both are used to compare two independent samples to see if there are significant differences between their values distribution. If there is then we reject the null hypothesis.

4.1 First Hypothesis – H_{A0}

Our first hypothesis was solely related to the LSP scores obtained. This hypothesis is as follows:

H_{A0} – The usability scores obtained using LSP for the same Website are similar across subjects with similar experience in using the Web.

We employed the Mann-Whitney test to compare the 10 scores obtained from experiment 1 to the sample's mean of 8.12 (see Fig. 2).

Fig. 2 shows that the both significances were below the 0.05 threshold, indicating that LSP scores could not come from the same distribution as the mean-based values. What this means is that LSP scores were not in fact similar across subjects for a given Website. These results provide evidence to reject the null hypothesis H_{A0} for experiment 1.

We repeated the same procedure for experiment 2, however this time we used the two-independent samples t-test to test our hypothesis since the LSP scores were normally distributed. Here the mean was 10.94 and we had 12 LSP scores.

Fig. 3 shows different results to those shown in Fig. 2, indicating that for experiment 2 LSP scores were similar across subjects for a given Website. These results did not provide evidence to reject the null hypothesis H_{a0}.

Our first experiment rejected the null hypothesis and our second experiment did not. The difference between these two experiments, apart from the subjects who volunteered to participate, is the Website evaluated. Experiment 1 used the University of Auckland's website, which was already well known to all participants. However experiment 2 used a website from another tertiary Institution, which was unknown to most participants. We believe that one possible explanation for the largely different LSP scores for experiment 1 may be a previous opinion towards the Website, which may have biased the results.

Test Statistics [b]

	LSPEXP1
Mann-Whitney U	20.000
Wilcoxon W	75.000
Z	-2.515
Asymp. Sig. (2-tailed)	.012
Exact Sig. [2*(1-tailed Sig.)]	.023 [a]

a. Not corrected for ties.

b. Grouping Variable: VAR00001

Fig. 2. Results for Mann-Whitney U test for Experiment 1 for H_{A0}

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
LSPEXP2	Equal variances assumed	5.831	.025	.001	22	.999	.0059	7.54079	-15.63269	15.64458
	Equal variances not assumed			.001	11.000	.999	.0059	7.54079	-16.59121	16.60310

Fig. 3. Results for T-test for Experiment 2 for H_{A0}

4.2 Second Hypothesis – H_{B0}

Our second hypothesis was related to the LSP and subjective scores. This hypothesis is as follows:

H_{B0} – Usability scores obtained using LSP are not significantly different from usability scores obtained via subjective opinion for the same Website, for subjects with similar experience in using the Web.

To test this hypothesis we had to compare LSP scores to the subjective scores. For the first experiment we used the Mann-Whitney test to compare the 10 LSP scores to another 10 subjective scores since LSP scores were not normally distributed (see Fig. 4).

Fig. 4 shows that the both significances were below the 0.05 threshold, indicating that LSP scores could not come from the same distribution as the subjective scores. The subjective scores were in fact much greater than LSP scores. What this result suggests is that LSP scores were significantly different from subjective scores, thus providing evidence to reject the null hypothesis H_{B0} for experiment 1.

We repeated the same procedure for experiment 2, however this time we used the two-independent samples t-test to test our hypothesis since the LSP and subjective scores were normally distributed (see Fig. 5).

Test Statistics [b]

	LSPEXP1
Mann-Whitney U	2.000
Wilcoxon W	57.000
Z	-3.804
Asymp. Sig. (2-tailed)	.000
Exact Sig. [2*(1-tailed Sig.)]	.000 [a]

a. Not corrected for ties.

b. Grouping Variable: VAR00001

Fig. 4. Results for Mann-Whitney U test for Experiment 1 for H_{B0}

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
LSPEXP2	Equal variances assumed	.838	.370	-8.433	22	.000	-69.9707	8.29699	-87.17763	-52.76381
	Equal variances not assumed			-8.433	15.437	.000	-69.9707	8.29699	-87.61186	-52.32959

Fig. 5. Results for T-test for Experiment 2 for H_{B0}

Fig. 5 shows similar trends to those shown in Fig. 4, i.e., that LSP scores were significantly different from subjective scores. This result also provides evidence to reject the null hypothesis H_{B0} for experiment 2.

Both experiments provided evidence to reject the null hypothesis H_{B0} and to support the alternative hypothesis H_{B1}, thus showing that usability scores obtained using LSP are significantly different from usability scores obtained via subjective opinion for the same Website, for subjects with similar experience in using the Web.

5 Summary and Discussion of Results

There are three types of validity that may influence the outcomes of an experiment: internal, external, and construct validity. Internal validity represents to what extent

conclusions can be drawn about the causal effect of the independent variables on the dependent variables. Except for type of Website, we have controlled all confounding factors. In addition, we have assigned subjects to treatments randomly. However. We are aware that using a website in experiment 1 which was well-known to the participants may have biased the results we obtained for that experiment. Unfortunately, we were unable to do anything about that since it was a technical problem that was outside our control.

Construct validity represents to what extent the variables precisely measure the concepts they claim to measure. Usability was measured using final scores from applying LSP or subjective assessment. However the set of usability features we used to calculate LSP was a subset of all usability features we had identified. We did not use the full set otherwise it would take subjects too long to carry out the evaluation, reducing even more our sample sizes.

External validity represents the domain to which a study's findings can be generalised. We used self-selected samples of students that not necessarily are representative of real users. However this was the only choice we had given the circumstances.

The results obtained for experiments 1 and 2 regarding hypothesis H_{A0} were contradictory. However given that experiment 2 used an unfamiliar website we believe that results obtained by this experiment be more representative, i.e., LSP scores are similar given the same website and subjects with similar experiences.

As for H_{B0} both experiments rejected the null hypothesis, suggesting that the usability scores obtained via LSP do not correspond to users' subjective perception of usability. Our results however provide no means of measuring which technique truly measures the usability of a website. The subjective scores tended to be a lot closer to the mean, and thus more likely to yield a repeatable result with a smaller range of values. However, this does not mean that the subjective scores accurately represent the true usability of the website.

6 Conclusions and Future Work

This paper has presented the results of two formal experiments that investigated Web usability measurement. Both experiments tested the same hypotheses. The first hypotheses tested to what extent LSP scores varied broadly given the same website and subjects with similar experience. The second hypothesis tested to what extent the scores obtained using LSP represent the subjective opinion users have regarding the usability of a website.

Our experiments rejected the second hypothesis, however presented contradictory results for the first hypothesis. Further replications of this experiment are necessary in order to validate further our findings. This will be the subject of future work.

References

1. Dujmovic, J. J. Criteria for computer performance analysis. Procs. ACM SIGMETRICS, Boulder, Colorado, United States, (1979).
2. Dujmovic, J. J. A cost-benefit decision model: analysis, comparison and selection of data management. ACM TODS, 12(3), (1987), 472-520.

3. Dujmovic, J. J. A Method for Evaluation and Selection of Complex Hardware and Software Systems. Procs. 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computer Systems, Turnersville, New Jersey, (1996).
4. Faulkner, X. Usability Engineering (1st ed.). Houndmills, Basingstoke, Hamphshire, London: Macmillan Press LTD, (2000).
5. ISO. Software Engineering - Product Quality, 9126-1 (pp. 25): International Organisation for Standardisation, (2001).
6. Olsina, L. A., & Rossi, G. Measuring Web application quality with WebQEM. IEEE Multimedia, *9*(4), (2002), 20-29.