The Damage Index: An aggregation tool for usability problem prioritisation.

Gavin Sim
CEPS
University of Central Lancashire
Preston
44 1772 895612
grsim@uclan.ac.uk

Janet C Read
Chici
University of Central Lancashire
Preston
44 1772 893285
jcread@uclan.ac.uk

ABSTRACT

The aggregation of usability problems is an integral part of a usability evaluation. Numerous problems can be revealed and given that there are usually limited resources for fixing or redesigning the system then prioritisation of the problem set is essential. This paper examines the prioritisation of usability problems from a single heuristic evaluation and multiple heuristic evaluations of Questionmark Perception, a computer assisted assessment application widely used within educational institutions. Two different methods for prioritisation are critiqued; one based on the severity ratings alone and the other on a Damage Index formula proposed by the authors. The results highlight the difference in ranking of problems dependent upon the approach taken. The Damage Index offers a method of systematically prioritising the usability problems in a repeatable way, removing subjectivity from this process, therefore offering improvements over just the reliance upon the severity ratings alone.

Categories and Subject Descriptors

H.5.2 [**User Interface**]: Evaluation / Methodology

General Terms

Measurement, Design, Human Factors,

Keywords

Usability, Consolidation, Aggregation, Usability Problems, Severity, Evaluator Effect.

1. INTRODUCTION

Usability evaluation methods are well researched and it has been claimed that usability is a key issue in HCI [1]. Various usability evaluation methods exist, for example the RITE method [2], heuristic evaluations [3], observations and experimental methods. The primary objective of usability evaluations is to identify usability problems within the system and in some instances such as the RITE method suggest improvements or solutions. However often there are only enough resources to fix a subset of the reported usability problems, therefore prioritisation is an important aspect of the process [4]. When performing an evaluation it is therefore important to maximise the yield per evaluation, ensuring that the usability problems reported offer sufficient coverage of the system. It is impossible to have closure on the problem set thus making it important to reveal as many real problems within the evaluation.

When conducting usability evaluations it is well documented that a single evaluator will identify less usability problems than two or more evaluators [5] [6]. The optimum number of evaluators was originally reported to be between 3 and 5 and it was reported that increasing the number of evaluators beyond 5 would result in a diminishing return, thus becoming inefficient. However this optimum number has since been questioned [7] and if novice evaluators are used or a complex system is being evaluated then 5 may be less than the optimum number required [8].

Once problems have been identified by evaluators, the usual process is to aggregate the individual reports into a single list of usability problems within the system. This process usually involves merging problems reported by multiple evaluators, retaining unique problems and discarding irrelevant problems. Once the problems are aggregated, dependent on the evaluation method, the problem set is prioritised to facilitate the redesign of, the system, or aid in identifying the problems that need to be fixed before the system is released. For example in a heuristic evaluation a severity rating is usually attached to each problem to identify the most severe problems. There are different severity rating scales that can be applied including Nielsen’s scale, a five star system [9] and a five point scale ranging from -3 to +1 [10]. Severity ratings have also been incorporated into other methods such as think aloud whereby once a problem had been identified the researchers asked the evaluators to attach a severity rating to the problem [13].

The reliability of evaluator judgements of the severity of a problem is low and it has been suggested that it would be not advisable to base any major investment of development time on the results of a single evaluator [20]. The reliability of severity ratings from multiple evaluators is also low in one study only 35% of the severe problems were rated as severe once [11] and in another study 56% of the problems were rated severe by only one evaluator [7]. Therefore if the data from usability evaluation studies are to be relied upon improvements need to be made to the aggregation and prioritising of the data sets from the evaluations. Without this resources may be wasted fixing unnecessary problems or trying to determine the problems that need to be prioritised before release.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference’10, Month 1–2, 2010, City, State, Country.

Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00.

Within HCI there has been little published in the area of aggregating and prioritising usability problems. It has been suggested that the literature on usability problems consolidation is mostly described at a coarse grade level [12]. For example in a study relating to accessibility and usability [13] there is no mention of how the problems were aggregated; the total number of problems are reported as are the aggregated problems with duplicates removed. In a heuristic evaluation of video games the authors simply stated that they analysed the problems and removed repeated problems from the same evaluator, however they did not describe how they merged problems between evaluators [14]. These authors do discuss that there was significant overlap in the problems found and provide examples of these. If the data is to be used to aid the redesign of the game the results need to be in a form that the developers can understand, one that lets them easily identify the severe issues and prioritise the limited resources.

This paper presents an aggregation tool to help facilitate the prioritisation of usability problems from a single evaluation and from multiple evaluation. In comparative studies of usability evaluation methods it has been widely reported that different methods find different problems [15]. Therefore to get a true indication of the usability of a system multiple evaluations may be necessary. Once the data has been captured, it needs to be merged and the problem set prioritised. This paper examines this process by examining the data from multiple evaluations of Questionmark Perception a computer assisted assessment application widely used within educational institutions.

2. Aggregation Tool

The use of severity ratings is important to aid the prioritisation and understanding of the most severe problems within a system. It has been suggested that the severity of a problem is a combination of three factors: frequency with which the problem occurs; impact if the problem occurs; persistence of the problem. Therefore in the study by Pinelle et al. [14] when they removed repeated problems by the same evaluator they were thus removing valuable data which could aid severity judgement. In another study 75 problems were allocated a severity rating by 4 evaluators on a five point scale and the mean score was 3 and there was little variation between the scores as the standard deviation was 0.8 [1] This would make it very difficult to prioritise the problem set and easily identify the problems with the potential impact. A common element amongst these evaluations is that they have a tendency to use the mean severity score and this will form the basis of the aggregation formula proposed.

Using a formula to prioritise the data set from an evaluation will enable the reliable prioritisation of problems in a systematic way. Using the notion that the damage or impact to an individual user of a system could be estimated based on the frequency of discovery and severity the following formula was devised see Figure 1

Figure 1 Damage Index Formula

$$DI = \frac{\bar{s} * n}{yN}$$

The formula is comprised of the following parameters:

- Damage Index = DI
- Mean Severity Score = $\bar{s}$
- Number of groups or individuals that identified the problem = n
- Upper bounds of severity rating scale = y
- Group size = N

The Damage Index would produce a ratio for each problem and the basis for the formula is problems with a high probability of being discovered and high severity rating are likely to cause the user the most difficulties. The mean severity rating for a problem would be calculated and this would be multiplied by the number of groups (if used within multiple evaluations) or by the number of individuals who reported it as a problem. As reported in the introduction there are a variety of different severity rating scales which can be applied therefore y represents the upper bound of the severity rating scale. For example if Nielsen's severity rating scale is used the upper bound is 4 (usability catastrophe). By not specifying a specific value this ensures that the formula is generic. The final parameter N is the group size, for example if used within a heuristic evaluation with 5 evaluators group size would be 5 or it could represent the number of evaluations.

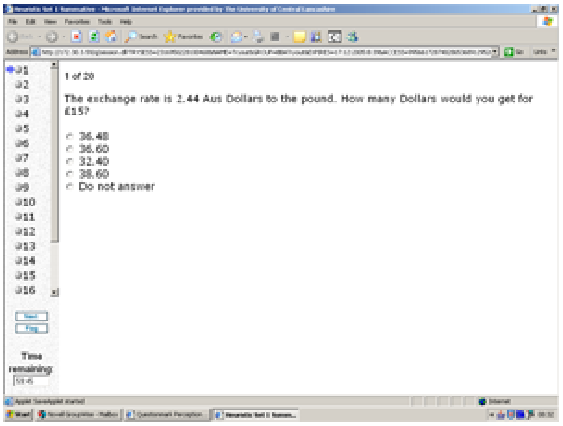
3. Single Heuristic Evaluation Case Study

A heuristic evaluation was performed on Questionmark Perception - the full results have been published [16]. The original study was an exploratory study exploring usability issues associated with Computer Assisted Assessment and whether context of use affected severity judgment. In this paper the data from the original study is re-analysed using the Damage Index proposed above.

Eight evaluators were recruited to the study. Four of the evaluators were lecturers in HCI and were thus considered to be experts in HCI as well as being familiar with the assessment domain (Double Experts). The other four evaluators were research assistants from within the Faculty of Design and Technology and had no prior knowledge of heuristic evaluations or of computer assisted assessment (this was asked informally) but may have contextual knowledge. The novice evaluators were given a brief introduction to the heuristic evaluation method before participating in the study. They were given an overview of the heuristic set, the procedure and the data capture forms.

The evaluators performed a heuristic evaluation on Questionmark® Perception version 3.4 using, to give meaning to the evaluation, a test that was created for the purpose of the evaluation see Figure 2.

Figure 2 Questionmark Perception Test Interface



1 of 20

The exchange rate is 2.44 Aus Dollars to the pound. How many Dollars would you get for £15?

○ 36.48
○ 36.60
○ 32.40
○ 38.60
○ Do not answer

Whilst completing the tasks, the evaluators were required to record any usability problems encountered on a form provided.

Once evaluators completed the task they then matched each problem to an appropriate heuristic and suggested a severity rating. The researcher collected in the completed forms. Due to time constraints it was not possible for the evaluators to aggregate their own problem sets.

3.1 Analysis

The analysis of the data was performed by just the first author of this paper. Each of the problems recorded by the evaluators was examined to establish whether it was a unique problem (one that no other person recorded). If a problem was recorded by more than one evaluator this was aggregated into a single problem. The aggregated list was returned to the individual evaluators to attach severity ratings to each problem and the mean severity rating for each problem was again calculated and rounded to the nearest whole number – this resulted in a new (mean) severity rating. This data was then used to calculate the Damage Index for each of the unique problems reported from the evaluation.

3.2 Results

For summative assessment there was a total of 48 recorded problems, these were then aggregated to leave 41 unique problems, with 5 being identified by more than one evaluator (some by two or more). The decision was made to only aggregate problems where there was a clear similarity between the reported problems. For example one evaluator reported *Would have been nice if one button did all the 'Do not want to answer' etc.* and another person stated *the buttons were too small.* This could have easily been aggregated into a single problem entitled '*problems with buttons*' but ultimately these two problems would require two separate fixes therefore they were deemed to need to be treated independently.

An example of a merged problem is that three evaluators reported that th*ere should be more spacing between the answers in multiple choice style questions* but just using slightly different terminology - this was reworded as *PR5 Poor presentation of text makes it difficult to read*. In another example, four evaluators expressed concern over *being penalised for spelling in text entry style questions*.

The frequency of the severity ratings for these problems, calculated by averaging the ratings from the evaluators are shown in Table 1.

Table 1 Severity Ratings for each reported problem.

Context	**Severity**				
	0	**1**	**2**	**3**	**4**
Summative	0	15	21	5	0

As can be seen in Table 1, there were 5 problems with a severity rating of 3 - 'major usability problem' and these were:

- PR1 - No option to quit
- PR2 - When all questions attempted finish appears. It exits without confirmation and doesn't check whether and flags are still set
- PR3 - A user thought they put in the correct answer but got an error message, and could not find a solution so had to quit
- PR4 - Lost exam answers message came up Page Expired
- PR5 - Poor presentation of text makes it difficult to read.

If the purpose of the evaluation was to aid the redesign of the software, these problems would be the first to be fixed but it would then become difficult to determine where resources should be utilized next, as there are, 21 problems classified with a 2.

The Damage Index formula was then applied to the problems. The most severe problem based on the Damage Index formulae was PR5 *Poor presentation of text makes it difficult to read,* this could clearly have an impact on test performance and may potentially cost students marks. This problem was also reported as one of the problems classified with a severity rating of 3 and would have likely been fixed if the results were based on severity alone. Each evaluator's severity rating for this problem is shown in Table 2.

Table 2 Evaluator Severity Rating of a Problem

Evaluator									
	E1	E2	E3	E4	E5	E6	E7	E8	Mean
Severity	3	4	3	3	1	2	2	3	2.63

The Damage Index was calculated using the frequency of discovery which in this instance was 3, the mean score, the upper bound of the severity rating which was 4 and the group size was 8. The calculation and Damage Index value is shown in Figure 3 below.

Figure 3 Damage Index value for PR5

$$0.25 = \frac{2.63 * 3}{4 * 8}$$

However the second most severe problem based on the Damage Index was *answers were marked wrong due to spelling mistakes,* and if the reliance was on severity ratings alone then this might not necessarily have been fixed or prioristised, as it ended up having a mean severity rating of 2 and was 1 of 21 problems with this rating.

The top five problems based on the Damage Index are:

- DI = 0.25 PR5 - Poor presentation of text makes it difficult to read
- DI = 0.20 PR6 - Answers were marked wrong due to spelling mistakes
- DI = 0.16 PR3 - A user thought they put in the correct answer but got an error message, and could not find a solution so had to quit
- DI = 0.09 PR2 When all questions attempted finish appears. It exits without confirmation and doesn't check whether and flags are still set
- DI = 0.08 PR1 - No option to quit

The Damage Index thus provides a different way of prioritising a problem set. When comparing the two approaches, 4 of the 5 problems in the top 5 by Damage Index sorting had a severity rating of 3 indicating a reasonable level of overlap. This would suggest that the Damage Index is able to identify the severe problems and potentially prioritise these in a more effective manner. Table 3 shows the frequency a problem was classified to a range within the Damage Index.

Table 3 Damage Index for each reported problem

	Damage Index						
	0-.020	.021-.040	.041-.060	.061-.080	.081-.10	.10-.20	.2+
Number of problems	0	8	19	10	1	2	1

3.3 Discussion

When used in a single evaluation the Damage Index enables the aggregated list of problems to be prioritised in a process that may reduce subjectivity. However there is inevitability still subjectivity in the initial aggregation process in determining whether two problems reported are the same.

There is however a difference between the problems that are ranked in the top 5 when comparing the performance of the Damage Index and severity ratings. The frequency of discovery is an important aspect of the Damage Index. This resulted in PR6 being ranked 2nd in the Damage Index whilst this was 1 of 21 problems with a severity rating of 2 and may not have been addressed in a redesign.

To improve the results of a usability evaluation it may be necessary to perform multiple evaluations. Research has shown that multiple evaluations reveal different problems [17] and it is not possible to have closure on the problem set. Therefore to improve the quality of the problem set multiple evaluations may be necessary and this will inevitably increase the complexity of aggregating and prioritizing the problem set.

4. Multiple Heuristic Evaluation Case Study

This study re-analysed data sets from a study to evaluate the usability of 3 commercial CAA applications TRIADS, WebCT and Questionmark Perception [18]. The original study was designed to compare the 3 applications and produced large data sets of usability problems for each of the 3 applications. For the purpose of this paper only Questionamark Perception is re-analysed and the Damage Index is applied.

In total 31 HCI undergraduate students participated in the initial evaluation. As per the single case study Questionmark® Perception version 3.4 was used to deliver the test using the same interface layout but with different questions.

The evaluation was performed over a three week period. In the first week, all the evaluators were given a brief overview of heuristic evaluations and taken through Nielsen's heuristics and the use of severity ratings in the lecture. In the second week the evaluators went to one of the computer laboratories within the Department of Computing to perform the heuristic evaluation. They were given a form on which to record the usability problems found. The form required the evaluators to state what the problem was, which heuristic(s) was violated, where each heuristic was violated, and how the problem was found. The form was based on a design described in [19] and it also required the evaluators to record the severity of the problem. One week later, the evaluators participated in the final stage of the study. At this point they were required to aggregate their results into a single list of problems.

4.1 Analysis

The analysis of the data was quite complex. At the beginning, the data was made up of 31 individual evaluation sheets, each containing a list of problems and severity ratings for Questionmark Perception.

In the second week of the study, the students clustered into small groups based on the software they had evaluated and aggregated the problems they had found. This resulted in 8 aggregated lists of usability problems, severity ratings and frequencies of discovery.

Following the student aggregation both authors of this current paper performed a card sorting exercise to merge the data between groups. This was a time consuming process taking several hours as each problem was analysed to see if it matched another problem from one of the other groups. A final aggregated list of problems was produced along with a mean severity score (calculated in the same way as in the first study reported in this paper).

4.2 Results

There were a total of eight different groups and the results from the evaluations are presented in Table 4 below.

Table 4 Number of problems reported by each group

		Stage 1	**Stage 2**

Group	Students	Problems	Problems
A	6	14	
B	4	15	
C	2	13	
D	3	13	
E	3	13	
F	4	13	
G	5	14	
H	4	15	
Total		110	77

In total there was 110 usability problems reported by the 8 groups and it is inevitable that there would be some overlap. The data set was further analysed by the authors and reduced to 77 through a process of merging duplicate problems. No problems were discarded during the aggregation process.

4.2.1 Between Groups

The mean severity rating can be calculated based on the aggregated data from the groups. For example Group D reported that *Text boxes are unclear how to answer them no prompt to type* and gave it a severity rating of 3, whilst the same problem was reported by Group E who gave it a severity rating of 2 and this was averaged giving it a severity rating of 2.5, then rounded to 3 to ensure it matched Nielsen's severity rating scale.

The mean severity ratings for each of the resulting 77 problems are shown in Table 5.

Table 5 Severity rating for each problem after stage 2

	Severity				
	0	**1**	**2**	**3**	**4**
Problems	0	11	32	21	14

With a large problem set it becomes even more difficult to prioritise the usability problems. In this instance there are 14 problems with a severity of 4 (usability catastrophe) and 9 of these problems were only identified by 1 of the 8 groups. The 5 problems which were reported by more than 1 group were:

- P1. You could finish the test and submit your answers even if some questions hadn't been attempted - should have prompted you.
- P2. Question 17 wording difficult
- P3. No key to explain the colour coding red = unanswered
- P4. Some questions are worth more marks
- P5. Test was out of 45 yet only 18 questions

If the Damage Index is then applied to the problem set, the frequency of discovery will be the number of groups that have discovered the problem. Factored into this there will also be the mean score, the upper bound of the severity rating, which was 4 and the group size of 8. The frequencies a problem was classified to a range within the Damage Index are displayed in Table 6.

Table 6 Between Groups Damage Index Score and Frequency

	Damage Index						
	0 - .04	.041 - .08	.081 - .12	.121 - .16	.161 - .20	.21 - .40	0.4+
Number of problems	11	32	9	11	4	5	5

The 5 problems with the highest Damage Index are:

- D1= 0.563 P6. - Q7 onwards does not specify how many boxes to tick
- DI = 0.563 P1. - You could finish the test and submit your answers even if some questions hadn't been attempted - should have prompted you
- DI = 0.531 P7 - q13-18 required perfect character entries else would be marked wrong
- DI = 0.50 P4 - Some questions are worth more marks
- DI = 0.41 P8 - 13-18 the input boxes had a drop down menu arrow, confusing the user

There are only two problems P1 and P4 that both the Damage Index and severity rating scale identified in the top 5. For example P6 the problem with the highest Damage Index had a mean severity rating of 3 shown in Table 7.

Table 7 Severity Ratings by Groups for P6.

	Group						Mean
	A	B	C	E	F	G	
Severity	2	4	4	3	3	2	3

P6 may not have been prioritized as a fix if the reliance was on mean severity ratings alone. This adds further support to the fact that the reliance on mean severity ratings alone may not be a sensible solution to prioritising the problem set.

4.2.2 Between Evaluators

The previous section analysed the data based on the number of groups, however it is also possible to evaluate the data based on the number of evaluators who participated in the heuristic evaluation which in this study was 31. The Damage Index formula would be modified with the group size now being 31 to represent each of the evaluators who performed the heuristic evaluation irrespective of which group they were allocated to.

Again the frequency in which a problem is classified to a range within the Damage Index is displayed in Table 8.

Table 8 Between Evaluators Damage Index and Frequency

	Damage Index						
	0 - .01	.011 - .02	.021 - .04	.041 - .08	.081 - .12	.121 - .20	.2+

Number of problems	8	20	27	8	5	4	5

Using this approach, the problem with the highest frequency of discovery was P6. *Q7 onwards does not specify how many boxes to tick* with a total of 17 evaluators reporting this problem. This was also the problem with the highest Damage Index value of 0.411 and the 5 problems with the highest Damage Index scores are presented below.

- D1= 0.411 P6. - Q7 onwards does not specify how many boxes to tick
- DI = 0.29 P1. - You could finish the test and submit your answers even if some questions hadn't been attempted - should have prompted you
- DI = 0.274 P7 - q13-18 required perfect character entries else would be marked wrong
- DI = 0.218 P9. - There wasn't any help information throughout the test
- DI = 0.21 P8 - 13-18 the input boxes had a drop down menu arrow, confusing the user

Four of the 5 problems were the same as the between group analysis of the problem set with the exception of P9 which was, in that case, ranked 7.

4.3 Discussion

When analyzing the data from multiple evaluations it is possible to rank based on frequency between groups or evaluators. The 3 problems ranked the highest were the same between groups and between evaluators suggesting a level of consistency in the identification of the most severe problems, this is important for the prioritization process.

At this stage it is not possible to state which is the most effective method when using data from multiple evaluations but there are clearly some differences in the ranking of problems. For example if 1 evaluator from each group identified the same problem and gave it a severity of 3 using the formula this would have a Damage Index 0.75 when calculated between groups. However if this was calculated between evaluators it would have a Damage Index value of 0.19 which is considerably lower than the between group calculation and therefore may be ranked lower in a list of priorities.

There were 9 problems with a severity of 4 but these were only reported by 1 evaluator and were ranked 23-31 on the between evaluators analysis and 17-25 between groups. This is still in the upper 3rd on the ranking of the problems and the reliance on the judgment of a single evaluator has been questioned [20] as they have been shown to be inaccurate, therefore frequency should play an important role in the prioritisation process.

One of the problems with the heuristic evaluation methodology is that the reported problems are predictions and unless falsification testing is performed then it difficult to establish if it is a real problem to the end user. If all problems are maintained within the problem set then inevitably false positives will be retained. If a single evaluator out of 31 has reported a problem and given it a severity rating of 4 then there is a chance that this could be a inaccurate classification to the severity rating scale or even a false positive. Therefore incorporating frequency of discovery into the formula may help minimize the inclusion of false positives and ensure that resources are not unduly allocated to fixing or redesign features, which would not have a serious impact on the end user.

5. Conclusions

Following a usability evaluation the aggregation and prioritisation of the data is an important process yet the methods are not clearly defined or researched. Without an effective mechanism resources may be inappropriately utilized on fixing trivial problems thus costing organizations time and money. This paper has presented a formula to aid the prioritisation of data from usability evaluations. The formula could be applied to a single evaluation or multiple evaluations to prioritise the problem set. The Damage Index is original within the context of usability evaluations methods and enables the prioritisation of data in a repeatable and quantifiable way, thus alleviating bias from the evaluators in the aggregation process. It offers a clear advantage over the reliance on severity scales alone by removing subjectivity from the process and ensuring frequency is factored into the process.

If used within inspection methods then falsification testing may play an important role in helping minimize false and inaccurate severity judgments. If a single evaluator reports a problem their understanding of the impact to the end user may be miscalculated thus artificially raising the severity prediction. Falsification testing is thus important when data is derived from inspection methods, a single reported problem should not be discarded as the probability of an item being found or reported is not equal [21]. Therefore falsification testing would help improve the quality of the corpus, aid the aggregation process and help establish the true severity of a problem to the end user.

An important aspect of the evaluation process is to ensure that the data capture forms enable the frequency to be reported and that problems can be traced back to an individual's data capture forms. Without this information it would be impossible to calculate the mean severity score and establish the frequency of discovery, which are integral to the application of the Damage Index.

The implications for future work are to look at the most effective method for calculating the Damage Index when merging data between multiple evaluations. It is unclear if the parameter group size should represent the number of evaluators or the number of studies. The results did reveal similarities in the prioritisation of the data but at this stage little is known on the effectiveness of this process. It may well be that the formula could be modified to incorporate a probably density function to factor in that some problems are simply harder to find than others.

The results presented in this paper are from a series of heuristic evaluations. It would be interesting to do a comparison of the data from different evaluations methods such as observation and heuristics to establish it the prioritised sets are similar.

The formula could be modified to be used in other domains, not just for prioritising the data from usability studies. For example it could be used in the area of computer security whereby problems are identified and ranked based on the severity of the potential threat [22] and areas such as system testing or acceptance testing.

6. REFERENCES

[1] Ssemugabi, S. and R. De Villiers. 2007. A comparative study of two usability evaluation methods using a web based e-learning application. in SAICSIT. 2007. Fish River Sun, South Africa: ACM.

[2] Medlock, M.C., et al. 2002. Using the RITE Method to improve products: a definition and a case study. in Usability Professionals Association. 2002. Orlando.

[3] Nielsen, J. and Molich, R. 1990. Heuristic evaluation of the user interface. in SIGCHI conference on Human factors in computing systems: Empowering people. 1990. Seattle: ACM.

[4] Hertzum, M. 2006. Problem Prioritization in Usability Evaluation: From Severity Assessment towards Impact on Design. International Journal Human Computer Interaction. 21(2): p. 125-146.

[5] Nielsen, J. and Landauer, T.K. 1993. A mathematical model of the finding of usability problems, in Proceedings of the SIGCHI conference on Human factors in computing systems. 1993, ACM Press: Amsterdam, The Netherlands. p. 206-213.

[6] Barnum, C. 2003. The 'magic number 5' Is it enough for web testing? Information Design Journal and Document Design. 11(2/3): p. 160-170.

[7] Jacobsen, N.E. 1998. The evaluator effect in usability studies: problem detection and severity judgements. in Human Factors and Ergonomics Society 42nd Annual Meeting. 1998. Chicago.

[8] Slavkovic, A. and Cross, K. 1999. Novice Heuristic Evaluations of a Complex Interface. in CHI 99. ACM.

[9] Yehuda, H. and McGinn. J. 2007. Coming to Terms: Comparing and Combining the results of multiple evaluators performing heuristic evaluations. in CHI. 2007. San Jose: ACM.

[10] Kantner, L., Shroyer, R. and Rosenbaum. S. 2002. Structured heuristic evaluation of online documentation. in Professional Communication Conference. 2002: IEEE.

[11] Hertzum, M. and Jacobsen, N. E. 1999. The evaluator effect during first-time use of the cognitive walkthrough technique. in HCI International. 1999: Lawrence Erlbaum.

[12] Law, E.L.-C. and Hvannberg, E.T. 2008. Consolidating Usability Problems with Novice Evaluators. in NordiChi. 2008. Lund: ACM.

[13] Petrie, H. and Kheir. O. 2007. The relationship between accessibility and usability of website. in CHI2007. 2007. San Jose: ACM.

[14] Pinelle, D., Wong, N. and Stach, T. 2009 Usability Heuristics for Networked Multiplayer Games. in GROUP 09. 2009. Sanibel Island: ACM.

[15] Desurvire, H., Kondziela, J. and Atwood. M. 1992. What is gained or lost when using evaluation methods other than empirical testing. in CHI. 1992. Monterey: ACM.

[16] Sim, G., Read, J.C. and Holifield. P. 2006. Using Heuristics to Evaluate a Computer Assisted Assessment Environment. in World Conference on Educational Multimedia, Hypermedia and Telecommunications. 2006. Orlando: AACE.

[17] Lavery, D., Cockton, G. and Atkinson, M.P. 1997. Comparison of evaluation methods using structured usability problem reports. Behaviour & Information Technology. 16(4/5): p. 246-266.

[18] Sim, G., Read, J.C. and Holifield. P. 2007. Heuristic Evaluations of Computer Assisted Assessment Environments. in World Conference on Educational Multimedia, Hypermedia and Telecommunications. 2007. Vancouver: AACE.

[19] Cockton, G., Woolrych, A. and Hindmarch. M. . 2004. Reconditioned Merchandise: Extending Structured Report Formats in Usability Inspection. in CHI 2004. 2004. Vienna: ACM.

[20] Nielsen, J. 1994. Enhancing the Explanatory Power of Usability Heuristics. in Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence. 1994. Boston: ACM.

[21] Woolrych, A., Cockton, G. and Hindmarch M. Falsification Testing for Usability Inspection Method Assessment. in HCI2004. 2004: Research Press International.

[22] Whitman, M.E., 2003. Enemy at the gate: threats to information security. Communications of the ACM, 46(8): p. 91-95.