JUS
JOURNAL OF USABILITY STUDIES

Vol. 2, Issue 2, February 2007, pp. 61-75

Heuristic Evaluation Quality Score (HEQS): a measure of heuristic evaluation skills

Shazeeye Kirmani
Infosys Technologies Ltd.
Electronics City, Hosur Road,
Bangalore, India, 560100.
Shazeeye_Kirmani@infosys.com

Shanmugam Rajasekaran
Infosys Technologies Ltd.
Electronics City, Hosur Road,
Bangalore, India, 560100.
Shanmugam_R@infosys.com

Abstract

Heuristic Evaluation is a discount usability engineering method involving three or more evaluators who evaluate the compliance of an interface based on a set of heuristics. Because the quality of the evaluation is highly dependent on their skills, it is critical to measure these skills to ensure evaluations are of a certain standard.

This study provides a framework to quantify heuristic evaluation skills. Quantification is based on the number of unique issues identified by the evaluators as well as the severity of each issue. Unique issues are categorized into eight user interface parameters and severity is categorized into three.

A benchmark computed from the collated evaluations is used to compare skills across applications as well as within applications. The result of this skill measurement divides the evaluators into levels of expertise. Two case studies illustrate the process, as well as its applications. Further studies will help define an expert's profile.

Keywords

Heuristic Evaluation, Interaction Design, Information Architecture, Visual Design, Navigation, Labeling, Content, Functionality, Showstopper, Major Issue, Irritant, Heuristic Evaluation Quality Score (HEQS), Heuristic Evaluation Quality Score Percentage (HEQS%).

Introduction

Heuristic Evaluation (HE) originally proposed by Nielsen and Molich in 1990 is a discount usability engineering method conducted by a group of three to five experts. The quality of heuristic evaluation is highly dependent on the skills of these experts (Kantner and Rosenbaum, 1997). This research identifies parameters and proposes a method to measure the skills of heuristic evaluators.

The effectiveness of Heuristic Evaluation has shown to increase significantly by involving multiple evaluators, but this also increases the cost. A single evaluator finds an average of 34% of the issues in the interface, with a range of 19% to 51% (Nielsen and Landauer, 1993). Based on six studies conducted by Nielsen and Landauer (1993), the optimal number of evaluators taking into account the benefit to cost ratio was four. They suggest that three to five evaluators are ideal, but they do not mention the level of expertise.

A series of Comparative Usability Evaluations (CUE) headed by Rolf Molich (Molich, 2006) compares usability approaches and provides practical suggestions to improve the efficiency and effectiveness of the approach. In CUE-4 17, experienced evaluation teams assessed the same website; nine chose usability testing and eight chose expert reviews. The study showed that there were no differences in the results between usability testing and expert (or heuristic) reviews.

In CUE-3 11, heuristic experts individually inspected a website, and then formed four groups to combine their findings into one group report. On average, only 9% of the issues overlapped between any two evaluators. The evaluators perceived the disparities to be multiple sources of evidence for the same issue and not disagreements. The confidence in their evaluations, though highly disparate, increased after the group discussion. The authors suggest that this reinforced confidence stems from a failure to appropriately distinguish specific problems to categories of problems, threatening the reliability of such evaluations. One of the reasons for this inappropriate distinction could be the level of heuristic expertise.

Heuristic evaluation, ideally done by heuristic experts, is also done by software developers, and sometimes, by non-experts. Because of this, it is important to study the quality of the evaluator's output. A study by Desurvire, 1994 showed that experts found 29% of the critical problems in an interface, compared to 12% found by software developers, and 6% found by non-experts. These findings indicate that expertise is important for an evaluation, and its quantification would remove ambiguity.

The importance of evaluators' expertise is clearly seen in the study by Athanasis and Andreas (2001). Regular specialists, double specialists, and novices were compared for their heuristic expertise. Regular specialists are heuristic experts who are familiar with the heuristics, but not familiar with the domain being evaluated. The domain could be healthcare in the case of evaluating a pharmacy portal, or banking in the case of evaluating an online banking portal. Double specialists are heuristic experts who are familiar with both the heuristics and the domain.

Results indicated that for a particular evaluation to find 75% of the issues, 15 novices were required, three to five regular specialists were required, and two to three double specialists were required.

A study by Jacobsen et al. (1998) emphasizes the evaluator effect as the potential threat to the reliability of usability studies. They mention that both detection of usability problems and selection of the most severe problems are subject to considerable individual variability. A possible solution to minimize this variability is to choose evaluators of similar skill levels.

Table 1: Steps to measuring heuristic evaluation skills

1. Identify the application, UI parameters and the severity ratings The choice of UI parameters depends on what skills you consider essential.
2. Standardize the scope of the evaluation The scope determines the boundaries within which the evaluation should limit itself.
3. Provide a knowledge transfer (optional) If the evaluators are not familiar with the application, a knowledge transfer will help evaluators understand the application.
4. Standardize the time for the evaluation All evaluators should be given the same amount of time to complete the evaluation.
5. Standardize the evaluation format A common format to do the evaluation makes it easy to measure skills.
6. Proceed to evaluate Evaluators should perform the evaluation to the best of their ability.
7. Use the individual evaluations to arrive at a benchmark Collate the individual issues of each evaluator eliminating repeated issues. Check these issues with a group of three or more heuristic experts to arrive at the benchmark.
8. Use this benchmark to measure the skills of each evaluator Individual skills are measured using this benchmark
9. Derive insights The performance of evaluators can be compared across and within applications.

Two of the limitations of heuristic evaluation are that evaluators are subjective, or that issues found may reflect the biases and perspectives of the evaluator. A study by Law and Hvannberg, 2002 attempted to solve this by strengthening the methodology using a usability test. However, such solutions might not always be available, and one has to work toward improving Heuristic Evaluation. A possible way could be to involve evaluators above a particular skill level.

When heuristic evaluation is done in a large scale production environment, such as the 200 applications each reviewed by three to five evaluators over the past two years in the company, there is a need for assessing evaluator's overall skill quality, as well as individual strengths and weaknesses. This assessment is also useful in identifying training areas, so that the evaluator's skills can be improved over time. Assessing evaluation skills over time would help assigning the right evaluator for a project. It would also lead to systematic certification.

Certification is critical because the most popular technique used by 76% of the usability community is heuristic evaluation (UPA Survey, 2005). A study (Nielsen, 1994) that involved 11 evaluators to heuristically evaluate an application showed a cost-to-benefit-ratio of 1:48, emphasizing the assessment of heuristic evaluation skills.

None of the studies mentioned previously have addressed the expertise or the profile of an evaluator. No research seems to have focused on this issue. We are trying to address the following questions in this study:

- Can the skills of heuristic experts be measured?
- If they can, how do we measure them?
- What are some applications of this skill measurement?

Method

Steps to Measuring Heuristic Evaluation Skills

The steps in Table 1 are an effective way to measure heuristic evaluation skills. These steps created by the authors evolved after over one year and 4 skill assessment programs.

1. **Identify the application, UI parameters and the severity ratings** – The eight UI parameters or aspects that define the interface (see Table 2) are skills important to the assessment. The application should be chosen such that it contains all the UI parameters. Severity ratings define the criticality in terms of the effort required to fix the issue. It can be a 3-point rating scale or a 5-point rating scale. Evaluators need to be given accurate descriptions of UI parameters and severity ratings to remove ambiguity in the assessment.
2. **Standardize the scope of the evaluation** – The scope or the extent of the evaluation is important in applications that are huge or are a part of a suite of applications. The scope can be narrowed to evaluate a part of the application or broadened to evaluate all the applications in a suite. The scope is defined by the number of scenarios (outline of a task) and/or the number of screenshots (picture of one screen of the application). The evaluators are given full access to the application, scenarios and screenshots of the key screens. A reasonable scope for a two-hour evaluation would be 6-10 scenarios and/or 20-30 screenshots. This is based on the 200 evaluations conducted by two to five evaluators over a period of two years in the company.
3. **Provide a knowledge transfer (optional)** – If the evaluators are familiar with the application, skip this step, or else provide a knowledge transfer along with a run through of the key scenarios that were identified in the previous step. Evaluators should be given time to run through the application independently to understand workflows and interactions to understand the application completely.
4. **Standardize the time for the evaluation** – The time taken to evaluate the application should be the same across all the evaluators.

Table 2. UI parameters that define the interface

UI Parameter	**Description**	**Typical Issues Found**
Information Architecture	Accurate structuring of information into groups best matching the mental model of users.	Absence of mutually exclusive groups, menus not prioritized, inappropriate menu width and depth
Interaction Design	Interaction between the user and the interface in areas such as the behavior of interface elements, error communication, choice of controls, etc	Absence of information communication and feedback status, absence of user control by not providing actions like undo, cancel, incomplete goal and procedure communication, search ineffectiveness, inappropriateness of controls
Visual Design	Consistency and appropriateness of layout, color, font, graphics, etc	Absence of visual hierarchy and visual harmony, not using space to relate similar groups
Labeling	Appropriateness of labels as per user vocabulary	Inappropriate labels for menus, sub menus, links, controls, pages, paragraphs, categories, presence of technical jargon and abbreviations
Functionality	Availability and appropriateness of functions to serve user goals	Missing functionalities, not combining functionalities, absence of optimal steps to execute a task, functionalities not designed as per user mental models
Content	Completeness and accuracy of information on the interface	Missing content, wrong content, inappropriate presentation of information for scanability
Navigation	Ease of flow across information spaces	Inadequate navigation aids for long pages, forms, charts, and related links
Other	Comprised of Branding (representing the identity of the company), Web Guidelines Accessibility (designing for the disabled) and other issues that did not fall into the categories above.	Branding- Different versions of the logo on the site. Web Guidelines- File size and type not mentioned for downloads. Accessibility- Frames used in the design do not support screen readers.

5. **Standardize the evaluation format** – A standardized evaluation format would ensure consistency in evaluations across evaluators making it easier to assess skills. For example, evaluators can present issues in a standardized format (Table 3) that includes the UI parameter, the severity and the heuristic violated.

Table 3: The evaluation format

Issue	The presence of two links called 'Help' increases the possibility of users clicking on the wrong link.
UI Parameter	Labeling
Severity	Major Issue
Heuristic violated	Error prevention

6. **Proceed to evaluate** – Evaluators are required to conduct the evaluation to the best of their ability in a distraction-free environment.
7. **Use the individual evaluations to arrive at a benchmark** – Collate the individual issues of each evaluator eliminating repeated issues. A group of three or more heuristic experts eliminate non-issues and come to a consensus of the severity rating for each issue. Steps are detailed under the subheading, "Benchmark".
8. **Use this benchmark to measure the skills of each evaluator** – Based on this benchmark individual ratings are calculated. Steps are detailed under the subheading, "Individual Assessment".
9. **Derive insights** – Analysis of heuristic evaluation skills over time will give insights into performance of evaluators across and within applications. Further applications of this technique in a range of areas have been discussed later.

A pilot evaluation would help to strengthen the assessment process. For example, a pilot would determine if the time allocated, the choice of UI parameters, and the scope of the evaluation is appropriate.

Description of Severity Ratings

Severity ratings help separate catastrophic problems from minor problems in order to estimate effort and resources required to fix the problem. The quality of the interface can be gauged by the number of severe issues. We have identified three categories based on Nielsen's ratings (Nielsen, 1994):

- **Showstopper:** a catastrophic issue that prevents users from using the site effectively and hinders users from accomplishing their goals.
- **Major Issue**: an issue that causes a waste of time, and increases the learning or error rates
- **Irritant**: a minor cosmetic or consistency issue that slows users down slightly. It minimally violates the usability guidelines.

Weight factors

A showstopper was given a weight of five points (see Table 4), a major issue a weight of three points, and an irritant a weight of one point. We used this system for two reasons. The first reason was to motivate evaluators to find more showstoppers, thus increasing their overall score, and the second reason was to give showstoppers more importance than major issues and irritants.

Four heuristic experts were given a five-point Likert scale, ranging from not critical to very critical. The definitions of the severity ratings were also given to them, and they were asked to rate the criticality of the severity rating. An average of the ratings for each severity rating formed the basis for assigning the weight factors.

Table 4: HEQS Formulae

Individual HEQS (or HEQS) = (Total Showstoppers for an individual)*5 + (Total Major Issues for an individual)*3 + (Total Irritants for an individual)*1
Benchmark HEQS = (Total Showstoppers of the benchmark)*5 + (Total Major Issues of the benchmark)*3 + (Total Irritants of the benchmark)*1 *Benchmark is a collation of all the unique and valid issues of all the evaluators.*
HEQS% = (Individual HEQS/ Benchmark HEQS)*100

Individual Assessment

Individuals were assessed by counting the frequency of the unique issues that they identified in each of the determined UI parameters (Table 5) and by counting the frequency in each of the severity ratings (Table 6).

Table 5: Issues based on UI Parameters

UI Parameter	Benchmark	Evaluator A
Interaction Design	38	14
Visual Design	23	5
Information Architecture	4	0
Functionality	17	1
Labeling	10	2
Content	7	2
Navigation	5	5
Other	4	0

Table 6: Issues based on severity ratings

Severity	Benchmark	Evaluator A
Showstopper	23	14
Major Issue	60	12
Irritant	25	3
HEQS (see Table 6)	Benchmark HEQS = 320	Individual HEQS = 109
HEQS%	100	30
Level		2

The sum of issues identified based on the UI parameters equaled the sum of issues identified based on the severity ratings. A score called the Heuristic Evaluation Quality Score (HEQS) was determined by multiplying the frequency for each severity rating with its respective weight factor. This takes into account the number of issues based on their individual severity ratings. For example, if an evaluator identified 10 showstoppers, 20 major issues and 5 irritants the Individual HEQS = (10*5) + (20*3) + (5*1) =115. The issues identified based the severity ratings can be used to derive the HEQS, and the issues identified based on the UI parameters can be used to assess individual strengths and weaknesses.

Benchmark

The benchmark is used to compare heuristic evaluators across applications and within applications. The benchmark is a collation of all the unique and valid issues of all the evaluators. Three or more heuristic experts look at all the comments of the heuristic evaluators, individually decide if each one of them is an issue or not, and if it is an issue, they allocate a severity rating to it.

The heuristic experts gather to discuss the issues and severity ratings collectively, and if a difference is present, they discuss to arrive at a consensus. An issue can be categorized under only one UI category and one severity rating.

There were inconsistencies in categorizing the issues under the incorrect UI parameter or severity rating. Less than 15% of the issues for both case studies discussed below were categorized under the wrong UI parameter. This is primarily because the evaluators were extremely familiar with these categories. If they were incorrectly categorized the experts would categorize the issue in the correct category and assess the evaluators. The inconsistencies in severity rating categorization were discussed later. A measure of their consistency, called the inter-rater reliability, is also captured. This measures the percentage of agreement among all the heuristic experts. A high score ensures a clear understanding and communication of all parameters.

To compare evaluators within an application is simple. For example, if the benchmark has 25 showstoppers, 30 major issues, and 10 irritants, then the Benchmark HEQS is 225 (see Table 4 for formulae). The higher the HEQS%, the better the evaluator’s performance.

To compare evaluators across applications, certain precautions need to be taken. First, the number of evaluators, as well as their expertise level, needs to be similar across both groups. This aids in standardizing the benchmark, and hence, one can compare across groups. Second, the conditions of the test need to be similar in terms of time, as well as the choice of parameters and severity ratings.

Levels of expertise

Four levels have been identified, each spaced at 25% intervals, accounting for level one, two, three, and four. These levels are arranged from minimum to maximum levels of heuristic skill expertise. The HEQS% of an individual falls into one of these four levels, which determines the level of expertise. For example, if the HEQS% of an evaluator is 51%, the evaluator belongs to Level 3. Though the levels are equally spaced in this study, future research can refine this further by exploring a nonlinear distribution.

Case Studies

Two case studies (three months apart) were conducted to validate the above method. For the first study, an evaluation of a website (Application 1), 18 volunteers participated. The second study was conducted on a web application (Application 2) and the same 18 volunteers participated.

All evaluators were from Infosys Technologies, Ltd. All had conducted heuristic evaluations previously, but had varying levels of expertise as determined in a previous assessment using the HEQS methodology. Forty-five percent belonged to Level 1, and 55% belonged to Level 2. All evaluators were familiar with the eight UI parameters and the three severity ratings.

The case studies were chosen because they were easily accessible and they did not require a knowledge transfer because all evaluators were familiar with them. The evaluators were given complete access to both the application and the screenshots of the key screens. The scope of evaluation for the first application (corporate website) was to evaluate the home page and the landing page of each of the items in the main menu. This corporate website had sections common to corporate websites, such as the 'About Us' and 'Services' section. Twenty-one screenshots were given to the evaluators.

The scope of the second application was to evaluate only the map section from the many applications that the company offered. The map section of the application helps users to find the route between two or more locations. Twenty-six screenshots were given to the evaluators. Each case study was evaluated in two hours in the standardized format based on the eight UI parameters and the three severity ratings defined previously.

The inter-rater reliability is shown in Table 7 for both the applications. Issue consensus refers to the percentage of times all three heuristic experts agree on whether the evaluators comments were issues. Severity Consensus refers to the percentage of times all three heuristic experts assigned the same severity rating to the issue. If there was a disagreement, the heuristic experts would mutually discuss and agree for both the issues and the severity ratings. Disagreements centered mostly between Showstoppers and Major Issues, accounting for 23% and 19% of the issues for Application 1 and 2, respectively. This result is discussed later in this paper.

Table 7: Inter-rater reliability of issue and severity consensus

Inter-rater Reliability	Application 1	Application 2
Issue Consensus	90%	93%
Severity Consensus	70%	74%

Results and Discussion

Case Study 1

The benchmark for Application 1 is shown in Table 8.

Table 8: Benchmark of Application 1 and 2

Benchmark	Application 1	Application 2
Showstoppers	5	23
Major Issues	72	60
Irritants	54	25
Benchmark HEQS	**295**	**320**
Interaction Design	31	38
Visual Design	34	23
Information Architecture	9	4
Functionality	6	17
Labeling	16	10
Content	12	7
Navigation	16	5
Other	7	4

Overall results indicated that the group found an average HEQS% of 24%, and the highest performer had an HEQS% of 38%. The strengths of the study group were visual design and information architecture, as they found an average of 24% and 20% of the issues of the their respective UI parameters of the benchmark. Functionality was a weakness, accounting for only 8 % of
the Functionality issues of the benchmark. Sixty-seven percent of the group belonged to Level 1 expertise, and 33% belonged to Level 2 expertise.

Case Study 2

The benchmark for Application 2 is also shown in Table 8. The group found an average HEQS% of 25%, and the highest performer had found an HEQS% of 38%. The strengths of the study group were content and navigation, amounting to 27% and 20% of the issues of their respective UI parameters of the benchmark. Their weakness was functionality, totaling 7% of the Functionality issues of the benchmark. Fifty percent of the group had Level 1 expertise, and 50% of the group had Level 2 expertise.

Comparing heuristic evaluation skills across applications

Based on the eight evaluators from the two case studies above, 50% did not change their level, 28% improved from Level 1 to 2, and 22% moved from Level 2 to Level 1. Also, for example, we can compare if Evaluator 5 is better than the others at evaluating Application 1. Evaluator 5 can also be compared to the evaluators who evaluated Application 2 (Figure 1). Some practical benefits for managers would include identifying an evaluator for a specific skill. For example, Evaluator 5 is skilled at identifying interaction design issues (Figure 2 and Figure 3). When overall heuristic evaluation skills are considered, one would choose evaluator 8 (Figure 1). Similarly, the training programs can be targeted based on the evaluator’s weakness. For example, evaluator 11 (Figure 2) has not identified any Information Architecture issues, and hence, can be trained in that area.

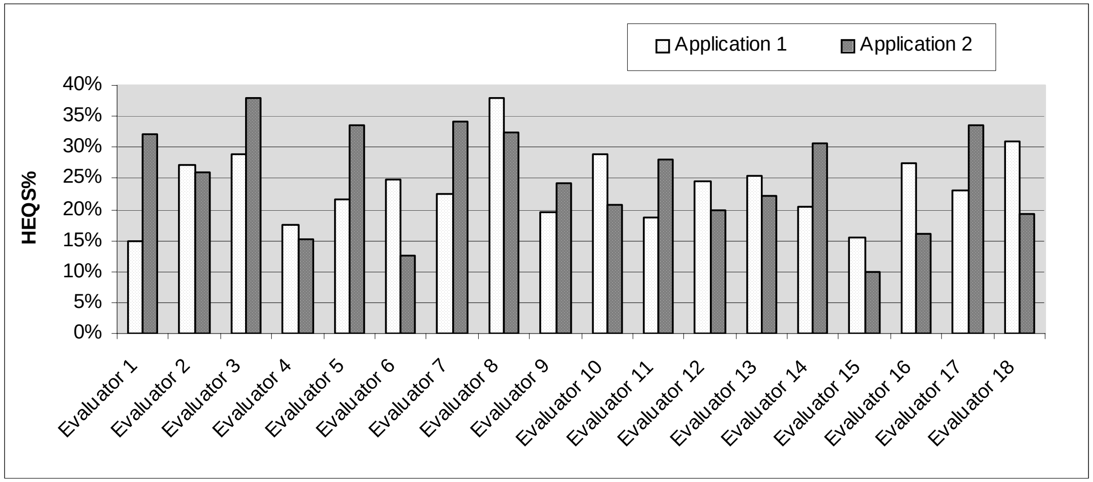

Figure 1: Comparing HE skills across applications

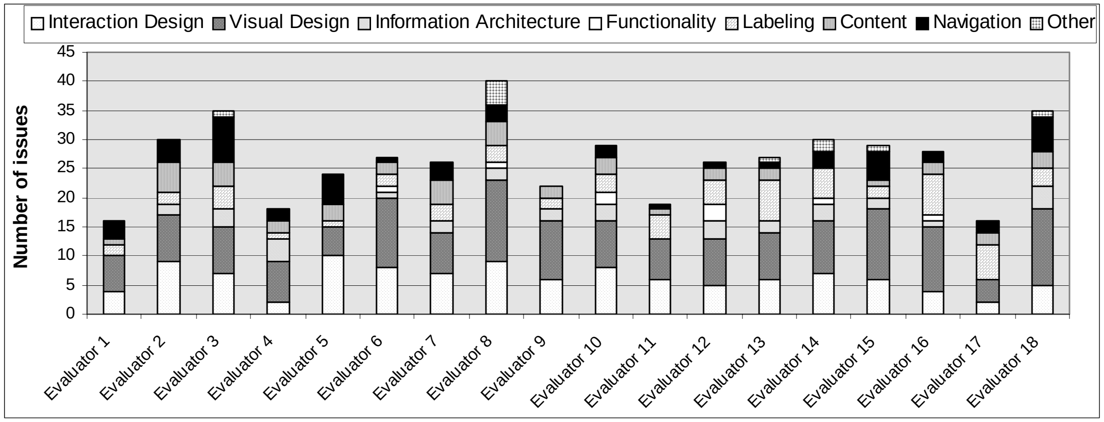

Figure 2: HE Skills based on UI parameters for Application 1

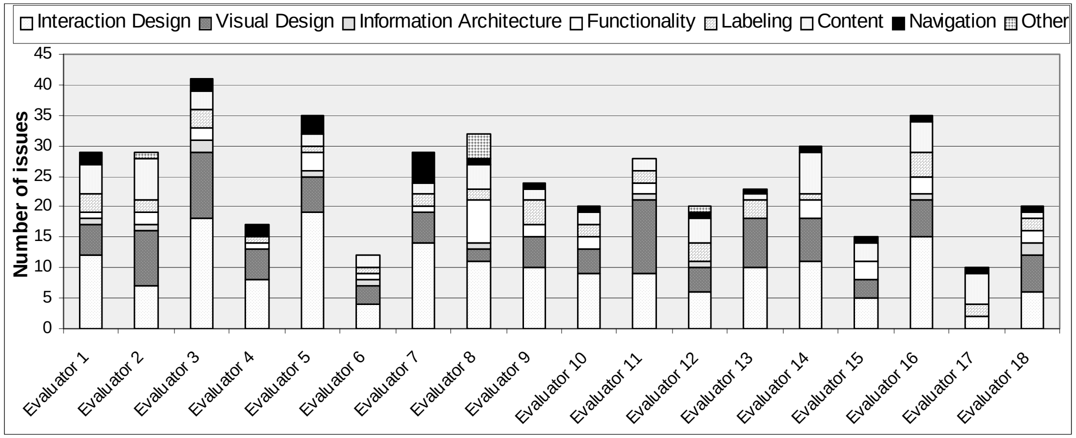

Figure 3: HE Skills based on UI parameters for Application 2

Managers can also identify which method of heuristic evaluation is better using the HEQS assessment. For example, two methods of heuristic evaluation, traditional HE and HE Plus (Chattratichart, and Brodie, 2002), can be compared, and the HEQS scores will identify the better method. A comparison of the two was done by Lindgaard et al. (2004), but they did not take into account the severity of the issues, which can be done using the HEQS methodology. Managers could also identify the quality of their evaluators at an international level. For example, the average HEQS% is 25% for the heuristic evaluators at a particular company, and 35% for the best evaluators in the industry.

Improving the Inter-rater reliability for Severity Consensus

As seen above, previous studies have shown that experts disagree when categorizing issues into severity categories (Hertzum et al., 2002; Jacobsen et al., 1998; Molich, 2006). An article by Bailey (2005) suggests that issues based on severity should be addressed by developers and not heuristic experts. The strengths of heuristic experts are the different perspectives they bring to the evaluation. These perspectives are also the limitation of this assessment. Heuristic experts' perspectives with respect to the evaluation context are not similar.

An issue that wastes time, requires a lot of learning or increases error rates is a Major Issue, and based on the context, the same issue can fall into different severity categories. For example, heuristic experts generally categorize the issue of advertisements having greater importance than the goals of the site, into a Showstopper. However when the stakeholders are taken into consideration, advertisements become a requirement that could be placed better, thus categorizing it as a Major Issue.

In another example, the issue concerning the abbreviations of medicine names on the site was categorized as a Major Issue, but pharmacists preferred the abbreviations as they were faster to understand.

Some suggestions for improving the inter-rater reliability for severity consensus include the following:

- Address severity ratings in a multi-disciplinary team that includes end users and stakeholders.
- Watch for trends and provide evaluators with a checklist under each severity that includes exceptions to the rule.
- Refine the ratings definitions over time, by observing categorization patterns.

Applications and Future Work

An important application of the HEQS methodology is to identify the overall skill level of evaluators and to identify their strengths and weaknesses. The data from this study can be used for targeted training of evaluators leading to a certification program eventually. One training exercise carried out showed an average skill improvement of 48.5% of the HEQS% (Figure 4).

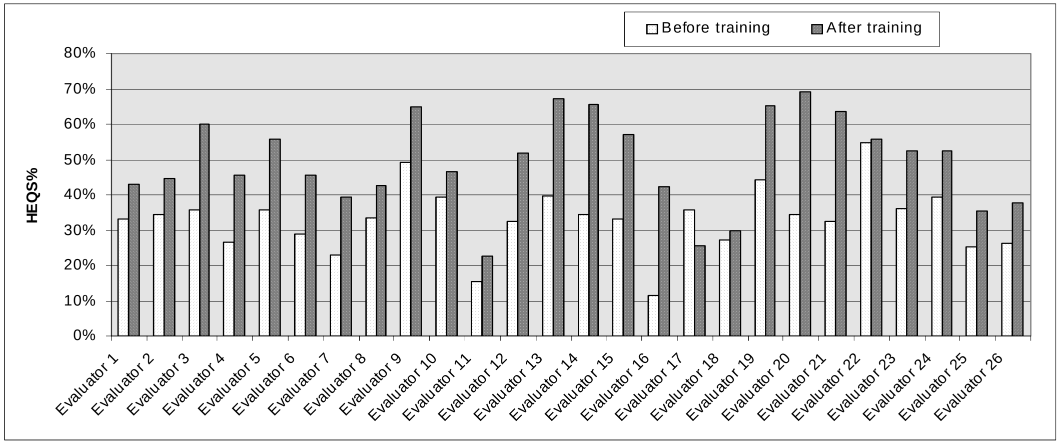

Figure 4: Comparing skills before and after training

Twenty six evaluators were asked to evaluate a website. The website provided information such as, seat reservation status on trains. Twenty-two screen shots and eight scenarios were given to evaluators to evaluate in two hours using the same eight UI parameters and three severity ratings. The evaluators' skills were assessed, and the benchmark found 10 Showstoppers, 49 Major Issues, and 24 Irritants, for a Benchmark HEQS of 221. The average HEQS for the group was 73, and the HEQS% was 33%. After this pre-training assessment, a training program followed. During training, the evaluators discussed the list of issues for the website and their corresponding severity ratings for each UI parameter. Common issues for Navigation, Visual Design, and Labeling were also discussed. (Lists for the other categories are being made.) This served as a reference for the next evaluation.

The common issues were gathered after summarizing more than 200 evaluations that were conducted in the company. Individual feedback was also provided.

The second evaluation was on the same website, but for a different section that handled general information on the different types of trains, railway departments, an so on. Twenty screenshots were given to the same 26 evaluators to evaluate. The evaluators had two hours using the eight UI parameters and the three severity ratings. The benchmark of this post training exercise contained 12 Showstoppers, 55 Major Issues, and 26 Irritants, for a Benchmark HEQS of 251. The average HEQS for the group was 124, and the HEQS% was 49%, resulting in an average improvement of 48.4%.

Conclusion

Further research in sharpening the severity rating by observing categorization patterns can help refine the HEQS model, and thus, define the profile of a heuristic expert more precisely. Currently, the authors are working on applying the experiences from this HEQS methodology to training.

Practitioner's Takeaways

- This paper offers a methodology to assess Heuristic Evaluation skills. Practitioners can use this methodology to identify an evaluator considering the context of the evaluation. For example, in an evaluation for an information providing website, the practitioner can choose an evaluator with required content skills.
- Assessment of skills using the HEQS (Heuristic Evaluation Quality Score) methodology can be customized as per the importance of UI parameters in an organization. For example, some organizations would consider visual design and information architecture as important skills pertaining to their environment, so they would tailor the methodology to suit their needs.
- Training programs can be targeted based on the evaluator's weakness, identified using the HEQS method. These can eventually lead to a certification program.

References

[1] Bailey, B. (2005). Judging the Severity of Usability Issues on Web Sites: This Doesn't Work. In http://www.usability.gov/pubs/102005news.html.

[2] Chattratichart, J. and Brodie, J. (2002). Extending the heuristic evaluation method through contextualization. Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society, HFES, September, Baltimore, pp. 641-645.

[3] Jacobsen, N. E., Hertzum, M., and John, B. E. (1998). The evaluator effect in usability studies: problem detection and severity judgments. In Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (Chicago, October 5-9, 1998), pp. 1336-1340.

[4] Lindgaard, G., Chatrattichart, J., Rauch, T., & Brodie, J. (2004). Toward increasing the reliability of expert reviews. Proceedings UPA 2004.

[5] Desurvire, H. (1994). Faster, Cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing? In J. Nielsen and R. Mack (Ed.) Usability Inspection Methods. New York: Wiley & Sons, Inc, pp. 173-199.

[6] Hertzum, M., Jacobsen, N. E., and Molich, R. (2002). Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations. CHI2002 Extended Abstracts, pp 662-663.

[7] Kanter, L., and Rosenbaum, S. (1997). Usability Studies of WWW Sites: Heuristic Evaluation vs. Laboratory Testing SIGDOC 97.

[8] Karoulis, A., and Pombortsis, A. (2001). Heuristically Evaluating Web-Sites with Novice and Expert Evaluators. Workshop on HCI. 8th Panhellenic Conference on Informatics, 8-10 Nov 2001, Nicosia, Cyprus.

[9] Molich, R. (2006). Comparative Usability Evaluation –CUE. In http://www.dialogdesign.dk/cue.html.

[10] Nielsen, J., and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. Proceedings ACM/IFIP INTERCHI'93 Conference. Amsterdam, The Netherlands, April 24-29, pp 206-213.

[11] Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. Proceedings of ACM CHI'94 Conference, Boston, MA, April 24-28, pp 152-158.

[12] Nielsen, J. 1994. How to Conduct a Heuristic Evaluation. In http://www.useit.com/papers/heuristic/heurstic_evaluation.html.

[13] UPA. (2005). UPA 2005 Member and Salary Survey. Usability Professionals Association. Bloomingdale, IL.

Acknowledgements

We would like to thank all the evaluators who participated in this study. Their enthusiasm is greatly appreciated.

Shazeeye Kirmani is a Senior Usability Engineer at Infosys Technologies, Limited. She received an M.S. in Human Factors from the State University of New York at Buffalo in 2004. She has worked extensively in the area of user testing and heuristic evaluation. She has anchored more than 100 heuristic evaluations in domains ranging from banking to retail from July 2005 to March 2006.

Shanmugam Rajasekaran is Head, Experience Design at Infosys Technologies, Limited. In the last financial year, the User Experience team delivered more than 250 projects for 90 Infosys customers across several domains. He is also actively involved in establishing quantitative processes in usability. He graduated in Product Design from National Institute of Design, Ahmedabad, India, in 1992.