



Heuristic Evaluation Quality Score (HEQS): Defining Heuristic Expertise

Shazeeye Kirmani

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

Abstract

This paper identifies the factors affecting heuristic expertise and defines levels of expertise permissible to conduct an evaluation. It aims to standardize skills or define heuristic expertise worldwide and also suggests ways to improve issue categorization.

An online heuristic evaluation competition was hosted on the World Usability Day website in November 2007 by Usability Professionals' Association (UPA), Bangalore. Twenty contestants from the U.S. and India with heuristic evaluation experience ranging from 0 to 10 years participated. Contestants were judged on a quantitative framework that assigns weights to the severity of the identified issues (Kirmani & Rajasekaran, 2007). Results indicated that the group with average heuristic experience of 2 years found a mean HEQS% of 8% in 1 hour. Previous studies identified that evaluators found a mean of 34% of issues but did not address issue quality (Nielsen & Landauer, 1993). Further studies on heuristic expertise quality would make the standards more reliable.

Keywords

Heuristic Evaluation Quality Score (HEQS), Heuristic Evaluation Quality Score Percentage (HEQS%), heuristic expertise, Interaction Design, Information Architecture, Visual Design, Navigation, Labeling, Content, Functionality, Showstopper, Major Issue, Irritant, Domain Experience, Usability Experience, Heuristic Experience, User Interface (UI) Parameter, Severity

Introduction

Heuristic evaluation is a discount usability engineering method involving a few evaluators who judge the compliance of an interface based on a set of heuristics. It is difficult for one evaluator to find all the usability problems with an interface hence a few evaluators, preferably between three to five evaluators, are suggested. This optimal range gives the best benefit-to-cost ratio (Nielsen & Landauer, 1993). 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 popular technique that is used by 76% of the usability community (UPA Survey, 2005) and has a high a cost-to-benefit ratio of 1:48 (Nielsen, 1994) emphasizes the assessment of heuristic evaluation skills. More so, evaluators are extremely confident of their abilities as experts. An online heuristic evaluation competition held in November 2007 by UPA, Bangalore, India asked contestants to rate themselves on a scale of 5, where 5 meant that they absolutely thought they would win the competition. Results showed that 85% of 20 contestants felt confident of winning the competition. They scored 4 or more on a scale of 5. This confidence or the inability to authenticate it threatens the quality of heuristic evaluation. Experts could misuse this false confidence, intentionally or unintentionally, to provide expertise of sub optimal standards limiting the usability of the applications that they evaluate. This issue can pose grave risk to the users of these applications who depend on these applications, in some cases, to save their lives. Hence, it is critical to quantify this expertise to ensure evaluations of a certain standard.

A framework to quantify heuristic evaluation skills was proposed by the authors (Kirmani & Rajasekaran, 2007). Quantification is based on the number of unique, valid issues identified by the evaluators as well as the severity of each issue. Unique in this context refers to a problem that could be repeated in more than one place but still counted as a single, unique problem with several instances. Unique, valid issues are categorized into eight user interface parameters and severity is categorized into three. The three categories of severity are showstoppers or catastrophic issues preventing users from accomplishing goals, major issues or issues causing users to waste time or increase learning significantly, and irritants or cosmetic issues violating minor usability guidelines. Weights of 5, 3, and 1 are assigned to showstoppers, major issues, and irritants respectively. A Heuristic Evaluation Quality Score (HEQS) is computed for each evaluator by multiplying the weight factor with the number of issues in that severity category. For example, Evaluator A has identified 2 showstoppers, 10 major issues, and 20 irritants his $HEQS = 2*5 + 10*3 + 20*1 = 60$. A benchmark of the collated evaluations of all the evaluators is used to compare skills across applications as well as within applications. If the benchmark HEQS is 200 then Evaluator A identified an HEQS% of $60/200$ or 30%. Skills are also computed for eight User Interface (UI) parameters to identify strengths and weaknesses of the evaluators. The eight parameters are Interaction Design, Information Architecture, Visual Design, Navigation, Labeling, Content, Functionality, and Other (for issues that do not fall into the first seven).

This metric has been used to compare the heuristic expertise of individual evaluators with other evaluators across or within applications to base evaluations on individual strengths. It has also been used to identify weaknesses and train evaluators in those skills. Measuring improvement based on training and tailoring training programs to groups or individuals based on this methodology are some other applications.

What has not been addressed in the previous study is a definition of heuristic expertise at a global level. This study aims to define these standards for the world wide usability community. It is also known that many such competitions will need to be conducted before these results can be generalized and this is a first attempt to do so.

In particular the following questions are addressed:

- What are the factors affecting heuristic evaluation expertise?
- What is the average expertise of heuristic evaluators?
- What level of expertise is required for one to conduct a heuristic evaluation?

Method

A world wide heuristic evaluation competition that was part of Usability Professionals' Association's (UPA) World Usability Day was hosted by UPA, Bangalore online in November 2007. The competition details were as follows:


- Nature of the website: A healthcare site where evaluators put in symptoms and the website provides advice.
- Scope of the evaluation: Five scenarios were given to evaluators.
 - Find all conditions related to a cough.
 - Check all symptoms associated with a cough.
 - Edit the symptoms.
 - Find related articles.
 - E-mail an article to yourself.
- Time to do the evaluation: 1 hour
- Demographic data collected: age, gender, experience, location, and confidence to win

Contestants had to sign an honor statement stating that they would not take more than an hour to complete the evaluation and that the evaluation was the sole effort of the contestant. The competition lasted for 2 weeks and entries were submitted in a particular format (see Table 1). The UI parameters and severity categories were taken from the initial HEQS paper (Kirmani & Rajasekaran, 2007).

Table 1. The Evaluation Format

Issue	The other 2 steps in the process of getting advice for a particular symptom are barely visible.
UI Parameter	Visual Design
Severity	Major Issue

Contestants were encouraged to participate via various methods such as sending links to the competition via individual e-mails, blogs, social networking sites, and usability communities. Anyone could participate. They were all directed to the World Usability Day website (see Figure 1) where the event was hosted online. The event page had a downloadable presentation with the demographic data collection form, evaluation format, scope of the evaluation, and judgment criteria. Each contestant used their own evaluation criteria to conduct the 1 hour evaluation. They had to e-mail their entries along with the filled demographic form to a given e-mail address.



World Usability Day
Making life easy!

Home About Tools Community Sponsors Press Contact Event Highlights **2007 Events**

World Usability Day 2007 Events

View by map View by country View by hour Webcasts & Interactive

Note: This event took place in the past.

Competition: Are you the world's best Expert Reviewer?

See more detail about this event.

Event Details

What is this competition about?
On the occasion of World Usability Day (8th November, 2007), Usability Professional's Association (UPA), Bangalore, is hosting a worldwide expert review competition.
Are you the world's best Expert Reviewer? Find out by participating in this competition!

Expert reviewers (or heuristic evaluators) claim to be 'experts' without quantitative proof.
A heuristic skill assessment called Heuristic Evaluation Quality Score (HEQS) was introduced earlier this year and published in the Journal of Usability Studies (Feb 2007, Volume 2, Issue 2).
All participants will be judged quantitatively using the criteria specified under HEQS.
Apart from identifying the world's best expert reviewer, data from this competition will also be used to arrive at a list of global review benchmarks, which will be published separately. This data will help organizations benchmark their reviewers at a global level.

Who can participate?
This competition is open to anybody who thinks they can identify issues that hinder a website's ease-of-use.
You may be located anywhere in the world. You just have to be passionate about making technology easy to use. You don't have to necessarily be a professional expert reviewer or heuristic evaluator to participate.

Competition Details
When: Thursday, 8th November, 2007 – World usability Day. Your completed entries have to reach us before the 8th of November ends in your local time zone.
Time to complete the evaluation: 1 hour

Format and Instructions:
On the 8th of November, you will evaluate one section of the WebMD website. You will find usability issues or issues that hinder the ease-of-use for a set of scenarios. Further instructions are located at the [UPA Bangalore website](#).
Submit your entries to heqs.wud2007@gmail.com.

Judgment Criteria

- All entries need to follow the given format, else they will be disqualified.
- Points will be awarded based on the process identified by the [HEQS paper](#) that was published in February 2007 in the Journal of Usability Studies.

If you win
Only a single winner will be announced.
If you win, you will receive a gift coupon of US \$ 400.

Any questions?
Contact UPA Bangalore at heqs.wud2007@gmail.com

Figure 1. Competition details and instructions hosted on the World Usability Day website

The competition was judged the same way as described in the previous HEQS paper (Kirmani & Rajasekaran, 2007). The judges were given a benchmark of a collation of all the issues (200 plus issues) of all the contestants. The three judges who were heuristic experts followed the process listed below.

1. Each judge rated each issue of the 200 plus issues as valid or invalid individually.
2. Each judge categorized the severity and UI parameter of each valid issue individually.
3. They got together to discuss each issue and its severity and UI parameter categorization.
4. If they did not agree on the issue validity, severity, or UI parameter categorization they discussed it together to arrive at a final consensus.
5. This finalized list of valid issues with their appropriate severity and categorization was used to judge each contestant.
6. Judges were requested to write down their thoughts on the categorization process to later discuss ways to improve it.
7. Each issue of a contestant's entry was matched to the finalized list and weights of 5, 3, or 1 were awarded based on the severity of the issue (5 for every showstopper, 3 for every major issue, and 1 for every irritant). Summing up the scores for all the issues one arrived at the HEQS for each contestant.
8. If contestants incorrectly categorized issues for severity or UI parameter they were re-categorized to arrive at their scores. Contestants were asked to enter the severity and UI parameter to gain insight into the categorization process. For example, the way

the issue was worded depended to a great extent to the way it was categorized. This qualitative data helped to improve the categorization process.

Demographic Data

Twenty contestants took part in the competition. Table 2 summarizes the demographic data.

Table 2. Demographic Data of the Contestants

Parameter	Average	Range
Age	28.4 years	22-34 years
Time spent as heuristic evaluator	23.7 months	0-120 months
Time spent as a usability practitioner	30.7 months	0-144 months
Time spent as a domain expert (i.e. healthcare)	4.2 months	0-24 months
Confident of winning (self rating on a scale of 5 where 5=absolutely win and 1=never win)	4.1	1-5
Gender	6 females and 14 males	
Location	6 states and 2 continents Karnataka, India Mumbai, India Gujarat, India Chennai, India New Jersey, US California, US	

Results

Overall results indicated that the group found an average HEQS% of 8% in 1 hour ranging from 2% to 17% (see Figure 4). This compares with previous studies of evaluators finding 24% and 25% (highest in both cases being 38%) in 2 hours (Kirmani & Rajasekaran, 2007). The numbers are slightly higher for the previous case studies as the average heuristic experience was higher (more than 30 months) than the average heuristic experience of the contestant group (23.7 months). The contestant group also included 4 contestants who have never been exposed to heuristic evaluations. Evaluators can be compared and their performance can be studied by looking at the issues identified based on UI parameter (see Figure 2) and severity (see Figure 3). For example, Evaluator 3 found twice as many interaction design issues, thrice as many content issues, and five times as many navigation issues as Evaluator 18. However, Evaluator 18 found five times as many showstoppers as Evaluator 3 indicating that Evaluator 3 is good at covering the breadth of issues (across UI parameters) while Evaluator 18 is good at covering severe issues.

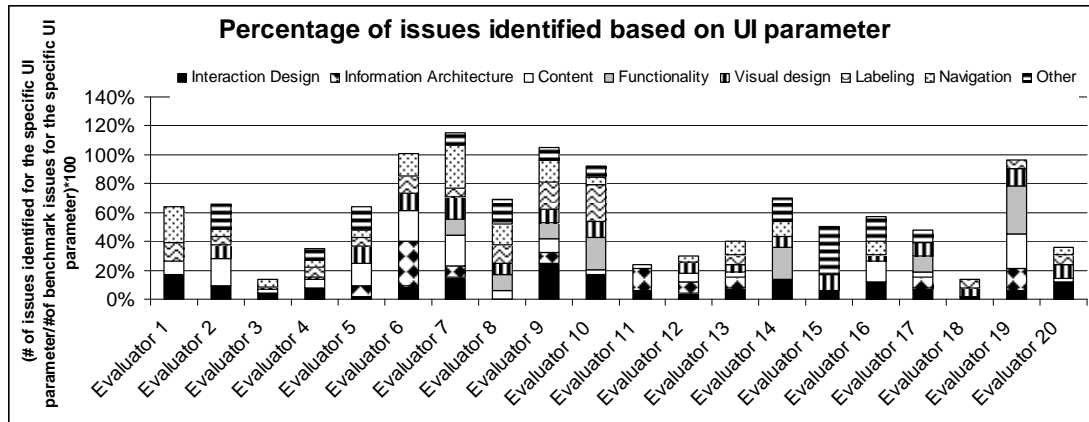


Figure 2. HE skills based on UI parameters

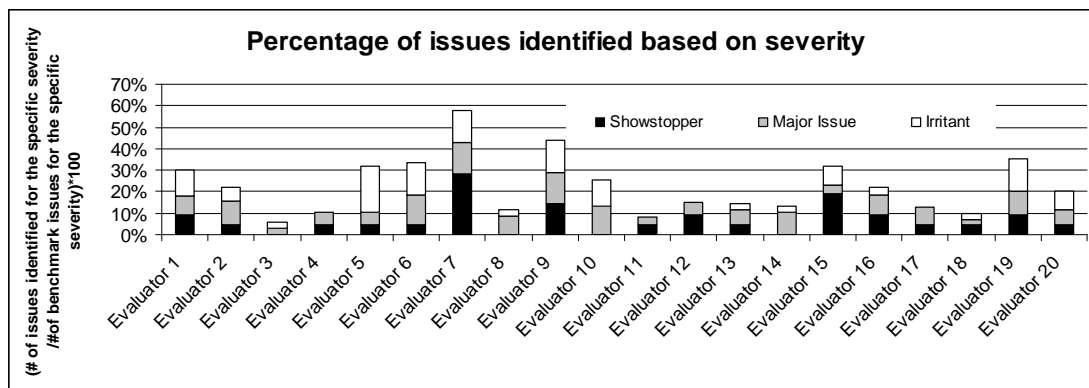


Figure 3. HE skills based on severity

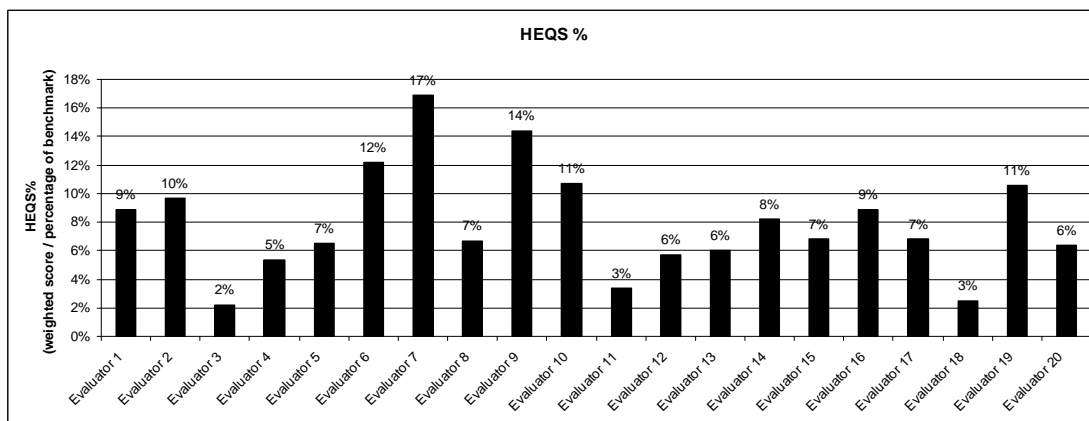


Figure 4. HEQS%

What is the average expertise of heuristic evaluators?

The average HEQS% is 8% for evaluations of 1 hour conducted by a group of evaluators with an average heuristic experience of 2 years and an average usability experience of 2.5 years.

What are the factors affecting heuristic evaluation expertise?

The following factors affect heuristic evaluation expertise:

- Usability experience: The relationship between usability experience and heuristic evaluation expertise is significant (see Table 3). Thirty percent of the variation between usability experience and heuristic evaluation expertise is related. The more usability experience the better is the quality of the evaluation.
- Heuristic experience: The relationship between heuristic experience and heuristic evaluation expertise is significant. Seventeen percent of the variation between heuristic experience and heuristic evaluation expertise is related. The more heuristic experience the better is the quality of the evaluation.
- Domain experience: Domain experience in this study did not significantly impact expertise. This could be due to the non-technicality of the website. Identifying conditions for a set of symptoms is understood world wide and does not require a lot of learning but other studies have shown that domain experts are better evaluators (Anthanasios & Andreas, 2001).
- Training: Training does impact heuristic evaluation expertise. Quality of the evaluation improves with training. A 48.4% improvement was seen in a study conducted on a group of 26 evaluators (Kirmani & Rajasekaran, 2007).

The following factors do not affect heuristic evaluation expertise:

- Age: Age does not affect heuristic evaluation expertise.
- Gender: Gender does not affect heuristic evaluation expertise.
- Self rating: Self rating or self proclamation of calling oneself an expert does not corroborate with heuristic expertise. Eighty-five percent of 20 contestants felt confident of winning the competition and rated themselves 4 or higher on a scale of 5.

This study did not shed light on site complexity or previous experience on the site. Future research should look at more complex examples and prior experience with the site.

Table 3. Correlation Analysis of Demographic Data

Parameter	Range	Average	Median	Significant/Not (at significance level of 0.1)
Gender	Female, Male	--	--	
Age	20 - 34 years	28.4 years	29 years	
Usability Experience	0 - 144 months	30.7 months	15 months	Significant
Heuristic Experience	0 - 120 months	23.7 months	13 months	Significant
Domain Experience	0 - 24 months	4.2 months	0	
Self rating and Confidence to win	1 - 5 (5 being "I will absolutely win")	4.1	4	

What level of expertise is required for one to conduct a heuristic evaluation?

Expertise can be divided into three levels:

- Below average evaluators: Evaluators finding an HEQS% of less than 8% are below average performers.

- Above average evaluators: Evaluators finding an HEQS% of 8% or more are above average performers.
- Exceptional evaluators: Evaluators finding an HEQS% of 15% or more are exceptional performers. Fifteen percent has been arrived at by selecting the top 5%, given the highest performers have been identifying HEQS% of 17% - 19% of issues in 1 hour.

It is known that any evaluator who identifies issues to improve the usability of an application is better than none, but I recommend that you choose 3-5 above average or exceptional evaluators to see an evaluation of high quality.

Improving severity and UI parameter categorization

From the inter-rater reliability in Table 4 we see that evaluators can categorize showstoppers consistently but are not consistently categorizing major issues and irritants.

Table 4. Inter-rater Reliability for Issues Based on Severity

Benchmark	Showstopper	Major Issue	Irritant	Non-issue
Number of unique issues	21	153	33	49
Complete consensus	91%	66%	55%	86%

Currently descriptions of severity (Nielsen, 1994) are seen in Table 5.

Table 5. Current Descriptions of Severity

Severity	Description
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.

After judging the competition notes on categorization were compared and we arrived at the following grid to improve the categorization by adding two dimensions: the user and the environment (see Table 6).

Table 6. Revised Severity Ratings

Severity	About the issue	Different users	Different environments	Yes/No*
Showstopper	Does the issue stop you from completing the task?	Can colorblind users interpret a colorful graph to complete a task?	Does the issue create an unstable environment?	
Showstopper Example	The "Submit" button is not working and hinders users from sending their forms.	If colors are the only form of communicating critical data to complete an online transaction, colorblind users are forced to abandon the task.	For a healthcare site it is critical that advice given pertains to the conditions chosen. Incorrect association can cause harm.	

Severity	About the issue	Different users	Different environments	Yes/No*
Major Issue	Does the issue cause you a major waste of time, increase your learning, increase the error rate, or violate a major consistency guideline?	Does the issue increase errors for older adults? Does the issue increase learning for all users?	Does the issue create an environment with a higher error or learning rates?	
Major issue Example	Using an "X" as an icon to zoom out breaks user mental models and increases errors considerably, especially in an environment where close is also denoted by an "X".	A low contrast between font and the background can cause an increase in error rates for older adults.	Providing smaller than usual buttons on a mobile interface where people are always moving can increase error rates considerably.	
Irritant	Does the issue involve a cosmetic error, slow you down slightly, or violate a minor consistency guideline?	Does the site not have visual appeal to teenagers?	Does the issue create an environment that slows you down slightly?	
Irritant Example	The label is "Symptom" when it actually should be plural as it has many symptoms.	If the colors are not young and vibrant (e.g., pink and yellow) for a site catering to teenagers it violates a cosmetic error.	If you are checking symptoms for your daughter, changing the content to cater to a different environment (third person) is helpful.	

*Answering positively to one or more questions is Yes.

From the inter-rater reliability in Table 7 we see that evaluators are not consistently categorizing information architecture issues.

Table 7. Inter-rater Reliability for Issues Based on UI Parameters

Benchmark	Information Architecture	Navigation	Labeling	Other	Visual Design	Interaction Design	Content	Functionality
Number of unique issues	13	20	16	12	52	52	33	9
Complete consensus	70%	85%	100%	92%	80%	92%	88%	78%

This could be due to the poor labeling of the group as Information Architecture in usability circles denotes structure and organization, navigation, and labeling. Hence, we have decided to re-label it as Structure and Organization (see Table 8).

Table 8. Revised UI Parameter

Current UI Parameter	Current Description	Redefined UI Parameter	New Description
Information Architecture	Accurate structuring of information into groups best matching the mental model of users.	Structure and Organization	Accurate structuring of information into groups best matching the mental model of users.

Limitations of this study

It is known that this study and its results are limited to the small sample size that has been used. Generalizing these results will require many more competitions with a diverse and larger sample size.

Conclusion

Further research with a more diverse and larger group would help in sharpening the reliability of these results. Tighter controls on variables affecting heuristic expertise to measure definite impact on heuristic expertise will also increase reliability.

Practitioner's Take Away

- A Heuristic Evaluation Quality Score (HEQS) can be used to quantify heuristic evaluation expertise to ensure evaluations of a certain standard. It is critical for evaluations to be of a certain standard to build reliability and trust among evaluators and ultimately provide the end users with high quality applications.
- The average HEQS% is 8% for evaluators taking 1 hour with heuristic experience of 2 years.
- Evaluations are recommended to be done by above average and exceptional evaluators. Above average evaluators have an HEQS% of 8% or more and exceptional evaluators have an HEQS% of 15% or more.

Acknowledgements

A special thanks to Intuit and Infosys Technologies Ltd. for sponsoring this competition. I would also like to thank Shanmugam Rajasekaran, Deepa Bachu, Amit Pande, Muthukumar, Anand Almal, Rajavel Manoharan, and Amit Bhatia for helping make the competition a success. Last but not least, I would like to thank the contestants who set aside valuable time and participated in the competition.

References

- Andreas, P. & Athanasis, K. (2001, November). Heuristically Evaluating Web-Sites with Novice and Expert Evaluators. Workshop on HCI. *8th Panhellenic Conference on Informatics*, Nicosia, Cyprus.
- Bailey, B. (2005, October). Judging the Severity of Usability Issues on Web Sites: This Doesn't Work. Retrieved March 2008, from <http://www.usability.gov/pubs/102005news.html>
- Chattratchart, J. & Brodie, J. (2002, September). Extending the heuristic evaluation method through contextualization. *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society, HFES* (pp. 641-645) Baltimore, Maryland.

- Jacobsen, N. E., Hertzum, M., & John, B. E. (1998, October). The evaluator effect in usability studies: problem detection and severity judgments. *In Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1336-1340) Chicago, Illinois.
- Kirmani, S. & Rajasekaran, S. (2007). Heuristic Evaluation Quality Score (HEQS): a measure of heuristic evaluation skills, *Journal of Usability Studies*, 2 (2), pp 61-75.
- Lindgaard, G., Chatrattichart, J., Rauch, T., & Brodie, J. (2004). Toward increasing the reliability of expert reviews. *Proceedings UPA*.
- Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations. *CHI2002 Extended Abstracts* (pp 662-663).
- Molich, R. (2006). Comparative Usability Evaluation–CUE. Retrieved March 2008, from <http://www.dialogdesign.dk/cue.html>
- Nielsen, J., & Landauer, T. K. (1993, April). A mathematical model of the finding of usability problems. *Proceedings ACM/IFIP INTERCHI'93 Conference* (pp. 206-213) Amsterdam, The Netherlands.
- Nielsen, J. (1992, May). Reliability of severity estimates for usability problems found by heuristic evaluation. *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems*. Monterey, California.
- Nielsen, J. (1994, April). Enhancing the explanatory power of usability heuristics, *Proceedings of ACM CHI'94 Conference* (pp 152-158) Boston, Massachusetts.
- Nielsen, J. (1994). How to Conduct a Heuristic Evaluation. Retrieved March 2008, from http://www.useit.com/papers/heuristic/heuristic_evaluation.html
- UPA. (2005). UPA 2005 Member and Salary Survey. Usability Professionals Association. Bloomingdale, Illinois.

About the Author



Shazeeye Kirmani

Ms. Kirmani is a Senior Usability Engineer at Infosys Technologies, Limited. She received her M.S. in Human Factors from the State University of New York at Buffalo in 2004.