

# Heuristic Evaluations of Bioinformatics Tools: A Development Case

Barbara Mirel and Zach Wright

University of Michigan  
(bmirel, zwright}@umich.edu

**Abstract.** Heuristic evaluations are an efficient low cost method for identifying usability problems in a biomedical research tool. Combining the results of these evaluations with findings from user models based on biomedical scientists' research methods guided and prioritized the design and development process of these tools and resulted in improved usability. Incorporating heuristic evaluations and user models into the larger organizational practice led to increased awareness of usability across disciplines.

**Keywords:** Usability, heuristic evaluation, biomedical research, organizational learning, user models.

## 1 Introduction

Assuring usefulness and usability—a perennial challenge in any software project—is particularly tricky in bioinformatics research and development contexts. Our NIH-funded bioinformatics center produces tools for systems biology analysis. The databases and tools enable biomedical researchers to interactively analyze genomic-level data for the purpose of uncovering systemic functional roles that candidate genes/gene products may play in susceptibility to a disease. Ensuring the usability of these tools is a challenge because we are not a software shop and must optimize the combination of academic and implementation specialties that we have available. The discount usability inspection method of heuristic evaluations is highly attractive.

We recognize that heuristic evaluations (HE) alone—the process of scoring tools for their concordance with usability standard—are insufficient for detecting and generating improvements for significant usability and usefulness advances [8]. Therefore, we integrate heuristic evaluations with three processes known to enhance their effectiveness: (1) Evaluators are familiar with the tools and users' query and analysis tasks; (2) heuristics—i.e., the usability principles against which tools are judged—are adapted to the domain and tasks specific to the tools, and (3) heuristics and interpretations of findings are informed by user models of researchers' analytical performances and goal-driven cognition [6].

Additionally, we recognize that assessments of our web-based bioinformatics tools must account for the support of more complex explorations than user interfaces (UI)/web pages originally targeted by usability inspection methods support. Toward this end, we combine heuristic evaluations with research, development, and other organizational processes. This integration facilitates our abilities to distinguish real

problems in the results, set priorities for fixes, and raise developers' awareness of user needs beyond surface fixes to better build for users' cognition in scientific analysis.

Our outcomes have been positive. We argue that for our bioinformatics tools, positive results hinge on combining domain-based, user-informed heuristic evaluations with organizational processes that break down boundaries isolating usability from development, modification request decisions, and UI design.

## 2 Relevant Research

Heuristic evaluations involve “evaluators inspect[ing] a user interface against a guideline, be it composed of usability heuristics or cognitive engineering principles, in order to identify usability problems that violate any items in the guideline”[8]. This method is known to produce many false positives and likely omissions of problems related to users' cognitive tasks. It nonetheless is one of the most popular usability assessment methods due to its low costs and efficiencies [2]. Thus it is important to improve the effectiveness of HEs without diminishing their benefits. Researchers have found several ways to achieve these improvements. They include conducting heuristic evaluations with many evaluators and combining them with evaluator training and reliability testing increase the effectiveness of HEs [10,12]. Heuristic evaluation results also improve when evaluators have prior knowledge of usability and the tool; when heuristics are adapted to domain tasks and knowledge; and when HE findings are compared with results from user performance studies [3]. Finally, improvements come from using sets of heuristics that are “minimal” (not overlapping) yet inclusive [10]. For example, some researchers have evaluators jointly consider heuristics and problem areas, thereby assessing to a “usability problem profile” [2]. Establishing an optimal set of heuristics, however, is still a black box.

To compensate for elusive “ideal heuristics,” many usability researchers advocate integrating findings from user performance studies with HE. Demonstrably, heuristic and user performance evaluations combined uncover more problems than either method does alone. Yet quality not just quantity of problems is critical. For better quality, some researchers claim that what is missing in Nielsen's standard set of heuristics is that they are not “related to cognitive models of users when they interact with system interfaces” [8]. Cognitively-oriented heuristics are especially important when tools support complex tasks. Recent attempts to construct heuristics that address cognition include Gerhart-Powel's [5] cognitive engineering principles and Frokjaer and Hornbaek's [3] principles based on metaphors of thinking. So far findings about the superiority of such heuristics have been mixed [4,8].

Running in parallel with these academic efforts, some studies by specialists in production contexts aim to improve the effectiveness of HEs by advantageously combining them with organizational processes. Hollinger [7], for example, reports on positive efforts at Oracle—against great organizational resistance at first—to combine bug reporting processes with HE findings, thereby “mainstreaming” reviews of outcomes. This mainstreaming increased usability awareness across different teams and functional specialties, incited interactive team discussions about usability, initiated tracking the costs and benefits of usability improvements, and resulted in fixing more usability defects. Moreover, results included “significant improvements in the quality of the user interface” [7].

Exploiting organizational processes is promising but, to the best of our knowledge, few production context studies report on combining HE with even more organizational processes than Hollinger [7] describes or on combining organizational processes with the established methods of improving HE outcomes by comparing them with usability performance findings, assuring evaluator familiarity with the tools, and adapting heuristics to the task domain.

### **3 Methods**

Our methods are tied to achieving the same effectiveness with HE that other researchers seek by combining them with other factors. Unfortunately, due to resource constraints we could not conduct extensive training of evaluators or involve numerous evaluators. We could, however, get several evaluators familiar with the tools, adapt and pilot test heuristics to our domain and tools, and introduce several new organizational processes. We also introduced the novel process of reframing surface problems found by HEs into more substantial problems-based on user models.

#### **3.1 Tools**

We report on heuristic evaluations of one open source, web-based query and analysis tool. The tool is the front end for querying our center's innovatively integrated protein interaction database. The query and analysis tool lets users query by gene(s), keyword(s), or gene list and provides tabular query results of relevant genes, attributes, and interactions. The tool is non visual but links to visualization tools.

#### **3.2 User Task Models**

User models were derived from longitudinal field studies of 15 biomedical researchers using our tools and others to conduct their systems biology analysis [9]. These models directed both our adaptations and interpretations of heuristic evaluations. The user models are unique in bioinformatics because they captures scientists' higher order cognitive and analytical flow for research aimed at hypothesizing and not only lower level tasks that are typically studied in usability tests, cognitive walkthroughs, or cognitive task analysis. Specifically, the user models capture moves and strategies for verifying accuracy, relevance, and completeness and uncovering previously unknown relationships of interest. These tasks involve manipulating data to turn it into knowledge through task-specific combinations of sorting, selecting, filtering, drilling down to detail, and navigating through links to external knowledge bases and literature. Additionally, to judge if genes and interactions are interesting and credible, scientists analyze high dimensional relationships and seek contextual cues from which to draw explanatory inferences. Ultimately, they examine conditions and causes in interactive visualizations, tools outside the scope of this article. This empirically-derived model of higher order cognition was critical to adapting standard Nielsen heuristics to our domain and tool.

### 3.3 Adapted Heuristics

We adapted Nielsen's standard set of 10 usability heuristics to our domain and uses of our tools to include the following: The presence of external links to multiple data sources and internal links to details and the large amounts of heterogeneous data in result sets; the core need for statistics and surrogates for confidence; and the variety of interactions needed for validating, sensemaking, and judging results.

### 3.4 Heuristic Evaluations and Evaluators

Three evaluators pilot tested the adapted heuristics with other query and analysis tools developed by our center to refine their applicability to the domain and users tasks. One evaluator is trained in usability and visualizations and the other two evaluators specialize, respectively, in portal architecture and systems engineering and in web administration and marketing communications. All were knowledgeable about the tools and moderately aware of users' tasks and actual practices through discussions with the usability director about field study findings. No reliability testing was done due to time constraints. Instead, inter-evaluator differences were analyzed by examining comments entered in the comments field in the instrument. After heuristic evaluations were conducted, outcomes and comments were summarized and grouped by agreement and severity. Relevant design changes were suggested.

### 3.5 Integration of Additional Processes

Concurrent with the heuristic evaluations, the following organizational and software development life cycle processes were instituted with enhanced usability in mind:

- Usability categories and severity levels were built into in the modification request (MR) system. Levels were: Minor, serious, major, critical, and failure, and they were coordinated with a newly instituted Technical Difficulty ranking.
- Operational processes were put into place for turning MRs into long term development priorities and for raising awareness of user models and their requirements. Our processes included forming a new committee for setting priorities composed of the directors of computer science, life sciences, and usability along with the lead developer and project manager.
- Informal and highly collaborative processes between developers, web designers, usability evaluators, and scientists were implemented to assure rapid prototyping and feedback
- A research project was initiated into design requirements based on heuristic evaluation findings and user models.

## 4 Results

### 4.1 Evaluation Outcomes

Conducting the heuristic evaluations took on average two hours/evaluator. Summarizing added another few hours to the effort. Sample summary outcomes are shown in Table 1. Those with agreed upon high severity are highlighted.

**Table 1.** Sample of results summarized from heuristic evaluations

Heuristic	Problem severity /agreement	Problem(s)	Design change
1. Currency of the tool web pages	High/agreement	No date present	Indicate last update to web pages
2. Readable text	High/agreement	Small font	12 point font
3. Hints for formulating a query for better results	High/agreement	No hints available	Need query hints when the query fails.
4. Able to undo, redo, go back	High/agreement	No history tracking;	Provide history tracking
5. Broken links	High/agreement	“Top of page” is broken	Fix [list of broken links]
6. Examples included and prominently	Range/no agreement (high to low)	Could use more examples and better emphasis	Add 1-2 (bolded) examples under the search box
7. Currency of the data; data sources cited	Range/no agreement (high to low)	Versions of dbs are listed but no dates of latest updates	Add a date for last updating to our database
8. Clearly shows if no results occur	Range/no agreement (high to low)	Shows, but the message isn’t clear	Change message to: [Suggestion]
9. Able to change result organization	Range/no agreement (high to low)	Sort is available but not apparent	Need note that columns are sortable
10. Vital statistics are available.	Range/no agreement (high to low)	What would those stats be?	No agreement
11. Information density is reasonable	Range/no agreement (high to lo)	A lot of whitespace; too many sections	Get rid of the 5 nested boxes;
12. Clear what produced the query results	Range/no agreement (high to low)	Should redisplay search term so user ties it to results	No agreement
13. Clear why results seem high or low	Middle/agreement	No explanations; I assume informed user knows why	Not clear where the search term is “hitting.”
15. Can access necessary data for validating	Low/ agreement	Not sure what the data would be	No agreement

As Table 1 shows, highly ranked problems involved broken and missing features and web page omissions that could be added without programming. Middle-ranked problems were tied more to user task needs and subjective issues such as what constitutes either “enough” support or the criteria scientists use for judging reliability/validity. Problems with little agreement about severity level were tied even more to evaluators having to project and evaluate the importance of scientists’ task needs in this domain. For example, evaluators varied widely in judging the importance of validation in scientists’ ways of reasoning and knowing. Some actual problems were not caught by the heuristic evaluations, especially those involving novel and unexpected ways users might interact with the interface. These findings were provided by the field studies. Additionally, evaluators’ comments and the summarized design changes ranged from precise to vague. Typically, design changes for familiar problems in

generic UI design were precise; those tied to user task models for systems biology and complex exploratory analysis were not.

## 4.2 Integrating Organizational Processes

Interpretations and the actions taken on the outcomes of heuristic evaluations took the following course organizationally. As noted in Methods, design changes were entered into the MR system and ranked for severity and degree of development effort. Low cost problems at the levels of failure, critical, major, and serious—e.g. broken links—were delegated and fixed immediately.

Concurrently, areas where the heuristic evaluation outcomes combined with problems pertinent to scientists' demonstrated practices in the field (as captured in the user models) were examined. From these analyses, important combinations of problems found by the HE surfaced—combinations that implied problems related to higher order cognitive task needs. For example, problems 3, 6, 8, 9, 12, and 13 in Table 1 were observed as a recurrent cluster in the field observations as part of scientists' higher order task for locating interesting genes and relationships expediently. For this task, scientists progressively narrow down results sets based on several meaningful attributes and on validity standards, such as genes/gene products enriched for neurodegenerative processes. Once combined, this set of HE problems revealed scientists' difficulty manipulating queries and output sufficiently to uncover potentially interesting relationships. Thus beyond easy fixes—e.g. column cues for sorting—deeper implications of a tool's actual usefulness were uncovered by the combined HE problems and user model.

Shaped by the user models developed at our center and by ongoing research in our into design requirements, issues like the example above were presented to the usability and development teams and then brought to the priority setting committee. For example, problems related to users being able to narrow down to interesting results led to realizations that the tool needed to provide a more powerful search mechanism, extensive indexing, and interfaces that allowed users to construct/revise queries using multi-dimensional. Another priority setting issue suggested by the HE outcomes and better understood through the user models was the need for specific types of additional content for users' validation purposes. Both needs received high priority. Additionally, as the software developers became more aware of the value of these usability techniques, we started to get requests for the heuristic evaluation instrument itself so that programmers could keep the criteria in mind while in the process of developing their software.

## 5 Discussion

Developing the heuristic evaluation instrument was an iterative process as the evaluators discovered its weaknesses and strengths during the course of evaluations. Many of the heuristics turned out to be redundant and were either combined or discarded. Close inspection of the tools also engendered new heuristics as evaluators noticed additional usability problems. Accompanying comments proved to be crucial and were made mandatory for any problems found in later evaluations. The severity

numbering system also proved to be too abstract and will be replaced by ratings that mirror the ones used in the MR system. Finally, some heuristics in the instrument proved to be too theoretical or complex to be useful (e.g. “salient patterns stand out”) and had to be removed or refined. Some of these difficult heuristics were less concrete and were often better suited to incorporation and analysis within the user model.

In tool assessments, heuristics alone identified isolated problems and a few inaccuracies. Combined with the user model, the heuristic evaluations enabled us to uncover problems related to integrated tasks associated with scientists’ higher order analysis and reasoning.

Evaluators’ written comments, omissions, imprecision in some proposed design changes, and lack of agreement about certain items were vital in cuing us to further examine particular problems or combinations of problems in light of the user models. Had time and resources permitted, reliability testing would have diminished disagreements. A positive unintended consequence of these disagreements, however, was that they revealed where developers’ awareness of user tasks was incomplete. For example, in the heuristic evaluations, comments about “the ability to change the organization of results” indicated that the tool did not make it obvious that columns could be sorted. The user model revealed, however, that the untransparent sorting was only one shortcoming related to this specific heuristic. In actual practice, scientists’ analysis and judgments required tools to provide a combined set of sorting-and-filtering interactions to rearrange results into multidimensional groupings—i.e. interesting relationships. Reframed to account for this need, this problem led to high priority, enhanced functionality. Unlike in Hollinger’s study, many usability problems—framed in ways that join heuristic evaluation outcomes and user models—were given high priority status.

For such achievements, collaborations across specialties were critical—formally and informally. Developers, web specialists, project managers, scientifically expert bioinformatics specialists, and the usability, scientific, and computer science directors all played distinct roles in shaping the perspectives needed for strategically determining and then implementing a better match between tools and systems biology tasks. In the process, people across specialties grew increasingly aware of each others’ perspectives and began slowly evolving a shared language for articulating them. This process is often termed “double-loop learning” and is essential for innovation [1]. One example of this cross-organizational learning is the software developers’ requests for the heuristics to help guide software development.

Vital to this learning and the common grounding on which it rests is the perennial challenge of assuring that heuristics are expressed in the right grain size and language. As with other research focused on this goal, our center’s efforts have highlighted places to make heuristics more concrete and ways to join outcomes with user models.

## 6 Conclusions

In our center’s case, collaborative communication, shared language, and greater awareness—i.e. double-loop organizational learning—were integrated into and developed from heuristic evaluations. We found a way to use this discount usability inspection method combined with user models and newly implemented organizational processes,

to reframe problems and to gain buy-in for short and long term usability improvements aimed at scientists' cognitive task behaviors. Heuristic evaluations coupled with user modeling revealed problems related to the higher order cognitive flow of analysis. Combined with organizational and software development processes that encouraged attention to usability, heuristic evaluations produced results and recommended changes that received high priority. Moreover, developers and directors who previously had not considered usability in choices they about knowledge representations or functionality now grew increasingly sensitive to the implication of their choices from a user perspective. Our center continues to refine the instrument and apply it to other tools and is simultaneously creating a complementary instrument for heuristic evaluation of interactive visualizations in bioinformatics tools.

## References

1. Argyris, C., Schön, D.: *Organizational learning II: Theory, method and practice*. Addison Wesley, Reading (1996)
2. Chattratichart, J., Lindgaard, G.: A comparative evaluation of heuristic-based usability inspection methods. In: *Proceedings of ACM CHI 2008 Conference*, pp. 2213–2220. ACM Press, New York (2008)
3. Cockton, G., Woolrych, A.: Understanding inspection methods: lessons from an assessment of heuristic evaluation. In: Blandford, A., Vanderdonckt, J. (eds.) *People & Computers XV*, pp. 171–192. Springer, Berlin (2001)
4. Frokjaer, E., Hornbaek, K.: Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction* 14, 1–33 (2008)
5. Gerhardt-Powals, J.: Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interactions* 8, 189–211 (1996)
6. Hartson, H., Andre, T.S., Williges, R.: Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* 13, 373–410 (2001)
7. Hollinger, M.: A process for incorporating heuristic evaluation into a software release. In: *Proceedings of AIGA 2005 Conference*, pp. 2–17. ACM Press, New York (2005)
8. Law, E.L.-C., Hvannberg, E.T.: Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In: *Proceedings of ACM NordiCHI 2004*, pp. 241–250. ACM Press, New York (2004)
9. Mirel, B.: Supporting cognition in systems biology analysis: Findings on users processes and design implications. *Journal of Biomedical Discovery and Collaboration* (forthcoming)
10. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.I. (eds.) *Usability Inspection methods*, John Wiley, Chichester (1994)
11. Nielsen, J.: Enhancing the explanatory power of usability heuristics. In: *Proceedings of ACM CHI 1994 Conference*, pp. 152–158. ACM Press, New York (1994)
12. Schmettow, M., Vietze, W.: Introducing item response theory for measuring usability processes. In: *Proceedings of CHI 2008*, pp. 893–902. ACM Press, New York (2008)

## Supplemental Material: Adapted Heuristics

Heuristic	Severity Rating 0 = no problem 5=major problem	Comments
<b>First Impression</b>		
Does the tool fit the overall NCIBI look and feel?	0 1 2 3 4 5 N/A	
Does it look professional?	0 1 2 3 4 5 N/A	
Is the tool appropriately branded with funding source and NCIBI, CCMB, and UM logos?	0 1 2 3 4 5 N/A	
Does the tool link back to UM, CCMB, and NCIBI?	0 1 2 3 4 5 N/A	
Is it clear what to do and what to enter? (limitations are clear, how to format query is clear, what options user has, if a user needs to enter terms from some taxonomy/ontology access to those terms is available for user to choose from)	0 1 2 3 4 5 N/A	
Are there examples shown and are they prominent?	0 1 2 3 4 5 N/A	
Is the display consistent with user conventions for web pages/apps?	0 1 2 3 4 5 N/A	
Is it clear why use the tool and to what purpose?	0 1 2 3 4 5 N/A	
Does it require minimal steps to get started quickly?	0 1 2 3 4 5 N/A	
Is the cursor positioned in the first field that requires entry?	0 1 2 3 4 5 N/A	
Is help readily available?	0 1 2 3 4 5 N/A	
Is it clear how current the data are?	0 1 2 3 4 5 N/A	
Is it clear how current the website is?	0 1 2 3 4 5 N/A	
Are data sources cited and identified?	0 1 2 3 4 5 N/A	
Are appropriate publications cited?	0 1 2 3 4 5 N/A	
Are there any broken links?	0 1 2 3 4 5 N/A	
Are the page titles (displayed at the top of the browser) meaningful and change for different pages?	0 1 2 3 4 5 N/A	
Are page elements aligned (e.g. in a grid) for readability?	0 1 2 3 4 5 N/A	
Is the site readable at 1024x768 resolution?	0 1 2 3 4 5 N/A	
Is the text readable? (e.g. size, font, contrast)?	0 1 2 3 4 5 N/A	
Does the page have appropriate metadata tags for search engines?	0 1 2 3 4 5 N/A	
<b>Search / Results</b>		
Is the length of processing time acceptable?	0 1 2 3 4 5 N/A	
Do adequate indicators show system status and how long it may take?	0 1 2 3 4 5 N/A	
Clearly shows if there are no query results?	0 1 2 3 4 5 N/A	
Clearly shows how many results query produces?	0 1 2 3 4 5 N/A	
Is it clear what produced the query results?	0 1 2 3 4 5 N/A	
Is it easy to reformulate query if necessary?	0 1 2 3 4 5 N/A	
Are there hints/tips for reformulating query for better results?	0 1 2 3 4 5 N/A	
If the query results seem high or low is it clear why?	0 1 2 3 4 5 N/A	
Are the results transparent as to what results are being shown and how to interpret it?	0 1 2 3 4 5 N/A	
Are the results displayed clearly and not confusing?	0 1 2 3 4 5 N/A	
Is there an ability to detect and resolve errors?	0 1 2 3 4 5 N/A	
<b>Interaction with Results</b>		
Is there an ability to filter or group large quantities of data?	0 1 2 3 4 5 N/A	
Is there an ability to change the organizations of results?	0 1 2 3 4 5 N/A	
Is there ability to undo, redo, or go back to previous results ?	0 1 2 3 4 5 N/A	
Are the mechanisms for interactivity clear?	0 1 2 3 4 5 N/A	
Is the logic of the organization clear?	0 1 2 3 4 5 N/A	
Are different data items (e.g. rows) kept clearly separate or delineated?	0 1 2 3 4 5 N/A	

If there are links is it clear where they go?	0 1 2 3 4 5 N/A	
If there are icons is it clear what they do?	0 1 2 3 4 5 N/A	
Do the link outs provide reliable return?	0 1 2 3 4 5 N/A	
Are the vital statistics/counts of information available?	0 1 2 3 4 5 N/A	
Do the names/labels adequately convey the meaning of items/features?	0 1 2 3 4 5 N/A	
Are data items kept short? Is there too much/little information?	0 1 2 3 4 5 N/A	
Is the density of information reasonable?	0 1 2 3 4 5 N/A	
Can you access the necessary data to assure validity? (e.g. sources)	0 1 2 3 4 5 N/A	
Can results be saved?	0 1 2 3 4 5 N/A	
Are the results available for download in other formats?	0 1 2 3 4 5 N/A	
Can the pages be easily printed?	0 1 2 3 4 5 N/A	
Is vertical scrolling kept to a minimum?	0 1 2 3 4 5 N/A	
Is there horizontal scrolling?	0 1 2 3 4 5 N/A	
<b>Comments</b>		
Additional comments go here		