

Categorizing the “Badness” of Usability Problems

Based on even a cursory review of the literature, it is evident that coming up with a viable method for categorizing the “badness” of usability problems is itself problematic within the usability community. With the exception of the double usability experts used in a Virzi study (pp. 465) and Nielsen’s use of multiple evaluators to yield “better-than-random agreement” (1994, pp. 50), studies have typically shown that usability professionals have great difficulty agreeing on the “badness” of identified usability problems. In some situations, this disagreement is observed even when controls are put in place to “increase agreement” (Lesaigle and Biers, pp. 587). Many discussions of the topic, such as those in Catani and Biers (pp. 1335) and Lesaigle and Biers (pp. 585), offer possible explanations for this disagreement but fail to provide guidance for rectifying the problem.

This paper describes some of the methods that have been proposed for categorizing usability problems, based on their “badness.” It also describes how each of the methods discussed would deal with a specific situation, and presents an argument for using one of these methods.

Methods for Categorizing Usability Problems

This section describes four methods that can be used to categorize usability problems, and provides some information about the benefits and limitations of each method.

Method 1: Severity Scales

The most widely used method for categorizing usability problems is a severity scale. Typical severity scales consist of 4-5 levels, each described by text and a numeric value. As one moves up or down on the scale, the severity of the problem increases or decreases. The factors taken into account when determining severity levels can include: perceived importance; frequency; market impact; persistence; criticality; probability; user impact; priority; and scope (global/local). Because “it is common to combine all aspects of severity in a single severity rating as an overall assessment of

each usability problem” (Nielsen, 1994, pp. 47), it can be confusing for usability professionals to compare different variations and to make a decision about which scale to use. This section describes a few variations.

Jakob Nielsen’s “Severity Ratings for Usability Problems”

Nielsen’s proposed severity scale is shown in Table 1 (1994, pp. 47).

Table 1 Severity Ratings for Usability Problems

Numeric Value	Textual Description
0	I don’t agree that this is a problem at all
1	Cosmetic problem only: need not be fixed unless extra time is available on project
2	Minor usability problem: fixing this should be given low priority
3	Major usability problem: important to fix, so should be given high priority
4	Usability catastrophe: imperative to fix this before product can be released

According to Nielsen, his severity ratings are based on three factors: frequency of occurrence; user impact; and the “persistence of the problem.” One can see from the textual descriptions that priority is incorporated into the scale, and Nielsen has also noted that market impact should be “assessed” (1994, pp. 47).

While Nielsen’s severity scale is no doubt very popular and widely used in the usability community, this author does not consider his scale easy to use. There are two reasons for this opinion. First, Nielsen does not provide enough guidance about which problems should fall into each category nor does he define ambiguous terms. For example, what is a “minor problem”? What qualifies as a “usability catastrophe” versus a “major usability problem”? Jeffries astutely notes that “add[ing] a prose description to a numerical judgement” is an important part of assisting usability professionals who must select a rating (pp. 289), but Nielsen’s severity scale illustrates that the content of the descriptions is also important. The second problem is that one cannot see how Nielsen incorporates frequency, user impact, persistence, and (possibly) market impact into the scale, nor what quantitative data (if any) he considers. For example, what frequency value is necessary to assign a usability problem a severity rating of 2? Furthermore, how do these three factors interact one

another? For example, does frequency have a greater weight than persistence? Given all these unanswered questions, it is not surprising that different usability professionals rate the same set of problems differently and that multiple evaluators are recommended for reliable categorizations (Nielsen, 1994, pp. 61). This scale (and many like it) rely heavily on the subjective judgement of the usability professional who assigns the rating.

Strict Calculations of Criticality, Probability, and Frequency

Rubin defined “criticality” as “the combination of the severity of a problem and the probability that the problem will occur” (Malaney, pp. 11). Categorizing usability problems based on criticality appears to reduce the amount of subjectivity inherent in severity scales like Nielsen’s by combining the severity rating with a calculable, quantitative value. Criticality has been calculated by each of the following formulas (from Malaney, pp. 12 and Günther, respectively):

Criticality = Severity + Probability of Occurrence, where the Probability of Occurrence = Percentage of total users that will be affected by the problem * Probability that a user from the affected group will experience the problem

Criticality = (Frequency Ranking * Severity Ranking) / 2

Unfortunately, Malaney’s formula seems daunting and time consuming, given that multiple variables require calculation. Table 2 illustrates how Günther’s more modest formula for calculating criticality can improve Nielsen’s severity ratings by removing some of the ambiguity:

Table 2 Chart for “Prioritizing Problems Found”

Severity	catastrophic problem	major problem	minor problem	cosmetic problem	not a problem
Frequency	> 90%	51-89%	11-50%	1-10%	<1%
Criticality Score	4	3	2	1	0
Fix Priority	Imperative	High	Low	--	--

Brooks described a scale (shown in Table 3) proposed by Lewis et al. in 1990, which relies solely on frequency data to categorize usability problems (pp. 269-270):

Table 3 Severity Scale for Usability Problems

Numeric Value	Textual Description
0	No users would have problems
1	Few users would have problems
2	More than half of the users have problems
3	Most users have problems

The Lewis et al. scale is similarly improved—though in the opposite way—when subjective severity ratings like Nielsen’s are added to the often questionable frequency and/or probability data obtained from small numbers of usability test participants (Virzi, pp. 463). Given these examples, it would appear as though merging quantitative data and subjective ratings can help stabilize methods that categorize problems using severity scales.

User and Market Impact

The Lewis et al. scale described above explicitly states how many users must experience a usability problem for it to be categorized at a certain level of severity, but does not indicate what to do about the impact the problem may have on the user. In a later study, Lewis seems to have recognized this, and states that “different participants might experience the same problem but might not experience the same impact” (pp. 371). He goes on to suggest that the “best strategy is to consider problem frequency and impact simultaneously to determine which problems are most important to correct rather than establishing a cutoff rule such as ‘fix every problem that appears two or more times.’ (Lewis, pp. 377). He also recognizes that he must indicate how usability professionals should judge user impact and provides four levels of classification based on “behavioral definitions” (Lewis, pp. 370). These behavioral definitions include data such as whether tasks were completed (and completed as expected), whether assistance was used, how often the problem was encountered, how long it took to recover, and whether the path to task completion was the most efficient (Lewis, pp. 370).

Other severity scales, such as the one proposed by Rubin (shown in Table 4), take user impact into account by incorporating this factor into their textual descriptions of each category (Andre, pp. 131; Malaney, pp. 11).

Table 4 Rubin's Severity Rankings

Severity Ranking	Severity Description	Severity Definition
1	Unusable	The tester is not able to or will not use a particular part of the product because of the way that the product has been designed and implemented.
2	Severe	The user will probably use or attempt to use the product here, but will be severely limited in his or her ability to do so.
3	Moderate	The user will be unable to use the product in most cases, but will have to undertake some moderate effort in getting around the problem.
4	Irritant	The problem occurs only intermittently, can be circumvented easily, or is dependent on a standard that is outside the product's boundaries.

Closely tied to the concept of user impact is market impact, or the impact that releasing a product with usability problems will have in terms of: complaint levels (Brooks, pp. 270); "profit, revenue, or expenses" (Wilson, 1999); "probability of loss of critical data" (Wilson, 1999), and so on. As previously stated, Nielsen also believed that market impact should be considered (1994, pp. 47). Brooks, however, is the only one who provides any suggestion about how market impact (specifically, complaint levels) can be considered when she states that "user tests need to be correlated against market data" (pp. 270).

Overall, the fundamental limitation of the "severity scales" methods is two-fold: there is no single method that incorporates all the factors that must be considered to accurately categorize usability problems; and there is not enough standardization and guidance about how usability professionals should relate quantitative data to subjective ratings and descriptions.

Method 2: Affinity Diagrams

Affinity diagramming is a completely different technique for categorizing usability problems. In affinity diagramming, each usability problem that has been uncovered during a usability test is written on a card or Post-It note. A usability professional or other facilitator then leads a cross-functional team as they collaboratively organize the usability problems into categories (Gaffney; Wilson, 1997). Affinity diagramming not only requires verbal participation, but also encourages every member of the team be physically involved by moving the pieces of paper (Stephen). The categories of

problems that eventually emerge are based on the team's understanding of the problems, and the way they perceive the relationships between those problems (McQuaid and Bishop, pp. 2). Once established, problem categories are reviewed to see if they can (and should) be consolidated or further divided (Gaffney). Once the categories are agreed upon, the team collectively names them and assigns priorities to them (McQuaid and Bishop, pp. 3; Wilson, 1997).

Affinity diagramming as a technique is beneficial for several reasons. First, the categories and priorities that result from this activity represent the communications and contributions of an entire development team (Nielsen, 1993; Stephen). Having everyone involved in this creative and decision-making process may make it easier for usability professionals to ask for changes that would improve the product's usability (Gaffney). Second, because affinity diagramming is essentially a guided brainstorming activity, it "enables [team members] to see patterns in the problems" and "envision solutions...to whole categories of problems" that might not have been considered otherwise (McQuaid and Bishop, pp. 2). Similarly, affinity diagramming can help reveal previously unknown requirements for the design of the product interface (Hackos and Redish, pp. 331). Third, affinity diagramming is a method that reduces a seemingly impossible task to one that is manageable. When there is a long list of problems, usability professionals (and others) may feel overwhelmed and not know where to begin. The simple act of transferring the problems from long lists to small pieces of paper appears to be helpful from a psychological standpoint (Gaffney; McQuaid and Bishop, pp. 2).

Affinity diagramming has limitations that are typical to most group activities. For example, this method requires that an alert facilitator keep the activity on track, ensure that certain individuals do not control the situation while others remain silent, and so on. For cross-functional teams consisting of more than 8 people, affinity diagramming may also present some logistical difficulties (Gaffney).

Method 3: Cost-Benefit Analysis

A cost-benefit analysis is a "technique used to determine the feasibility of a project or plan by quantifying its costs and benefits" (InvestorWords.com). A cost-benefit analysis can be done in a number of ways depending on how the terms "cost" and "benefit" are defined, which costs and benefits are considered, and the quantitative measurements that an organization's decision makers use to determine acceptable return on investment (Solution Matrix Ltd.). For usability problems, the "costs" are typically the monetary resources that must be expended by the company producing the product to fix the problem, and the "benefits" are those "users will experience if the

problems are fixed.” The result of a cost-benefit analysis is a categorization of usability problems into a high cost/low cost and high benefit/low benefit matrix (McQuaid and Bishop, pp. 3).

One advantage of performing a cost-benefit analysis is that it provides a consistent structure by which development teams (including usability professionals) can examine the importance of a particular usability problem (McQuaid and Bishop, pp. 3). Other methods allude to the fact that priority must be determined, but fail to supply any guidance in terms of criteria. A cost-benefit analysis also requires that user impact be taken into account, as well as explicitly calculated from data obtained during a usability test (Wilson, 1997). Another important benefit of using a cost-benefit analysis to categorize and prioritize usability problems is that it organizes the data into a form that is easy for managers and decision makers to understand. It assists usability professionals in justifying their recommendations and “building a strong, successful business case” for usability improvements (Solution Matrix Ltd.).

The primary disadvantage of performing a cost-benefit analysis is also one of its primary benefits; namely, the use of quantitative measures. First, to determine the benefit (also referred to as “impact”) that a usability change will ultimately have on users, usability professionals must first have “collect[ed] quantitative measures, like time to complete a task, time spent in errors, and number of errors” during a usability test (versus some other possibly cheaper and quicker usability method). They must also expend time and energy analyzing that data (Wilson, 1997). Additionally, it seems as though an accurate determination of benefit requires a usability test of an alternative design (to obtain quantitative data that can be compared to data obtained from the original). Second, the cost-benefit analysis may also present problems when the criteria for making business decisions is not well established or is inconsistent within an organization. It may be that different decision makers are used to making judgements based on their (possibly experienced) opinions rather than on a set of standard quantitative measures. In other words, it may still be difficult to get consensus on what changes should be made when a cost-benefit analysis is presented.

Method 4: User Action Framework (UAF)

Andre et al. recognized the need for a categorization scheme that would “minimize individual differences in reporting by providing practitioners with a standard process” (pp. 1). Building on earlier, similar taxonomies that had good points but remained somewhat problematic, they developed the User Action Framework (UAF) (Andre et al. pp. 2). The UAF is described by its creator as “a theory-based, interaction-style-independent structured knowledge base of usability issues and concepts” (Andre, pp. 157). Essentially, the UAF is a decision tree containing various

usability categories and subcategories, which usability professionals “traverse” to determine a problem’s proper category (Andre et al., pp. 1). The categories and subcategories available in the UAF knowledge base are “comprised of usability concepts, issues, and guidelines” (Thompson, pp. 59) and include those that deal with time, tasks, objects, and user interaction (Andre et al., pp. 2-3). In early studies, the UAF appears to yield more consistent problem categorizations by usability professionals than other methods, and any remaining variations are attributed to differences in the interpretations of problem descriptions (Andre, pp. 75).

The UAF is beneficial method for classifying usability problems for several reasons. First, the UAF does not require any time consuming or difficult statistical calculations; everything is encapsulated into the knowledge base. Categorizing problems is as simple and painless as using a decision tree. Second, the UAF allows usability professionals to traverse different paths yet converge on the same, final categorization of the problem (Andre, pp. 76). This eliminates some of the individual differences in problem categorization but still takes into account the experienced judgements of usability professionals. Third, software tools could be built around the UAF that would allow for the automation of decision tree traversals, further reducing the time it would take to categorize problems.

There does not appear to be literature in which the categories and subcategories that comprise the UAF are explicitly listed. For example, does the UAF take into account problem frequency, user impact, and so on? How are these factors considered and weighted? Perhaps the creators of the UAF are purposefully guarding this information (for competitive reasons), but without it, the overall comprehensiveness of the UAF cannot be thoroughly evaluated. However, even if one assumes that the UAF knowledge base is inadequate, it seems likely that usability professionals could contribute to it, possibly improving an already promising categorization method.

Method Application

Two of the five participants in a usability test have a problem that causes them to be unable to complete a task. The other three don't have the problem.

The situation described above provides two pieces of information: frequency (that is, how many participants had a problem) and user impact (that is, the problem resulted in users being unable to complete a task). Using Nielsen’s “Severity Ratings for Usability Problems” (shown in Table 1) to categorize this problem presents some difficulty, but is far from worthless. Although not explicitly stated in the textual descriptions, the user not being able complete the task seems to indicate at least a major

usability problem (level 3). Because the importance of the task to the overall use of the product and to the user is not described, however, it is not immediately clear whether or not the problem should be promoted to level 4. Given that only 2 out of 5 users experienced the problem, it is likely that usability professionals would be more inclined to leave the problem at level 3.

In theory, Günther's chart (shown in Table 2) takes some of the guesswork out of Nielsen's method by adding a criticality score. Unfortunately, it fails in practice. A calculation of frequency for this scenario indicates that 40% of users experienced the problem. If used alone, this frequency factor would cause the problem to be categorized as minor (level 2). Günther provides no information about how to get a "frequency ranking" from the percentage, so one is left to guess by the chart that this would probably be equivalent to a frequency ranking of 3. When severity (level 3) is added, the criticality score becomes $(3 \times 3) / 2 = 4.5$, a value far enough off the chart to promote the problem to "catastrophic."

Note: The next set of categorization methods add some confusion because they reverse the meaning of the numeric values. While Nielsen and Günther use a high number for more severe problems, Lewis and Rubin use a low number.

The Lewis et al. "Severity Scale for Usability Problems" (shown in Table 3) makes it easy to categorize this problem (as level 1, few users have problems), but the categorization seems pointless. If Lewis' behavioral definitions are added, the problem is still categorized as a level 1 ("scenario failure") problem, but is more helpful because it indicates the user impact. The scenario failure category is defined as one in which "the problem caused *the* participant to fail to complete a scenario by either requiring assistance to recover from the problem or producing an incorrect output" (Lewis, pp. 370; emphasis added). Note that this definition (as well as Lewis' other behavioral definitions) allude to "the" participant, and provide no indication as to whether the number of users experiencing a problem is a factor in its categorization. Perhaps the number of users who experienced the problem is irrelevant, given that it was a "show stopper" for at least one (who in a usability test might represent some portion of the total user population). The use of Rubin's "Severity Rankings" (shown in Table 4) is identical to Lewis' later version for this case; because "the" user could not complete the task, the severity ranking would be level 1: Unusable.

To determine the market impact as suggested by Brooks, a usability professional would need access to market data that indicated how many complaints their organization received about the product and problem in question. If the 5 participants used in the test were truly representative of a larger user population, one could expect that 40% of all users would experience a problem that prevented them from completing a task, and are likely to complain about this product. From a business perspective, the

possibility of having 40% of users unhappy with the product would seem unsatisfactory. The problem would ideally be assigned a priority that is high enough to be fixed before the product's release.

There is not enough information provided in the situation description to be able to conduct an affinity analysis or a cost-benefit analysis, and it is unknown as to how the User Action Framework (UAF) would categorize the problem.

An Argument for Using an Integrated Categorization Method

Each of the categorization methods described in this paper have strengths and weaknesses. None appear to be complete in terms of the factors they consider nor in terms of how they make use of the quantitative data obtained from a usability test. Given these problems, this author is of the opinion that a categorization method that incorporates as many strengths of the available methods as possible and is as easy for usability professionals to use would be the best choice.

Although there is currently no way to tell which factors are incorporated into the User Action Framework (UAF) nor how they are qualified, the UAF seems to be the strongest method in terms of both its comprehensiveness and ease of use. The integrated categorization method described in McQuaid and Bishop also appears to be a sound method. Although they only divide their method into two steps: "categorizing the problems" using affinity diagramming and "prioritizing the problems" using a cost-benefit analysis, McQuaid and Bishop are actually doing more than merging two methods (pp. 2-3). Their descriptions of performing these steps illustrate that other factors, including user impact ("how important it is to fix the categories from the users' perspective"), market impact ("how much time, effort, and cost the client must expend to fix the problem") (pp. 3), problem scope (pp. 1), and priority are also considered.

Conclusion

Although disagreement within a particular field of study can be beneficial and stimulating, the categorization of problems based on their "badness" is a serious impediment for usability professionals. If problems are to be properly considered, usability test reports must be organized in a fashion that includes a statement about

their effect. As Jeffries indicates, "making an explicit severity assessment...help[s] developers prioritize changes under typical time-constrained conditions," and "...helps the evaluators assess whether the proposed solution is consistent with the size of the problem it solves" (pp. 286). Having a "clear definition of severity and rating categories" would also give more credibility to our usability test reports, erasing any misconception that problem severity is based solely on subjective judgements (Catani and Biers, pp. 1335). Additionally, having a standard method by which all usability professionals categorize problems would "allow [for] comparisons between the results of different inspection methods" and perhaps advance (or change the direction of) the usability profession (Desurvire, pp. 197). Until usability professionals converge not only on the method used but also on the criteria used to assign usability problems to a particular category, this dilemma will continue to plague our profession and hinder our otherwise valuable work.

References

Andre, T.S. (2000). "Determining the Effectiveness of the Usability Problem Inspector: A Theory-Based Model and Tool for Finding Usability Problems." Doctoral Dissertation submitted to Virginia Polytechnic Institute and State University, Blacksburg, Virginia. Accessed 12 October 2002. <http://scholar.lib.vt.edu/theses/available/etd-04122000-09440030/unrestricted/andre.pdf>.

Andre, T.S., Belz, S.M., McCreary, F.A., and Hartson, H.R. (2000). "Testing a Framework for Reliable Classification of Usability Problems." In Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting, Santa Monica, CA: Human Factors and Ergonomics Society. (Also online.) Accessed 12 October 2002. <http://people.cs.vt.edu/~hartson/UAF%20reliability/hfes%202000%20uaf%20reliability.pdf>.

Brooks, P. (1994) "Adding Value to Usability Testing." In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, John Wiley and Sons, Inc. New York.

Catani M.B., and Biers, D.W. (1998). "Usability Evaluation and Prototype Fidelity: Users and Usability Professionals." In Proceedings of the Human Factors Society 42nd Annual Meeting, pp. 1331-1335.

Desurvire, H.W. (1994) "Faster; Cheaper! Are Usability Inspection Methods as Effective as Impirical Testing?" In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, John Wiley and Sons, Inc. New York.

-
- Gaffney, G. (1999). "Affinity Diagramming" page. Part of Information and Design's Usability Techniques series. Accessed 12 October 2002.
<http://www.infodesign.com.au/usability/affinitydiagramming.html>.
- Günther, S. (2002). "Usability Inspection Methods" page, "Prioritizing Problems Found" section. Accessed 12 October 2002. http://www.f4.fhtw-berlin.de/~s0391333/MMA/Usability_Heuristics_Basics.htm.
- Hackos, J. T., and Redish, J. C. (1998). User and Task Analysis for Interface Design. John Wiley & Sons, Inc.
- Jeffries, R. (1994) "Usability Problem Reports: Helping Evaluators Communicate Effectively with Developers." In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, John Wiley and Sons, Inc. New York.
- Lesaigne, E.M. and Biers, D.W. (2000). "Effect of Type of Information on Real-time Usability Evaluation: Implications for Remote Usability Testing." In Proceedings of the IEA 2000/HFES 2000 Congress (6); pp. 585-588.
- Lewis, J. (1994). "Sample Sizes for Usability Studies: Additional Considerations." In Human Factors (36), pp. 368-378.
- Malaney, N. (2001). "A Practical Guide to Usability Testing: Usability Guide: Chapter 5: After Usability Testing" page. Accessed 12 October 2002.
<http://home.ix.netcom.com/~amalaney/IndStudy/Home.html>.
- McQuaid, H.L. and Bishop, D. (2001). "An Integrated Method for Evaluating Interfaces" page. In the Usability Professionals' Association 2001 Conference Proceedings. Accessed 12 October 2002. http://www.maya.com/WeAre/Staff/papers/MCQUAIDBISHOP_UPA2001_paper.pdf.
- Nielsen, J. (1993). "UPA'93: Usability Professionals Association Annual Meeting" page. Accessed 12 October 2002. <http://www.useit.com/papers/tripreports/upa93.html>.
- Nielsen, J. (1994) "Heuristic Evaluation." In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, John Wiley and Sons, Inc. New York.
- Solution Matrix Ltd. (2002). "Cost Benefit Analysis: A Solution Matrix Ltd. Mini-whitepaper." Part of the Business Case Analysis page.

- Stephen, D. (2001). "SENG (Software Engineering) 611 Technique Review: Organizing and Prioritizing Requirements and Formal Specifications" page. University of Calgary. Accessed 12 October 2002. <http://sern.ucalgary.ca/~daves/SENG611/organize.htm>.
- Thompson, J.A. (1999). "Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation." Masters Thesis submitted to Virginia Polytechnic Institute and State University, Blacksburg, Virginia. Accessed 12 October 2002. <http://scholar.lib.vt.edu/theses/available/etd-121699-205449/unrestricted/thesis.pdf>.
- Virzi, R.A. (1992). "Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough?" *In Human Factors* (34) 2, pp. 457-468.
- Wilson, C. (1997). "Usability Techniques: Analyzing and Reporting Usability Data." In the October, 1997 issue of Usability Interface. Accessed 12 October 2002. <http://www.stcsig.org/usability/newsletter/9710-analyzing-data.html>.
- Wilson, C. (1999) "Reader's Questions: Severity Scales." In Usability Interface (5), 4 and <http://www.stcsig.org/usability/newsletter/9904-severity-scale.html>.

