JUS
JOURNAL OF USABILITY STUDIES
Vol. 10, Issue 1, November 2014 pp. 17-25

The Relationship Between Problem Frequency and Problem Severity in Usability Evaluations

Jeff Sauro
Principal
Measuring U
201 Steele Street
Suite 200
Denver, Colorado
United States
jeff@measuringu.com

Abstract

The relationship between problem frequency and severity has been the subject of an ongoing discussion in the usability literature. There is conflicting evidence as to whether more severe problems affect more users or whether problem severity and frequency are independent, especially in the cases where problem severity is based on the judgment of the evaluator. In this paper, multiple evaluators rated the severity of usability problems across nine usability studies independently using their judgment, as opposed to data driven assessments. The average correlation across all nine studies was not significantly different than zero. Only one study showed a positive correlation between problem frequency and severity. This analysis suggests that researchers should treat problem severity and problem frequency as independent factors.

Keywords

Usability problems, usability testing, inter-rater reliability, intra-rater reliability, correlation, severity ratings

Introduction

User experience professionals are responsible for helping software developers make decisions about what to fix in an interface to make it more usable (Dumas & Redish, 1999). A central method in identifying improvements comes from formative usability testing whereby representative users attempt some tasks on an interface and a usability evaluator notes the problems and makes recommendations. The identification of usability problems is the key output of this type of usability test (Nielsen, 1993; Rubin & Chisnell, 1994). A simple problem description along with screen shots and some context under which the problem was discovered may provide sufficient detail to make decisions when the problems are understandable and noncontroversial.

However, it is often the case that usability tests generate more usability problems than development teams can address. To prioritize which issues developers need to fix, user experience professionals should account for the two critical but separate elements that test participants encounter in a usability test: problem frequency and problem severity. For example, it could be that 1 out of 10 participants had a problem with a financial website that inadvertently posted his or her private information for everyone to see (low frequency and high severity). Or 9 out of 10 participants in the same study might have been mildly irritated by having to deselect an option to receive marketing communications from a financial firm (high frequency and low severity).

Measuring the frequency of a problem is generally straightforward. The number of participants in a usability test who encounter a problem is divided by the total number of participants, which generates the proportion encountering the problem. For example, if 1 out of 5 participants encounter a problem, the problem frequency is .20 or 20%. There are nuances to reporting problem frequency, such as how to deal with participants who encounter the problem multiple times, sometimes called persistence (Nielsen, 1993), and how to deal with different evaluators discovering different problems (Lewis, 2012). When these nuances have been accounted for, it is straightforward to compute a problem percentage.

Rating the severity of a problem is less objective than finding the problem frequency. First, there are a number of ways to assign severity ratings and there tends to be disagreement between evaluators when assigning severity (Nielsen, 1993). While there have been a number of different severity rating systems proposed over the last few decades, in general, each method proposes a similar structure: a set of ordered categories reflecting the perceived impact the problem has on the user, from minor to major (Hertzum, 2006). Lewis (2012) made the distinction between judgment driven and data driven severity ratings. Judgment driven ratings rely on the stakeholders of the study to determine how impactful the usability problems are. Data driven ratings use criteria such as ease of correction, likelihood of usage, and impact on task completion (Hassenzahl, 2000). For example, Nielsen (1993) proposed a 5-point scale from cosmetic (1) to catastrophic (4) with zero indicating no usability problems. Rubin and Chisnell (2008) proposed a 4-point scale from unusable (4) to irritating (1). Dumas and Redish (1999) also proposed a 4-point scale from a subtle issue (4) to a larger issue that prevents task completion (1). The latter two approaches are examples of data driven prioritization.

Rubin and Chisnell (2008) proposed the concept of criticality that combined problem severity and frequency. He proposed using a 4-point scale for frequency instead of an actual observation: 1 represents an estimated problem frequency of less than 10%, 2 represents an estimated problem frequency from 11% to 50%, 3 represents 51% to 80%, and 4 is > 90%. By adding together the severity and frequency rating, problems can range from 2 to 8. A rating of 5 therefore can mean a problem that affects a lot of users but is minor or a problem that affects very few users but is severe.

The relationship between problem frequency and severity has been the subject of an ongoing discussion in usability literature. Most notably, Virzi reported that more severe usability issues tended to happened more frequently (Virzi, 1990, 1992). In Virzi's 1990 study, three undergraduate students, under his supervision, observed 20 students attempting 21 tasks on a software calendar. The evaluators identified 40 unique usability issues and then rated the severity of the issues independently using a 7-point scale from low impact to high impact. Differences in severity ratings were then reconciled by the evaluators until they agreed on a

severity rating. A positive correlation between problem severity and frequency was reported (r = .46).

To minimize bias introduced in severity ratings by evaluators' prior knowledge of problem frequency, Virzi conducted a follow up study using a voice response system (1992). Twenty participants attempted seven tasks, and two usability evaluators identified 17 usability problems. Each problem was assigned a severity rating using a 3-point severity scale (low, medium, or high). Six separate usability experts who were familiar with the software were given descriptions of the problems without any frequency information and rated the severity using the same 3-point scale. The mean correlation between the six experts and the independent evaluators was r = .46. The average correlation between just the six experts was r = .33. The correlation between problem severity and frequency from the 17 problems was, unfortunately, not reported. A conclusion from these findings is that practitioners conducting usability tests would need fewer users to detect more severe problems. In his studies, virtually all the problems rated high in severity were found with the first five participants. This is important as many lab-based usability studies are run with a small number of participants, typically between 4 and 10 (Sauro, 2010).

However, in attempting to replicate Virzi's (1990, 1992) findings, Lewis failed to find a similar relationship between the frequency of a problem and its severity (1994). Lewis examined the usability data from 15 participants attempting tasks on productivity software (word processing, spreadsheets, calendar, and mail). A total of 145 usability problems were observed and classified by the same observers on a 4-point data-driven severity scale (1= scenario failure to 4 = inefficiency). The correlation between severity and frequency was nonsignificant (r = .06). Lewis recommended treating severity and frequency as independent. That is, a usability problem is just as likely to be one of low severity as it is of high severity.

In a study that compared the results of a heuristic evaluation with user testing, Nielsen (1994) reported that 11 evaluators identified 40 problems in a prototype of a telephone integration system. Of these problems, 17 were confirmed by four participants in a usability test. Nielsen reported a statistically significant correlation of r = .46 between the number of users having a problem and the evaluators mean severity ratings. This correlation was the same magnitude as found by Virzi (1990). This study differs from the Virzi and Lewis studies as the severity ratings were based on problems identified by evaluators, not problems uncovered from observing the users.

In a study by Woolrych and Cockton (2001), one evaluator of a PowerPoint drawing editor observed 12 users encountering 16 total problems. Problem severity was assigned using a 3-point scale based on time wasted or task failure. The authors did not report the correlation but provided a matrix that showed which user encountered each of the 16 problems and the severity rating the single evaluator assigned that incident. Hertzum (2006) calculated the correlation between severity and frequency from this data and found a negative nonsignificant correlation (r = -.29). Also unlike the Virzi and Lewis studies, the severity rating was calculated by averaging together each participant's assigned severity rating. For example, on problem two, seven participants (44%) encountered the problem, five were assigned a high impact severity, one a middle impact, and one a minor impact.

Hertzum also calculated the correlation from data reported in a study by Jacobsen, Hertzum, and John (1998). In this study, four evaluators observed videotape of four users thinking aloud as they attempted tasks on a multi-media authoring system. He found a positive and significant correlation (r = .46) between the severity and frequency of identified problems. All evaluators observed all four users so therefore had clear knowledge of problem frequency when assigning severity ratings.

In a study by Law and Hvannberg (2004), two evaluators observed 19 participants using an educational content application. In total the participants identified 88 problems, and the evaluators found a small positive correlation (r = .10).

In total, six studies provided an average correlation between frequency and severity of r = .22 (Fisher transformed values of .46, .06, .46, -.29, .46, and .10, for each study respectively), suggesting a small positive relationship between problem frequency and severity.

In many of the studies, however, the evaluators judging severity also had knowledge of problem frequency, especially in the studies by Jacobsen et al. (1998), Woolrych and Cockton (2001), Nielsen (1994), and Law and Hvannberg, (2004). One of the reasons it is believed Virzi found a correlation between problem frequency and severity is that even though problem frequency was removed from the problem description, it's possible that information contained in these paragraph long descriptions can contain information about how widespread an issue is. For example, a problem with a global navigation element will likely impact more users than something on a more obscure screen. This was a concern also brought up by Lewis (1994).

Despite the widespread use of usability testing in industry and opinions about whether more severe problems occur more often (and are, therefore, easier to detect with fewer participants), there is a shortage of studies in the literature to help resolve the issue—despite the discussion continuing for over twenty years.

The purpose of this study was to assess the relationship between problem frequency and problem severity using nine usability problem sets across different application types, different evaluators, and (where possible) to minimize the bias of knowing how many users encountered a problem when assigning severity ratings. It was hypothesized that the correlation between problem frequency and severity is nonsignificant.

Methods

We investigated the relationship between frequency and severity using datasets collected from nine usability tests (as mentioned in this paper as Study 1 through Study 9) on websites and mobile applications. There were three types of usability tests conducted: in-person moderated, remote moderated, and remote think aloud (where unmoderated participants from Usertesting.com were trained to think aloud). For the moderated in-person tests, a test facilitator and note taker sat together and interacted with the participant in the same usability lab in Denver, Colorado. For the remote moderated studies, the participants attended by joining a Citrix GoTo Meeting, and they shared their screens with the test facilitator while attempting the tasks in the study. Both in-person and remote moderated usability tests were conducted by a single facilitator and note taker. However, there were different levels of interaction with the facilitator and note taker for the three types of usability tests, as described below. For remote moderated studies, the facilitator interacted with the participant over the phone. For the remote think aloud study, participants were given tasks to attempt on a website and asked to think aloud while attempting the tasks. Each participant's screen and audio were recorded using the website Usertesting.com, but the participants had no interaction with a facilitator. The usability test sessions lasted between 45 and 90 minutes with tasks lasting between 5 and 20 minutes. Details about the nine studies are shown in Table 1 below.

Table 1. Description of the Nine Usability Datasets

ID	Device	Interface Tested	Participants	Evaluators	Issues	Tasks
1	Desktop	Ecommerce website	17	4	75	6
2	Desktop	Sports merchandise website	12	3	37	4
3	iPhone	Cable provider App	7	4	36	6
4	iPhone	Cable provider App	7	2	24	6
5	iPad	Cable provider App	5	2	25	9
6	iPad	Cable provider App	5	2	32	9
7	iPad	Cable provider App	6	2	37	10
8	Desktop	Antivirus Web-App	20	2	29	8
9	Tablet	Ecommerce Website	20	3	36	6

Evaluators

There were two to four evaluators who assigned severity ratings to the usability problems. All evaluators worked in the same company and interacted with each other on the studies. These evaluators represented a mix of experience with observing and coding usability problems. Two junior evaluators had observed 30 to 50 total users across four usability studies over a two-year period. Two senior evaluators had observed at least 200 users across 30 usability tests over the prior two-year period. At least one senior evaluator was involved in assigning severity ratings for each usability test. The evaluators all had familiarity with all products and domains but were not considered subject matter experts.

Participants

The 99 participants ranged in age from 18 to 66 and had a mix of experience with the devices and websites or applications being used. For all studies, there was a total of 53 female participants and a total of 46 male participants. All participants were from the US and spoke English. Participants were recruited using online advertisements and compensated between $50 and $100 for their time.

Measures

Usability problems were recorded using an Excel spreadsheet. The facilitator and note taker identified the problems and problem frequency in each study. The total number of participants who encountered the problem divided by the total number participating in the study generated the problem frequency.

Problem severity was assigned in all cases using three levels: 1 = minor, causes some hesitation or slight irritation; 2 = moderate, causes occasional task failure for some users or causes delays and moderate irritation; and 3 = critical, leads to task failure or causes user extreme irritation. For more details on different problem severity rating systems and an example of the systems see Sauro (2013).

Procedure

Problem frequency information was removed from the problem descriptions before the evaluators independently assigned the problems a severity rating. Each usability study had at least two evaluators independently rate the severity of the problems. One of the evaluators for each study was also the test facilitator and therefore had some knowledge of problem frequency—even though the frequency information was removed when severities were assigned. Problem severities were then aggregated and an average problem severity was generated, as recommended by Nielsen in his 1992 study. For example, a problem from Study 8 was "Software screenshots appeared interactive," which received a moderate severity rating (2) from the experienced evaluator/test facilitator and a minor (1) from the less experienced evaluator who did not observe the sessions. The average problem severity rating from these two evaluators was 1.5. The average problem severity was then correlated with the problem frequency for each of the problems identified by study using Pearson correlations.

Problem severity is arguably an underlying continuous distribution and using only three categories may artificially reduce the correlation between frequency and severity. Correlations were averaged after using the Fisher transformation (see, for example, Bobko, 2001). To assess the inter-rater reliability, correlations between evaluators were calculated and averaged. To assess intra-rater reliability (how consistent the same evaluator is in assigning ratings), evaluators for one of the studies repeated their severity ratings after a delay of one day. To help reduce the bias of having knowledge of problem frequency, a senior evaluator who had no involvement in the project and who doesn't work with the other evaluators independently assigned problem severities for one of the studies. This evaluator (not the author) has over 25 years of experience conducting usability tests.

Results

The correlations between frequency and severity for each of the nine studies are shown in Table 2 below. Only one of the datasets had a correlation significantly different than zero (Study 8).

The average inter-rater reliability was r = .52 (ranging from r =.02 to r = .84). The intra-rater reliability for Study 9 from three evaluators who assigned the same ratings one day apart was r = .58 (the transformed average of .48, .47, and .74).

Table 2. Pearson Correlations and Inter-Rater Reliabilities

Study ID	Pearson *r*	Inter-rater *r*
1	0.02	0.37
2	0.29	0.33
3	0.26	0.45
4	-0.04	0.09
5	-0.29	0.80
6	0.22	0.84
7	-0.32	0.64
8	0.44*	0.46
9	-0.04	0.25
Mean	0.06	0.52

Note: To estimate what the correlation would be if ratings were made on a more continuous scale, the polychoric correlation was also generated (Uebersax, 2006). The polychoric correlation for the datasets ranged from -.39 to .47 with an average correlation of .056, providing similar results and the same conclusion as the Pearson correlation. Studies 7 and 8 had polychoric correlations statistically different than zero, compared to just Study 8 being statistically significant using the Pearson r.
* Statistically significant at the p < .05 level.

Recommendations

The primary research question was whether problem frequency and problem severity are correlated. Across the nine usability studies, only one study had a positive correlation that was statistically different than zero. The average correlation across these nine studies was r = .06 which is not statistically significant from zero. Of the nine studies, five had positive correlations and four had negative correlations. This study adds to the existing studies in the literature in two ways. First, efforts were made to keep problem frequency estimates separate from the evaluators making problem severity judgments. This was a confounding variable in most of the other studies found in the literature. Second, this study included nine datasets with sample sizes that were larger than the typical sample sizes in the literature review. This range of participants, from 5 to 20, along with different interface types and evaluators increased the chances of detecting a correlation between frequency and severity, if one existed.

This data suggests that there isn't strong evidence that more severe problems happen more frequently than less severe ones. Looking at multiple studies using different devices, facilitators, and evaluators minimizes the reliance on a single study with its flaws and idiosyncrasies to draw a conclusion about the relationship between frequency.

Caulton (2001) argued that the heterogeneity of a user population has an impact on problem identification. For example, in the international study by Law and Hvannberg, (2004), the authors found that problem detection rates were associated with different cultural backgrounds (languages and countries). These differing cultural backgrounds formed a heterogeneous user group that may have influenced problem detection rates. This connection to heterogeneous groups may explain a reason for low correlations between severity and frequency. However, in our study, all participants across the nines studies, while representing different skills and ages, came from the US and were likely more homogenous. In other words, we would expect a higher correlation given the homogeneity of the user groups and yet still failed to find one.

All things considered, having more points in a severity scale should increase the reliability of the severity rating. A 3-point scale was used across all nine studies. The results of the polychoric correlation analysis, which corrects for fewer scale points, suggests that increasing the number of severity levels would not affect the results of no correlation.

Limitations

Despite efforts to reduce the influence of problem frequency on severity ratings, at least one evaluator in each study had some knowledge of problem frequency. In addition, even if an evaluator who assigned the severity ratings was not a facilitator or note taker, it is likely some of the evaluators have knowledge of frequency when assigning severities (e.g., from sharing the same office and discussing projects in general). Given this bias toward finding a positive correlation between frequency and severity, it is surprising that the average correlation was just slightly above zero and four of the nine studies had nominally negative correlations. We expected it to be at least statistically different than zero.

In actual practice it is difficult to control for the impact of problem frequency when assigning severity ratings. It's often the case that only a single evaluator facilitates and reports the findings. Under these circumstances the concepts of problem frequency and severity may get comingled. That is, with the knowledge that a problem affects a lot of users, even a trivial one, it just seems to be more critical, even if the impact is minimal on the experience. While this is a problem for assessing the correlation between frequency and severity, it's probably not that harmful in practice. In addition, despite having detailed problem descriptions, it is difficult to have evaluators provide severity ratings when those evaluators did not observe and code the usability problems. This is a problem also mentioned by Jeffries (1994) and presents a challenge for researchers estimating the frequency and severity independently.

The generally low reliability between and within evaluators is consistent with earlier studies that found relatively low correlations between evaluators (e.g., Catani & Biers, 1998; Jacobsen et al., 1998; Lesaigle & Biers, 2000). The low reliability of the estimates can be improved by averaging the ratings from independent evaluators. See both Nielsen (1994) and Bobko (2001) for details on using the Spearman-Brown prophecy formula for how increasing evaluators will increase the reliability of the severity ratings.

It is likely the severity scheme used in this study was more judgment than data driven as evaluators didn't have systematic data-driven indicators, like task completion rates and proportion of the product affected, available when assigning severity ratings. There is some evidence that the more judgment based ratings have lower reliability than data driven schemes (Lewis, 2012). If severity ratings themselves are inconsistently applied, between and within evaluators, it becomes harder to establish a relationship between severity and frequency. Future research could use a more calibrated severity system with higher intra-rater and inter-rater reliability as well as isolating evaluators from any frequency information to look for correlations.

Conclusion

The analysis of these nine studies suggests there is no correlation between problem frequency and problem severity in usability problem sets. There is as much evidence that critical problems happen *less* frequently than minor problems as there is that problems occur more frequently. With little evidence supporting a correlation, it suggests the first few users are not more likely to uncover the more severe issues—contrary to one of the findings of Virzi (1992) and in support of the findings of Lewis (1994). This analysis does not contradict the other important findings of Virzi and Lewis—that the first few users are likely to uncover the most *frequent* problems—but suggests practitioners can still use a small sample size to identify the more common usability issues.

Practitioners should not expect a correlation between frequency and severity. They might see such a correlation in any given study, but they are just as likely to see no correlation or a negative one. The only safe strategy for a practitioner is to assume that the only driver of discovery is frequency and to never assume that high impact problems will necessarily have high frequency.

Tips for Usability Practitioners

The following are points practitioners can take away from this study:

- Practitioners should track both the frequency and severity of problems uncovered in a usability study.
- The severity of a problem is independent of how many users a problem will likely affect (frequency).
- In general, different usability evaluators will uncover different problems and rate the severity differently.
- Where possible, multiple evaluators should be used to rate the severity of problems independently; the average of the severity ratings is a more stable measure of severity than any single evaluator.

References

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management*. Thousand Oaks, CA: Sage Publications.

Catani, M. B., & Biers, D. W. (1998). Usability evaluation and prototype fidelity: Users and usability professionals. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 42*(19), 1331–1335. doi: 10.1177/154193129804201901

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*(1), 1–7.

Dumas, J., & Redish, J. C. (1999). *A practical guide to usability testing*. Portland, OR: Intellect.

Hassenzahl, M. (2000). Prioritizing usability problems: Data-driven and judgment-driven severity estimates. *Behaviour & Information Technology, 19*, 29–42.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction, 21*(2), 125–146.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability studies: Problem detection and severity judgments. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting, 42*(19), 1336–1340. doi: 10.1177/154193129804201902

Jeffries, R., (1994). Usability problem reports: Helping evaluators communicate effectively with developers. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 25–62). New York: Wiley.

Law, E. L., & Hvannberg, E. T. (2004). Analysis of combinatorial user effect in international usability tests. *In Proceedings of CHI 2004*. Vienna, Austria: ACM.

Lesaigle, E. M., & Biers, D. W. (2000). Effect of type of information on real time usability evaluation: Implications for remote usability testing. *Proceedings of the Human Factors and Ergonomics Society, 44*(37), 585–588. doi: 10.1177/154193120004403710

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors 36*(2), 368–378.

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (4th Edition), pp. 1267–1312). New York, NY: John Wiley.

Nielsen, J. (1992). Reliability of severity estimates for usability problems found by heuristic evaluation. *In Posters and Short Talks of the 1992 SIGCHI Conference on Human Factors in Computing Systems* (pp. 129–130). New York, NY: ACM. doi: 10.1145/1125021.1125117

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Morgan Kaufmann.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 25–62). New York: Wiley.

Rubin, J. & Chisnell, D. (2008) *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd Edition). Indianapolis, IN: Wiley Publishing Inc.

Sauro, J. (2010). How many users do people actually test? *Measuring Usability*. Retrieved October 1, 2013, from http://www.measuringu.com/blog/actual-users.php.

Sauro, J. (2013). Rating the severity of usability problems. *Measuring Usability*. Retrieved October 26, 2013, from http://www.measuringusability.com/blog/rating-severity.php .

Uebersax, J. (2006). The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement. Retrieved from: http://john-uebersax.com/stat/tetra.htm .

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. *In Proceedings of the Human Factors Society 34th Annual Meeting*. Santa Monica, CA: Human Factors Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*, 457−468.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), Proceedings of IHM-HCI 2001 Conference: Vol. 2 (pp. 105- 108). Toulouse, France: Cépadèus.

About the Author

Jeff Sauro
Mr. Sauro is the Principal and founder of Measuring U (measuringu.com). He is author of five books, including *Quantifying the User Experience* and the forthcoming *Customer Analytics for Dummies*. He has worked for Oracle, PeopleSoft, Intuit, and General Electric and holds a Masters from Stanford in Learning, Design & Technology and is completing his PhD in Research Methods & Statistics at the University of Denver.