Evaluating Severity Rating Scales for Heuristic Evaluation

Sascha Herr
Human-Computer Interaction Group
University of Bamberg
96045 Bamberg
sascha.herr@uni-bamberg.de

Nina Baumgartner
Human-Computer Interaction Group
University of Bamberg
96045 Bamberg
nina-yvonne-elisabeth.baumgartner@stud.uni-bamberg.de

Tom Gross
Human-Computer Interaction Group
University of Bamberg
96045 Bamberg
tom.gross@uni-bamberg.de

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2892454

Abstract
The heuristic evaluation is a widely applied discount usability evaluation method. Experts use the method to identify usability issues in interfaces and to rate their severity in order to establish a prioritisation of resource allocation. However, in practice, there are often large discrepancies between the individual severity ratings of experts, indicating challenges with the rating process and doubtful accuracy of ratings. This paper discusses these challenges by drawing from research on psychometrics, proposes solutions and reports preliminary findings of an empirical online study.

Author Keywords
Usability; Heuristic Evaluation; Rating Scale; Severity.

ACM Classification Keywords
H.5.2.e. Information Interfaces and Representation (HCI), User Interfaces, Evaluation/Methodology.

Introduction
Heuristic evaluation is widely applied in practice. Experts inspect a user interface by working through predefined representative tasks guided by established heuristics in order to find usability issues. The primary objective of the heuristic evaluation is to find all existing usability issues, the secondary objective is to quantify the severity of the found issues. The latter is

Nielsen Scale

- ○ I don't agree that this is a problem at all
- ○ Cosmetic problem only, need not be fixed unless extra time is available on the project
- ○ Minor usability problem: fixing the problem should be given lower priority
- ○ Major usability problem: Important to fix, should be given high priority
- ○ Usability catastrophe: Imperative to fix before release.

Practitioner's Scale

- ○ **Minor**: The problem causes some hesitation or slight irritation
- ○ **Moderate**: The problem causes occasional task failure for some users; causes delays and moderate irritation
- ○ **Major**: The problem leads to task failure; causes extreme irritation

Individual Factor Scale

Frequency How high is the number of users affected by the problem?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Difficulty How hard is it to overcome the problem for a user?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Workflow Impact How disturbing is the problem for the workflow of the user?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Persistence How often does one user encounter the problem?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Frustration How frustrated is the user due to the problem?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Market Impact How high is the impact of the problem on the popularity of the product?	Very low	☐ ☐ ☐ ☐ ☐	Very high
Fixing Effort How high is the effort for developers to fix the problem?	Very low	☐ ☐ ☐ ☐ ☐	Very high

Figure 1. Severity scales investigated in our study.

very important to prioritise the fixing efforts for the issues and to allocate development resources accordingly. However, research [7] as well as our own experiences with heuristic evaluation from usability consulting projects suggest that there is a lack of a common standard for the rating process. Ratings often differ tremendously between experts—a finding that is known as the evaluator effect [3].

In the light of these discrepancies, doubts about the accuracy of severity ratings emerge. Quick solutions to increase agreement between evaluators such as letting them discuss their ratings in the group or rate the severity entirely within a group are flawed since undesirable social effects such as dominance can occur that are detrimental to rating accuracy (e.g., groupthink [5]).

In this paper we suggest improvements to the rating process by incorporating findings from literature on psychometrics specifically with regard to the applied severity rating scale. Rating scales can induce various cognitive effects, biases, and errors [2]. Designing the rating scale to reduce them is likely to improve the accuracy and precision of ratings. In the following, we show challenges with prominent approaches, propose an improved severity rating scale, and report an online study that compares scales and evaluates our solution.

Our contribution is threefold: First, we show that the applied scale can indeed have an effect on the rating process. Second, we propose a more valid and informative rating scale. Third, we show that the time costs in terms of applying the new scale is relatively minor compared to its positive effects.

Severity Rating Scales

In this section we introduce several rating scales (cf. Figure 1 and Figure 2). We first identify challenges with the most documented and well-known severity rating scale—the Nielsen Severity Scale. We then discuss a scale that is found in practice—the Practitioner's Scale. Finally, we argue for a new individual factors scale.

Nielsen Severity Scale

There are many ways in which a rating scale can be the source of biases [2]. Careful consideration of the properties of a scale can reduce these biases and thereby improve the accuracy of ratings. For instance, the unipolar severity scale proposed by Nielsen [9] has five points associated with numerical values: 0 not a problem; 1 cosmetic problem; 2 minor problem; 3 major problem; and 4 usability catastrophe. From a psychometric perspective, there are several pitfalls associated with this format:

(1) In its current form, the scale is unbalanced, which is not justified without strong a priori evidence that most evaluators have an attitude towards one side of the scale. This biases ratings towards the side where more points are offered [2].

(2) It could be argued that the first option "not a problem" is conceptually not a part of the severity of a problem and therefore might hinder cognitive processing and introduce biases.

(3) There is evidence that the optimal number of points for rating scales is five to seven, in order to ensure reliability, validity, and discriminating power [11], which the scale does not provide with only four points for the severity construct.

Severity Scales Properties

Nielsen (N):

- Range: 5-point
- Categories: not a problem, cosmetic problem, minor problem, major problem, usability catastrophe.
- Unidimensional

Practitioner (P):

- Range: 3-point
- Categories: Minor, Moderate, Major
- Unidimensional

Individual Factors (IF):

- Range: 5-point
- Categories: Semantic differentials "very low" to "very high"
- Multidimensional: Frequency, Difficulty, Workflow Impact, Persistence, Frustration, Market Impact, Fixing Effort

Figure 2. Overview of severity scale properties.

(4) Severity is a multidimensional construct, which is poorly addressed in the scale. Nielsen [8] advises to take three factors into account: frequency, impact, and persistence. Furthermore, he proposes that the assessment of the market impact is important. Despite the acknowledgement of the multidimensionality of severity, practitioners typically provide only one overall rating.

(5) There is still a need for clarification of which factors comprise the severity construct. There might be more factors involved than yet discussed.

Practitioner's Scale
We conducted a preliminary study to investigate how practitioners rate the severity of usability problems. We interviewed five usability professionals and posted the following question in popular usability focused forums:

"I would like to ask you for your best practices for rating the severity of a usability problem when you apply a heuristic evaluation or an expert review.

1. *What kind of rating scale do you employ to assess the severity of usability problems?*

2. *Can you describe it (how many levels does it have, how are those levels named) and give your opinion on it?*

3. *Which factors or dimensions of the problem do you consider for coming up with the rating?"*

The gathered responses indicate that many professionals refer to the Nielsen scale, but also prominently conceptualised their own severity scale, which focuses on simplicity and fast conduction of the rating process. For this purpose, many professionals use only three categories to assign the severity of a problem: minor, moderate, major.

Individual Factor Scale
Many factors play a role in determining the severity of a usability issue, highlighting the multidimensionality of the severity construct. Informed by the preliminary study, body of literature, as well as psychometric methodologies, we propose an individual factor scale to improve rating accuracy. Specifically, the use of a summated rating scale entails advantages in validity and reliability of ratings [10]. For this rating scale, we employed seven factors that are estimated individually.

1. **Frequency**: The frequency of how often the problem arises in the population.

2. **Difficulty**: The difficulty for one user to overcome the problem.

3. **Workflow Impact**: The impact that the problem has on the workflow of the user.

4. **Persistence**: The persistence with which one user is faced with the problem.

5. **Frustration**: The frustration that emerges when a user encounters the problem.

6. **Market Impact**: The impact the problem has on the popularity of the product.

7. **Fixing Effort**: The effort for developers to fix the problem.

Fly-out Menu Mouse-Over-Difficulties

Description:
The website of an equipment seller has a navigation bar with many levels. Users have to focus the mouse precisely to get to the section they would like to visit by moving horizontally. If the mouse is moved too much vertically, the menu loses focus and users have to start from the beginning.

Figure 3. Example problem case.

These factors are rated on a 5-point semantic differential scale ranging from "very low" to "very high", addressing pitfalls 1 and 3 of the Nielsen scale. On theoretical grounds, multiple item scales are more reliable than single indicators since they individually represent different aspects of the construct [12], which alleviates pitfall 4. While not explicitly tackled in the Individual Factor scale pitfall 2 can be solved by extracting the "not a problem" option from the severity scale and offering it as a separate choice. Regarding pitfall 5, the proposed factors provide a first step in defining the multidimensionality of usability problem severity. The factors encompass different perspectives on the problem. Frequency forces evaluators to consider the overall reach of the problem. Difficulty, workflow impact, and persistence follow the pure single user perspective. Frustration targets the affective component of a usability problem. Market impact induces a business perspective. Finally, fixing effort establishes a developer perspective.

The advantage of the Individual Factor scale is twofold. First, by making the judgement of individual factors explicit, evaluators are more aware of their own criteria for establishing the ratings, which should improve accuracy and lead to better recommendations. Second, the additional data could provide usability professionals more grounds for discussion and developers more information for prioritising on certain aspects.

Method

We conducted an empirical online study to compare and evaluate the three scales.

Participants
In total, 103 participants completed the online questionnaire. No sampling restrictions were established, however, we focused on advertising on usability-related communities. The age ranged from 18 years to 56 years (M=31.9, SD=8.3). 48% of the participants were female. Participants' average years of experience in the field of human-computer-interaction was 6.9 (SD=6.4). For sample overview, we identified all participants with an experience of 10 or more years as experts [1], participants with one or less years as novices, and the remaining as intermediates, resulting in 36 novices, 34 intermediates, and 33 experts.

Material
We used the Soscisurvey platform to conduct our online study. We developed 32 problem cases that represented typical usability problems found in our own usability consulting projects. These consisted of a short problem title, a description and an example screenshot (cf. Figure 3).

Experimental Design
We employed a between subject design with one independent variable: scale type. Participants were randomly assigned to one scale type condition. Consequently, they were presented with problem cases along with one of the three scale types Nielsen (N), Practitioner's (P), or Individual Factor Scale (IF), with which they had to rate the severity of the problem. Problem case order was randomised. Two experts from science and practice with both over 20 years of experience in HCI were requested to establish a ground truth with all three scales for the 32 problem cases, which enabled us to investigate the scale effect on rating accuracy. The dependent variables were the case

Scale	N	Mean	SD
N	39	64.0	9.7
P	34	54.9	12.1
IF	30	55.8	7.8

Table 1. Overview of mean ratings per scale.

Scale	N	Mean	SD
N	39	28.1	6.7
P	34	39.3	8.8
IF	30	23.4	2.5

Table 2. Overview of mean deviations from ground truth ratings per scale.

Scale	N	Mean	SD
N	36	35.8	25.7
P	32	28.4	17.0
IF	25	46.8	23.5

Table 3. Overview of average case rating time per scale.

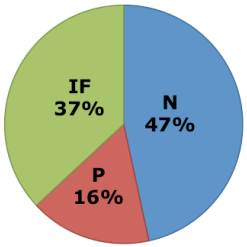

Figure 4. Overview of scale preference.

ratings, deviation from ground truth ratings, average case rating time, and scale preference.

Results

Ratings were treated as interval data, in line with Nielsen and general practical purposes [6, 8]. In order to standardise the scale ranges, the POMP-method was applied to obtain ratings from 0-100. In the following, we report on the effect of scale type on the ratings themselves, rating accuracy, the rating efficiency, and the scale preference.

Effect of Scale Type on Severity Ratings

We compared the average ratings of the 32 problem cases across the scale types (cf. Table 1). One-way independent ANOVA revealed a significant effect of scale type on severity ratings, $F(2, 100)=8.99$, $p<.001$. Tukey post-hoc tests showed that the Nielsen scale significantly produced higher ratings than the Practitioner's (p = .001) and the Individual Factors (p = .004) scales. In an effort to investigate our claim of pitfall 2 with the Nielsen scale—that is, that the option 'not a problem' is conceptually not part of the severity—we also ran the analysis with restricted range for the Nielsen scale (excluding 'not a problem' ratings) and obtained comparable, non-significant differences. This confirms our claim and argues for not including the 'not a problem' category in the severity scale as Nielsen suggested, especially on the notion that mean values in a report can lead to overestimation of problem severity on a quick glance.

Effect of Scale Type on Rating Accuracy

In order to establish a measure for rating accuracy we computed the average rating deviation from the ground truth rating of our two experts. Here, the IF scale offers the least deviation from the ground truth rating, while P scale differs the most (cf. Table 2). Due to violation of the homogeneity of variance assumption, non-parametric analyses were used. Kruskal-Wallis Test results confirmed a significant difference between the deviations, $H(2)=49.6$, $p<.001$. Mann-Whitney tests establish a significant difference between all the scale conditions, $p's<.001$. The results indicate that the IF performs best with regard to accuracy, while Nielsen is second best and P performs worst. The low standard deviation also suggests higher precision.

Effect of Scale Type on Rating Efficiency

Since the heuristic evaluation is a discount usability method that is quickly applied, the efficiency of how ratings can be established is of importance. While N and P scales request only one rating, participants have to perform seven ratings with the IF scale, which could defeat the purpose of the method of being low cost and quick to perform. Therefore, we investigated the average time participants needed to rate a case (cf. Table 3). Since in long online studies, participants are prone to take a break during a session, we cleaned the data of obvious outliers [4], resulting in the loss of 10 participants. Conform to expectations, P scale took the least time to rate a case, N scale was moderately fast, and IF scale was slowest. ANOVA indeed establishes a significant difference in average rating time, $F(2, 90)=4.66$, $p=.012$. However, Tukey post-hoc tests revealed only a significant difference between P and IF scale ($p=.009$). In the light of these results, it is astonishing that the IF scale, which formally requires seven rating processes in contrast to just one, only has an increased processing time of 30% compared to the N scale, and 64% compared to the P scale.

Preferred	N	P	IF
Novice	17	6	13
Intermediate	18	3	13
Expert	13	8	12
Total	48	17	38

Table 4. Overview of scale preference per experience group.

Preferred/ Assigned	N	P	IF
N	19	5	15
P	19	8	7
IF	10	4	16
Total	48	17	38

Table 5. Overview of preferred and assigned scale type.

Scale Preference

At the end of the online questionnaire, we gave participants a short description of all three scales, so that they could get an impression of the two scales they did not use during the study. We asked them, which of the scales they would prefer for their daily work (cf. Figure 4 and Table 4). Nearly half of the participants (47%) chose the Nielsen scale, while 37% chose the Individual Factor scale, and 16% the Practitioner's scale. It is noteworthy that 42% of the participants chose their assigned scale as favourite (cf. Table 5). Participants could give reasoning in a free text field for their preference decision. A lot of participants argued that the Nielsen scale is established, common, and easy to understand and apply. The Practitioner's scale was in general regarded as very quick to apply, but suffered from crude categories that did not fulfil the participants' need for finer grained judgments. The Individual Factor scale was regarded as very detailed and informative, but participants believed it to be time-consuming and at first hard to get into a rating flow, since the dimensions need to be understood by heart.

Conclusion

Our online study compared three different rating scales for the severity rating process of heuristic evaluation. We developed a new rating scale that improves on established concepts by incorporating principles from psychometrics and evaluated it against common rating scales. Our results showed that the applied scale has an influence on the ratings in terms of overestimation of severity with the Nielsen scale. In terms of rating accuracy, we observed that ratings obtained with the Individual Factor scale are the most accurate. The consistently lower standard deviations of the Individual Factor scale hint at less disagreement between evaluators. Rating efficiency is worst with the Individual Factor scale, but the trade-off between gained information and higher validity vs. 30% slower processing time might be well worth it. Nevertheless, the most preferred scale is the Nielsen scale; with the IF scale being the second favourite and the Practitioner's scale trailing behind.

The appeal of the heuristic evaluation is its cost-efficient and easy approach. It would be misdirected to overly complicate the method. However, adaptations to the severity scale can improve accuracy and information gain substantially, while introducing relatively small costs of processing time. In the light of our results, practitioners might consider more fine-grained and empirically valid approaches such as our Individual Factor scale for rating the severity of usability problems. Even though the established Nielsen scale fairs relatively well in our evaluation, the danger of overestimation is present.

Further research should continue investigating the multidimensionality of the severity construct. A sound framework could guide evaluators in finding important usability problems as well as improve severity ratings further. Moreover, the weighting of individual factors might differ for different product areas and purposes and should be investigated. Understanding the interplay between precision and accuracy in heuristic evaluation is a promising endeavour for research. As heuristic evaluation is widely applied, improving the method is likely to yield far spreading positive outcomes.

Acknowledgements

We thank the members of the Cooperative Media Lab in Bamberg and all the participants of the study.

References

1. K. Anders Ericsson. *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports and Games*. Lawrence Erlbaum, New Jersey, 1996.
2. Hershey H. Friedman and Taiwo Amoo. Rating the Rating Scales. *Journal of Marketing Management*, 9, 3 (1999), 114-123.
3. Morten Hertzum. Problem Prioritization in Usability Evaluation: From Severity Assessments towards Impact on Design. *International Journal of Human-Computer Interaction*, 21, 2 (2006), 125-146.
4. David C. Hoaglin and Boris Iglewicz. Fine Tuning some Resistant Rules for Outlier Labeling. *Journal of American Statistical Association*, 82 (1987), 1147-1149.
5. Irving L. Janis. *Victims of Groupthink*. Houghton Mifflin, New York, NY, 1972.
6. Thomas R. Knapp. Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy. *Nursing Research*, 39, 2 (1990), 121-123.
7. Rolf Molich, Jennifer McGinn and Nigel Bevan. You Say "Disaster", I Say "No Problem": Unusable Problem Rating Scales. Extended Abstracts of the *Conference on Human Factors in Computing Systems - CHI 2013* (Apr 27-May 2, Paris, France), ACM, N.Y (2013), 301-306.
8. Jakob Nielsen. Severity Ratings for Usability Problems (1995), (accessed 11 January 2016), https://http://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/.
9. Jakob Nielsen. Usability Inspection Methods. John Wiley & Sons, New York, NY, 1994.
10. Elazar Pedhazur and Leora Pedhazur Schmelkin. *Measurement, Design and Analysis*. Lawrence Erlbaum, Hillsdale, NJ, 1991.
11. Carolyn C. Preston and Andrew M. Colman. Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, 104 (2000), 1-15.
12. Mark Shevlin, Jeremy N. V. Miles and Brendan Bunting. Summated Rating Scales. A Monte Carlo Investigation of the Effects of Reliability and Collinearity in Regression Models. *Personality and Individual Differences*, 23, 4 (1997), 665–676.