Rater Bias: The Influence of Hedonic Quality on Usability Questionnaires

Stefanie Harbich and Sonja Auer

Siemens AG, CT IC 7, User Interface Design, Otto-Hahn-Ring 6, 81730 Munich, Germany
stefanie.harbich@dokusoft.de
sonja.auer@siemens.com

Abstract. In this study of various evaluation-instruments, subjects fulfilled several tasks on two different operating systems and answered several questionnaires, among them AttrakDiff™ and ISONORM 9241/10, and objective measures were taken. A correlation between the "hedonic quality - identity"-scale of the AttrakDiff™ and the ISONORM 9241/10 was found. As the ISONORM 9241/10 measures usability as described in ISO 9241-10 and not hedonic quality, the hedonic quality seems to have an influence on the tester ratings of usability. This is supported by the finding, that the hedonic quality does not correlate with the objective measures and therefore does not have any effective influence on the efficiency component of usability.

1 Introduction

Usability testing often relies on questionnaires, as they are known to be easy to handle, reliable, statistically objective, economical and easy to evaluate. They allow products to be compared and usability questions can be answered in an effective way. Though questionnaires have the advantage of being highly efficient and low in cost, they have disadvantages, too, like adulterations by biases in answering the checkmark items.

A well proven effect is the halo effect. After subjects judged one main aspect of the tested software, which had a quite big effect, they tend to judge all other aspects dependent on that main aspect. They seem to be unable to differentiate between different categories and therefore rate them all in an equal way [1]. This may negatively affect usability ratings, as one aspect of a software may easily cause a deep impression on the testers, which disables them to rate other aspects objectively.

Besides usability, an additional aspect in testing products is their appeal. Hassenzahl developed a questionnaire called AttrakDiff™ using a semantic differential with the four scales pragmatic quality, appeal, hedonic quality - stimulation and hedonic quality - identity [2]. Hedonic quality comprises the fulfillment of the need for novelty and change and the need to communicate and express oneself through objects [3]. "Hedonic quality - stimulation" means the human need for individual development, i.e. improvement of knowledge and skills. Another human need - identity - is the expression of the self through objects by identifying with them [2].

The concept of hedonic quality is different from the concept of usability, so these two should not correlate. But as this quality is an attribute of the tested software that

M.F. Costabile and F. Paternò (Eds.): INTERACT 2005, LNCS 3585, pp. 1129 – 1133, 2005.

needs to be rated on a very subjective basis, it may very easily bias the users perception of other attributes like the usability of a software. In this case, testers would not rate usability itself but would be influenced by the hedonic quality of the software.

Are participants influenced by the hedonic quality of a product when rating its usability? To test this, a usability test was conducted. Two questionnaires were employed, one measuring usability, one measuring hedonic quality. These subjective measures were complemented by a set of objective measurements, namely performance time, clicks and success in task performance. These objective methods operationalize effectiveness and efficiency as defined in ISO 9241-11 [4].

The questionnaire for measuring usability was the ISONORM 9241/10 questionnaire [5], which represents an operationalisation of the ISO 9241-10 [6]. As hedonic quality is not part of ISO 9241-10, this questionnaire is not supposed to measure anything similar to hedonic quality. Thus a high correlation between objective task performance measures and the ISONORM 9241/10 results should be expected, if the questionnaire measures were not influenced by hedonic quality.

Considering the possibility of an influence of hedonic quality on the ratings of a usability questionnaire because of the halo effect, it may be presumed that a higher correlation between the hedonic scales of the AttrakDiff™ and the ISONORM 9241/10 will occur than between the hedonic scales and the objective measures.

2 Method

32 clerks participated in this study, 30 female, 2 male. Their average age was 42 years, with a minimum of 24 and a maximum of 58 years, s = 9.3.

Participants were told they should test the two operating systems Windows XP Professional and SuSE Linux 9.2 with KDE 3.3. Every participant was currently working with Windows and had not previously worked with Linux. This ensured a predictable result to the effect that Windows would receive better ratings. Participants were given the same nine tasks for each system. Two observers sat behind a one-way mirror and watched the testers and their monitors by cameras and a scan converter. They rated the task achievement, stopped the needed time and counted the clicks.

Participants were instructed to work through the tasks and questionnaires listed in their testers' manual. Tasks were for example: "Save the attachment of this mail to [path] without renaming." or "Open the data browser and copy the file [filename] to [path]". After having completed the tasks participants answered the AttrakDiff™ and the ISONORM 9241/10.

Success was rated "1" (without errors), if the task was completed faultlessly, i.e. straightly without any mistake. "2" (noncritical errors) was assigned, if the participant completed the task successfully within seven minutes and without aid of the instructors. The observers rated success "3" (critical error), if the participants gave up or did not complete the task successfully within seven minutes.

In order to normalize the time and clicks, as they may depend on the specific hard- and software used, the times and clicks of four 'experts' were taken. These 'experts' knew the two systems and how to solve the tasks. Every expert's task was rated the

best achieveable value. The participants' data was divided by the experts' value to calculate a ratio of time and clicks.

Some participants gave up before they completed the task or quitted because they thought erroneously they did complete the task. As these time and clicks measures do not represent the real time and clicks, that would have been measured, if the tasks were completed, every value corresponding with a critical error ("3") was substituted by a missing.

3 Results

All measures used were able to distinguish between the two systems. This holds for the usability questionnaire and averaged objective usability measures as well as for averaged hedonic quality.

Table 1. Means and standard deviations for usability ratings of the two operating systems Windows and Linux

Method	Windows (n = 32)		Linux (n = 32)		total (n = 64)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Usability questionnaire	4.90 *	0.92	4.56	0.83	4.73	0.89
Hedonic quality - stimulation	4.07 **	0.87	4.59	0.95	4.33	0.94
Hedonic quality - identity	4.78	0.63	4.60	0.70	4.69	0.66
Success	1.74 **	0.29	2.06	0.21	1.90	0.30
Time	4.10	1.65	4.44	1.50	4.27	1.57
Clicks	1.70	0.45	1.63	0.53	1.66	0.49

Note. Usability questionnaire and hedonic quality scales based on Likert-type scale with "1" indicating strong disagreement and "7" indicating strong agreement. Windows' means with asterisks differ significantly from Linux' means.

* $p < .1$

** $p < .01$

Table 2. Comparison of correlations between usability questionnaire and hedonic qualities and correlations between hedonic qualities and objective measures

Method		Hedonic quality				Hedonic quality - identity				Hedonic quality - stimulation			
		Obj.	Success	Time	Clicks	Obj.	Success	Time	Clicks	Obj.	Success	Time	Clicks
	r	-.14	.10	-.02	-.27	-.14	-.02	-.06	-.22	-.15	.21	.03	-.31
Usability questionnaire	HQ .39	*	**	*	-								
	HQ-I .54					**	**	**	*				
	HQ-S .21									-	-	-	-

Note.
Obj.: Objective measures (mean of success, time and clicks)
HQ: Hedonic quality (mean of identity and stimulation)
HQ-I: Hedonic quality - identity
HQ-S: Hedonic quality - stimulation

**: $p < .01$
*: $p < .05$
-: $p > .05$

Table 3. Correlations between usability questionnaire and objective measures

Method	Usability questionnaire
Objective measures	-.02
Success	-.18
Time	.08
Clicks	.04

Correlations between the different methods support the rater bias hypothesis: Table 2 shows high correlations between usability and hedonic quality (mean of "hedonic quality - identity" and "hedonic quality - stimulation" after a Fishers-Z-transformation), r = .39, but low correlations between hedonic quality and object0ive measures, r = -.14. Surprisingly, as can be seen in Table 3, there is almost no correlation between the usability questionnaire and the (after a Fishers Z - transformation) averaged objective measures, r = -.02, and between the usability questionnaire and the single objective measures, i.e. success, r = -.18, time, r = .08, and clicks, r = .04.

Comparing the single correlations in Table 2, the correlation between usability and hedonic quality, r = .39 is significantly bigger than the correlation between hedonic quality and objective measures, r = -.14, $p < .05$ (one-tailed). This holds especially for the correlation between the objective measure success and hedonic quality, r = .10, versus the correlation between the usability questionnaire and hedonic quality, $p < .01$ (one-tailed), and also for the objective measure time, r = -.02, $p < .05$ (one-tailed), but not for the objective measure clicks, r = -.27, $p > .05$ (one-tailed).

As for the averaged hedonic quality, the correlation between "hedonic quality - identity" and the usability questionnaire, r = .54, also is significantly bigger than between "hedonic quality - identity" and averaged objective measures, r= -.14, p< .01 (one-tailed). This again is true for success, r = -.02, p<.01 (one-tailed), and time, r = -.06, $p < .01$ (one-tailed), and clicks, too, r = -.22, p<.05, (one-tailed).

These findings suggest an influence of hedonic quality (specifically the "hedonic quality - identity") on the usability questionnaire, but not on the objective measures. There is no evidence for such a difference in influence of "hedonic quality - stimulation" on the usability questionnaire compared to the objective measures. The correlation of "hedonic quality - stimulation" and the usability questionnaire, r = .21, does not significantly differ from the correlation of "hedonic quality - stimulation" and the averaged objective measures, r = -.15, or success, r = .21, or time, r = .03, or clicks, r = -.31, $p > .05$.

The results were analyzed separately for the two operating systems, too. Similar effects were found.

4 Discussion

The hedonic quality and usability questionnaires correlate to a big extent, whereas there seems to be no correlation between the usability questionnaire and objective measures, even though the ISONORM 9241/10 questionnaire and the objective measures indicate to measure usability and not hedonic quality. A possible explanation is the halo effect of hedonic quality on usability. Testers are influenced by the hedonic quality of a software and rate usability depending on their ratings of hedonic quality.

Hedonic quality means "hedonic quality - identity" here. The "hedonic quality - stimulation" does not have such a big effect. Participants rated usability higher, when

they identified highly with the tested software. When they were stimulated by the soft-ware, they rated usability higher, too, but to a lesser extent.

Another explanation may be that the objective instruments measure something different from usability, as the usability questionnaire and the objective measures do not correlate. This seems unplausible, as the objective measures operationalize the ISO 9241-11 - definition of efficiency and effectivity. Solely satisfaction was not covered by the objective measurements time, clicks and success. Maybe satisfaction influences both usability and hedonic quality in an extensive way. Of course, this would be a halo effect, too, as satisfaction then outshines other aspects of the software.

For the future, the influence of satisfaction on usability questionnaires *and* the hedonic quality should be analyzed.

Depending on the goals of testing, usability questionnaires should be used and analyzed carefully. Though they seem to rate usability in an objective way, other aspects may influence these ratings and give a false impression of the usability of the product.

References

1. Bortz, J. & Döring, N. (2002). Forschungsmethoden und Evaluation, Berlin: Springer.
2. Hassenzahl, M., Burmester, M.& Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In J. Ziegler & G. Szwillus (Eds.), Mensch & Computer 2003. Interaktion in Bewegung (pp. 187-196). Stuttgart, Leipzig: B.G. Teubner.
3. Hassenzahl, M., Beu, A. & Burmester, M. (2001). Engineering Joy. In: IEEE Software, 2-8, (pp 70-76).
4. ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. Brüssel: CEN.
5. Prümper, J. (1993). Software-Evaluation based upon ISO 9241 Part 10. In: T. Grechenig & M. Tscheligi (Eds.), Human Computer Interaction (pp 255-265), Berlin: Springer.
6. ISO 9241-10 (1996). Ergonomic requirements for office work with visual display terminals Part 10: Dialogue Principles. Brüssel: CEN.