From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments

Marc Hassenzahl
Darmstadt University of Technology
Institute of Psychology
Steubenplatz 12, 64293 Darmstadt, Germany
hassenzahl@psychologie.tu-darmstadt.de

Nina Sandweg
Siemens Corporate Technology
Competence Center User Interface Design
Otto-Hahn-Ring 6, 81730 München, Germany
nina.sandweg@siemens.com

ABSTRACT
In many cases, practitioners and researchers of Human-Computer Interaction and Usability Engineering rely on users' subjective product quality assessments. Such an assessment of, for example, perceived usability is believed to summarize previous experiences made with the according software. The present study shows that such summary assessments of perceived usability do not reflect a whole experiential episode, but rather its most recent incidents. Additional measurement strategies, such as repeated measurements of perceived usability throughout the experiential episode, are explored.

Author Keywords
user experience; perceived usability; evaluation; questionnaires

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces: Evaluation/methodology.

INTRODUCTION
In the past decade, researchers in the field of judgment and decision-making became increasingly interested in how individuals form retrospective summary assessments of experiential episodes (see [2] for an overview). Your last evening with friends, for example, may have started with a tasty dinner, followed by a parlor game and may have ended in a trendy bar close-by. The question "How much did you enjoy the evening?" requires you to integrate the joy (or pain) derived from the various single experiences made in the course of the evening into a single summary assessment. The most straightforward assumption is that individuals combine all the intensities of joy (or pain) experienced during the episode. In the example given above, the joy derived from the dinner may be 70 (on a 0-100 scale), 20 from the parlor game and 90 from the nightcap. The whole evening would then be a 60.

However, research showed that summary assessments do not correspond particularly well to the averaged experienced intensities (e.g., [4]). This inspired the study of how various other characteristics of an experiential episode, such as amount of variation in intensities or the trend of intensities (improving, deteriorating), relate to summary assessments.

The question of how experiences are transformed into summary assessments seems relevant to Human-Computer Interaction in general and product evaluation specifically. In many situations, questionnaires are used to assess quality aspects of a software product, such as its usability, or users' satisfaction. A common practice is to combine subjective measures of usability with a usability test. Usually, participants first work through a series of tasks. At the end of the series, they are asked to evaluate the software on the basis of the experiences just made. This requires the transformation of an experiential (usage) episode into a summary product quality assessment.

The present paper explores how the intensity of experiences relates to summary assessments of a software product's quality. Specifically, we study how the experience of mental effort while working with a software relates to summary assessments of its perceived usability.

METHOD
Twenty-one individuals (10 female, 11 male, median age = 41, min = 24 max = 57) participated in a usability test for a new version of CONNEXX (see [6]), a software tool for the configuration of hearing instruments. The tests were carried out in the UK (6) and Australia (15). The majority of participants were audiologists.

Each participant had to work through a series of typical tasks. Immediately after completing a task, mental effort (ME) was measured with the subjective mental effort questionnaire (SMEQ, [7]; see also [1]). The SMEQ is a single rating scale ranging from 0 to 150. Verbal anchors such as *hardly effortful* or *very effortful* facilitate the rating process. This procedure resulted in individual ME profiles consisting of seven[1] measurements for each participant (see

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

[1] Originally, nine measurements were taken. However, the first was excluded from the analysis, because the according "task" was only meant to be an icebreaker. Participants were instructed "to have a look around" without specification of a particular task goal.

Figure 1 for an example). On average, a single usability test session took about 2 hours.

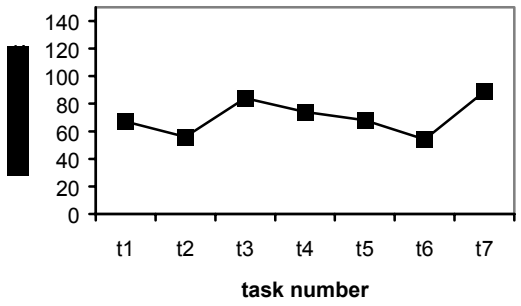

Figure 1: Example mental effort (ME) profile (participant 2).

Several predictors were derived from the seven ME measurements (see Table 1). The "intensity of the whole experiential episode" was captured by the average across the individual ME measurements, "variation in intensity" by their standard deviation. "End intensity" was represented by the ME measurement at t7 (see Figure 1). "Peak/End" was calculated as the proportion of the maximum ME to the ME at t7. In the example given in Figure 1 the proportion is 1, because end ME and maximum ME are the same. The proportion is smaller than 1, if the end intensity is less then the maximum. Generally this is a way to capture whether the experience became more positive over time or not. A second way to express this trend is to calculate a rank correlation between position in sequence and intensity of ME. A positive correlation expresses deterioration (i.e., mental effort increases over time). A negative correlation expresses improvement (i.e., mental effort decreases over time).

After having completed all tasks, participants were asked to assess the product's usability with the help of seven 7-point semantic differential items ("human – technical", "simple – complicated", "practical – impractical", "cumbersome – efficient", "predictable – unpredictable", "confusing – clear", "unruly – manageable", see [5] for an earlier application of the scale). A *summary perceived usability* (sPU) value was calculated for each participant by averaging across the seven single items (internal consistency was satisfactory, Cronbach's $\alpha = .85$).

In addition, two items were used to assess perceived usability parallel to mental effort. After measuring mental effort, participants were asked to assess the product on the attributes "unruly – manageable" and "cumbersome-efficient"[2]. They were further instructed to rate the product according to the most recent experience (i.e., task). Specifically, they were presented with their last rating and asked to revise it accordingly. An *integrated perceived usability* (iPU) value was calculated as the average across all single PU measurements over time (i.e., 2 items x 7 measurements).

RESULTS

Table 1 shows the correlation of the five features derived from the mental effort (ME) data with the summary (sPU) and integrated perceived usability (iPU) judgments. The data of two participants were removed from the analysis because they artificially inflated the correlations.

Feature	Summary PU	Integrated PU
Intensity of whole episode	-.42[m]	-.42[m]
Intensity of whole episode (end intensity excluded)	-.36	-.41[m]
Variations in intensity	-.24	-.25
End intensity	**-.49***	-.28
Peak/End	-.34	.09
Trend (improvement or deterioration)	-.38	**-.47***

Note: m) $p < 0.10$; *) $p < 0.05$; N= 9

Table 1: Correlations between various features of the experiential episode and the summary and integrated perceived usability (PU)

Summary PU correlated highest with the last ME measurement, i.e., the higher the intensity of mental effort at the end of the usage episode (t7), the lower was sPU. Integrated PU, however, correlated highest with the trend of the experiential episode. The more positive the rank correlation between ME measurements and their position in the episode, the lower was iPU. In other words, increased mental effort over time (i.e., deteriorating experience) led to lower iPU values, whereas less mental effort over time (i.e., improving experience) led to higher iPU values.

For both, summary and integrated PU, the correlation with the intensity of the whole episode was substantial and marginally significant. This seems to contradict the general finding that summary assessments do not capture the whole experiential episode particularly well. However, as long as ME at t7 (i.e., end intensity) is also a part of the mean ME (i.e., intensity of the whole episode) this finding is not astonishing. Indeed, the elimination of the ME measurement at t7 from the intensity of the whole episode

However, mental effort measurements are not meaningful without a particular goal. The last measurement was excluded, because participants were aware of the fact that it would be the last measurement. This knowledge may have prompted people to incorporate more than their immediate task experience into their rating.

[2] We did not use the full seven-item scale after each task in order to make the whole procedure less time consuming.

decreased the correlation with summary PU, whereas the correlation with integrated PU remained intact (see Table 1, row 3).

To better understand the way participants base their summary judgments of PU on features of the experiential episode, a stepwise regression analysis was performed. Summary PU was indeed best predicted by the end intensity of the experiential episode alone (adjusted $R^2 = .20$, $F = 5.50$, $p < 0.05$). The less intense mental effort at the end of the episode was the higher was the summary PU value. In contrast to summary PU, integrated PU was best predicted by the trend (improving, deteriorating) of the experiential episode (adjusted $R^2 = .18$, $F = 4.85$, $p < 0.05$). The stronger the deterioration of the experiential episode was, the lower was the integrated PU value.

To identify ways of measuring perceived usability that better reflect the average intensity of the entire experiential episode, one may determine the combination of available PU measurements, which predicts the average mental effort best. Predictors were integrated PU, summary PU and the rate of change in PU over time (i.e., standard deviation of PU measurements t1 to t7). The best model was a combination of summary PU and the rate of change in the single PU measurements (adjusted $R^2 = .38$, $F = 6.54$, $p < 0.01$). A low average mental effort implied a high summary PU ($\beta = -.43$, $t = 2.33$, $p < 0.05$) combined with a low rate of change in PU over time ($\beta = 0.52$, $t = 2.82$, $p < 0.05$).

DISCUSSION

The summary assessment of a product's perceived usability (PU) is primarily based on the end of the previous experience. This is most easily explained as a memory effect, a so-called *recency effect*. Individuals construct their summary assessment on the basis of what comes to their mind about the episode they just experienced. The more recent a detail, the more easily it comes into mind. Thus, the intensity of mental effort experienced towards the end of the episode tends to carry more weight in a subsequent summary assessment.

On the basis of the present results one may question the common practice of collecting summary usability assessments based on an experiential episode. However, this is not meant to imply that subjective usability ratings are redundant *per se*. They still reflect what an individual thinks about a product at a particular point in time and will thus influence related attitudes or even overt behavior (e.g., time spent with the product in the future). The challenge is to find additional measurement strategies that better reflect essential features of the entire experiential episode.

One such additional strategy is to repeatedly measure usability during the experiential episode. Indeed, the integrated perceived usability value (i.e., the mean of PU measurements t1 to t7) is best captured by the trend (improving, deteriorating) of the experiential episode. As long as the trend is a central feature of the *entire* episode, integrated PU might be considered as "better" than summary PU. It does not capture the averaged (or summed) experienced intensities, but it reflects at least an important aspect of an experience, namely whether using the software becomes less or more demanding over time.

If the goal of a quality measurement is to capture the averaged (or summed) experienced intensities, a mixed measurement strategy seems the most promising, namely the summary assessment of PU combined with an indicator of the rate of change of PU. However, the last component has to be treated carefully. The explanatory power of rate of change in PU over time might be due to its specific way of measurement. In the present study, participants were asked to revise their former PU assessment on the basis of the new experiences gained while working through the most recent task. By that, participants were encouraged to rather think about *changes* in the software's usability over time than to make independent assessments solely based on the experiences made in each single task. Future studies should explore other ways to assess perceived usability directly while experiencing the software. Those simple measurements and derived indicators could even be an alternative to the regularly recommended use of objective data (e.g., number of errors, problem handling time). Objective data is often hard to obtain in industrial settings because of time and budget restrictions or conflicts between the reliable measurement of, for example, task execution time and the desire to collect qualitative, design-relevant data.

There is one limitation of the present study to be discussed in more detail. Mental effort may not be a variable rich enough to represent an "experience" at all. All in all, the percentage of variance in summary and integrated PU that was explained by the way mental effort was experienced over time was rather low. This implies that additional experiential or even non-experiential aspects may play an important role in judgments of perceived usability, such as stimulation or prior judgments of similar or other products that serve as standards against which a new product is contrasted.

To conclude, the question of how experiences are integrated into retrospective, summary assessments of those experiences and further translated into summary assessments of a software's perceived usability may prove helpful in better understanding the determinants of a quality experience. Research on preferences for sequences of outcomes, for example, repeatedly demonstrated that an improving sequence of otherwise identical outcomes is preferred to a deteriorating one (e.g., [5]). Although the actual outcomes are identical, improvement adds extra value to a sequence. Similar aspects may be crucial in the context of Human-Computer Interaction. Perceived usability, for example, may rather be a consequence of whether handling becomes less difficult over time than a consequence of level of difficulty *per se*.

In addition, the present research prompts awareness of the problem of memory effects in subjective judgments of usability. Practitioners as well as researchers only rarely attempt to evaluate products by measuring how judgments develop over time. Most of the time they rely on the participants' ability to integrate experiences into summary assessments. This might not be the most promising strategy.

ACKNOWLEDGMENTS

We like to thank Eduard Kaiser, Siemens Hearing Instruments, and Heinz Bergmeier, Siemens Corporate Technology, for their support.

REFERENCES

1. Arnold, A. G. Mental effort and evaluation of user interfaces: a questionnaire approach. In H.-J. Bullinger and J. Ziegler (Eds.), *Proceedings of the HCII '99 international conference on human-computer interaction, vol. 1*. Mahwah, NJ, Lawrence Erlbaum (1999), 1003-1007.
2. Ariely, D and Carmon, Z. Summary assessment of experiences: The whole is different from the sum of its parts. In G. Loewenstein, D. Read, and R.F. Baumeister (Eds.), *Time and decision. Economic and psychological perspectives on intertemporal choice*. New York, Russell Sage, 2003.
3. Hassenzahl, M. The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction, 13* (2002), 479-497.
4. Kahneman, D. Evaluation by moments. Past and future. In Kahneman, D., Tversky, A. (Eds.), *Choices, values and frames*. New York, Cambridge University Press, 2000.
5. Ross, W. T. and Simonson, I. Evaluations of pairs of experiences: A preference for happy endings. *Journal of Behavioral Decision Making, 4* (1991), 273-282.
6. Sandweg, N., Pedell, S., Platz, A., Schneider, K.-P., Honold, P., Hermann, D., and Kaiser, E. Re-design of CONNEXX hearing instruments fitting software. In H. Luczak, A. E. Cakir, and G. Cakir (Eds.), *Proceedings of the 6th international conference on Work With Display Units.* Berlin, ERGONOMIC Institut für Arbeits- und Sozialforschung (2002), 275-276.
7. Zijlstra, R. Efficiency in work behaviour. *A design approach for modern tools*. Delft, Delft University Press, 1993.