VIDEO-RATING: A RELIABLE AND VALIDE EVALUATION METHOD FOR THE MAN-COMPUTER-INTERACTION (MCI).

Matthias RAUTERBERG
Department of Computer Science, FB 10
University of Oldenburg, P.O.Box 2503
D-2900 Oldenburg, West-Germany

Keywords: video rating evaluation method man human computer interaction human machine communication psychomental stress software ergonomy interface design

INTRODUCTION

Important for the design of interactive software is a good measure for psycho-mental stress. Howard and Murray (1987) indentified the following main types of evaluation: expert based, theory based, subject based, user based, market based. The subject-based evaluation technique collects data on the three levels of the user: physiological, behavioural and at least the cognitive/affective level.

This work is a methodical study and explicates a video-rating method to explore and measure the observable and spontaneous behaviour during interactive programming.

To measure psycho-mental stress with questionnaires at special break-points is in contradiction to a continuous online non-reactive measurement. On the other side: for the online measurement of e.g. the pupil diameter or the electro-encephalogramm as an indicator for the emotinal quality of the information processing one needs an artficial experimental setting and an expensive registration set. In contrast to this the video recording procedure is easy to handle and allows a continuous measurement of psycho-mental stress, too.

METHOD

The sample was composed of ten male FORTRAN-programmers (with age between 18 and 30 years) working at the University of Hamburg (Rauterberg, 1981).

The experiment was divided into three phases: instruction phase (30 minutes), MCI: problem-solving phase (60-90 minutes), interview and explanation phase (20 minutes).

The goal of this work is the construction and test of a video-rating evaluation method. Therefore a stressful situation has to be defined. The following three known aspects (Köchling, 1985) guarantee that a quite stressful experimental setting was constructed:

time pressure: the programmers had only 60 minutes to solve the problem without losing their reward of 30.-DM[1]; after this time the reward was reduced continuously; the deadline for time of problem solving was after 90 minutes;

negative feedback: after every solution trial the time left for completion and a remark about the faulty programme was shown on display;

high mental load: the programmers had to debug the iterative algorithm of the complex problem: "The Tower of Hanoi"; this algorithm was incorrectly programmed as a FORTRAN-subroutine.

Later the video tapes were rated by eleven students of psychology after a 2-hour instruction phase. To avoid uncertainty about the semantics of the 15 bipolar rating scales (see table 1) each rater could look all the time on a description of all scales. The whole video-rating phase extended over three hours on three forenoons. So this video-rating method needs about 100 workinghours.

[1]) This research was supported by Grant from "Gesellschaft zur Förderung der angewandten Psychologie" in Hamburg.

In: A. S. Adams, R.R. Hall, B. J. McPhee & M. S. Oxenburgh (eds.), Proceedings of the Tenth Congress of the International Ergonomics Association (Vol. II, pp. 633-635). Sydney (AUS): IEA Press

Table 1. A bipolar scale of the 15 rating scales composing the rating sheet.

concentrated	-3..-2..-1..0..+1..+2..+3	not concentrated

The rating-sheet consisted in 25 different permutations of the 15 bipolar scales.

The videorecord of each programmer was truncated into three five-minutes sections (at the beginning: time-1, middle: time-2, and end: time-3 of the MCI-phase). These sections were accidentally presented to the raters. To estimate the retest reliability five sections were presented twice.

RELIABILITY

To estimate the reliability coefficients a two-factorial variance-analysis with complete repented measurements was computed (Winer,1971). To estimate the retest-reliability (REL-RETEST) the ratings of the double video-presentations were correlated.

Three reliability coefficients were computed: REL-HORST, REL-EBEL, REL-RETEST (Langer and Schulz von Thun, 1976).

The different sources of object variance for REL-EBEL and REL-HORST are:

- VP = variance between programmers;
- VM = variance between repeated measurements;
- IPM = interaction between programmers and time.

The error variances are:

- VR = variance between raters;
- IPR = interaction between programmers and raters;
- IRM = interaction between raters and time;
- IPRM = interaction between programmers, raters, and time.

The reliability coefficients were computed as follows:

$$\text{REL-EBEL} = 1 - \frac{(IPR+IRM+IPRM)}{(VP+VM+IPM)} \quad (1)$$

$$\text{REL-HORST} = 1 - \frac{(VR+IPR+IRM+IPRM)}{(VP+VM+IPM)} \quad (2)$$

Table 2. The arithmetical means of all rating scales for the three five-minutes sections (beginning: time-1; middle: time-2; end: time-3) of the MCI-phase; the results of the variance analysis and the 3 different reliability-coefficients.

no.	Rating-scales (-)	(+)	means time-1	time-2	time-3	V.A.-factor: Q.S.	df	TIME F	alpha	REL-HORST	REL-EBEL	REL-RETEST
1	silent	-talkative	-0.02	+0.60	-0.79	105.7	2	3.18		.95	.98	.97
2	bored	-engaged	+1.35	+1.49	-0.33	224.1	2	11.63	1 %	.88	.94	.86
3	tranquil	-active	+0.19	+1.09	+0.01	74.2	2	2.60		.88	.91	.85
4	noisy	-quiet	+0.07	-0.32	+0.67	54.3	2	2.75		.85	.95	.99
5	concentrated	-not concentrated	-1.34	-0.67	+0.25	138.7	2	3.66	5 %	.85	.95	.93
6	disappointed	-assured	+0.23	-0.32	-1.22	116.0	2	5.57	5 %	.84	.95	.96
7	slow	-quick	+0.40	+0.39	-0.60	71.6	2	3.27		.83	.94	.84
8	patient	-impatient	-0.02	+0.84	+0.35	40.6	2	1.62		.80	.92	.98
9	irritated	-peaceful	+0.01	-0.43	-0.03	13.0	2	0.46		.72	.89	.95
10	restricted	-open-minded	+0.65	+0.91	+0.27	22.8	2	2.68		.65	.82	.50
11	anxious	-courageous	+0.51	+1.01	+0.50	18.6	2	1.50		.59	.82	.62
12	uncertain	-self-assertive	+0.16	+0.11	-0.47	25.9	2	1.35		.52	.81	.49
13	earnest	-cheerful	-1.20	-1.12	-1.47	7.4	2	1.76		.32	.64	.77
14	tense	-relaxed	-0.95	-1.26	-0.87	9.2	2	0.91	*	.27	.82	.99
15	attractive	-unattractive	-0.35	-0.27	-0.07	4.5	2	0.68	*	.15	.59	.60

The scales marked with a star (*) in table 2 have coefficients REL-HORST less than 0.30, and were excluded from further analysis.

VALIDITY

At the beginning and at the end of the MCI-phase the programmers had to answer the ZERSSEN-Subjective-Feeling-Scale (ZSFS). After each compilation trial they had to

In: A. S. Adams, R.R. Hall, B. J. McPhee & M. S. Oxenburgh (eds.), Proceedings of the Tenth Congress of the International Ergonomics Association (Vol. II, pp. 633-635). Sydney (AUS): IEA Press

respond online to a bipolar selfrating scale about their actually subjective-emotional feeling (MCI-SF; 0='very unpleasant'; 9='very pleasant').

The highly significant difference (X_{pre}=9.4, sd=11.4; X_{post}=2.3, sd=11.8; p<0.5%) of the pre-post comparision in ZSFS supports the assumption that this experimental setting guarantees a high content validity.

The values in MCI-SF also show a high significant decrease ($X_{1.min}$=5.1, sd=2.5; $X_{60.min}$=3.2, sd=3.3; p<0.5%) of "pleasant feeling" from 1. to 60. minute of the MCI-phase.

The external validity can be estimated by the following question given in the explanation phase: "To what extent is this experimental situation equivalent to your daily work situation?" (scale-range: 0% - 100%). Nearly half of all programmers estimated this correspondence at 80%.

To determine the internal validity a hierarchical cluster-analysis was computed over all measured scales (video-rating, personality-questionnaires, subjectiv-feeling scales, self-rating scales, etc.). Six dimensions are found to be relevant (see table 3).

Table 3. The six dimensions of the rating sheet.

psycho-mental stress	(rating-scales: 3, 5, 8, 9, 13;	$R_{average\text{-}correlation}$= 0.80);
psycho-emotional stress	(rating-scales: 6, 12;	$R_{average\text{-}correlation}$= 0.70);
emotional stability	(rating-scales: 10, 11;	$R_{average\text{-}correlation}$= 0.63);
engagement	(rating-scale : 2;	$R_{average\text{-}correlation}$= 0.59);
interactivity	(rating-scale : 7;	$R_{average\text{-}correlation}$= 0.67);
verbal-behaviour	(rating-scales: 1, 4;	$R_{average\text{-}correlation}$= 0.71).

CONCLUSION

The main subject of this work is the construction and test of an evaluation method. Therefore the means of the stress scores of this methodical study attained in this experimental setting are not interpretable as real data of MCI work conditions. In order to determine the latter this video-rating evaluation method ought to be applied to real MCI work conditions. But the results of this methodical study show that this video-rating evaluation method is able to measure the psycho-mental stress in real life situations, e.g. at MCI work places with high reliability and sufficient validity.

This video-rating evaluation method for measuring psycho-mental stress is not only restricted to the field of MCI. It is applicable to many other research fields, too.

REFERENCES

Howard, S. and Murray, D.M., 1987, A taxonomy of evaluation techniques for HCI. In Human Computer Interaction - INTERACT '87, (Edited by H.-J. Bullinger and B.Shackel) (IFIP) (North-Holland: Elsevier Science Publishers), pp 453-459.

Köchling, A., 1985, Bildschirmarbeit - Gesundheitsregeln und Gesundheitsschutz, (Köln: Bund-Verlag).

Langer, I. and Schulz von Thun, F., 1974, Messung komplexer Merkmale in der Psychologie und Pädagogik, (München Basel: Ernst Reinhardt).

Rauterberg, M., 1981, Psychomentale Belastung in der Mensch-Computer-Interaktion, (Diploma thesis)(Hamburg: Fachbereich Psychologie).

Winer, B.J., 1971, Statistical principles in experimental design, (New York: McGraw Hill).

In: A. S. Adams, R.R. Hall, B. J. McPhee & M. S. Oxenburgh (eds.), Proceedings of the Tenth Congress of the International Ergonomics Association (Vol. II, pp. 633-635). Sydney (AUS): IEA Press

Designing a Better World

INTERNATIONAL ERGONOMICS ASSOCIATION

10TH INTERNATIONAL CONGRESS

SYDNEY, AUSTRALIA 1-5 AUGUST 1988

PROCEEDINGS Vol. II

In: A. S. Adams, R.R. Hall, B. J. McPhee & M. S. Oxenburgh (eds.), Proceedings of the Tenth Congress of the International Ergonomics Association (Vol. II, pp. 633-635). Sydney (AUS): IEA Press

In: A. S. Adams, R.R. Hall, B. J. McPhee & M. S. Oxenburgh (eds.), Proceedings of the Tenth Congress of the International Ergonomics Association (Vol. II, pp. 633-635). Sydney (AUS): IEA Press