

A Survey of Interface Goodness Measures

HITL Technical Report R-94-1

Jerry Prothero
Human Interface Technology Laboratory
prothero@hitl.washington.edu

March 16, 1994

Abstract

The development of interface design as an engineering discipline has been hampered by the absence of general, robust, and quantitative measures for the usefulness of interfaces. In the absence of such measures, it is difficult to optimize an interface for a particular task, and to make precise statements about the relative advantages of different types of designs. This article briefly surveys the literature on interface goodness measures.

1 Introduction

A chronic problem in the field of interface design is difficulty in measuring how well an interface performs. In the absence of such measures, the development of good interfaces tends to be an art, with little assurance when one is done that the final product is the best possible interface for the given task.

The evolution of interface design from an art to engineering would seem to depend on the improvement of the measures used to test interfaces. This is important not only for the testing of individual interfaces, but also for the development of reliable design principles and an understanding of the range of their applicability.

Rather than trying to solve the measures problem, this article is instead a survey of the state-of-the-art.

As discussed below, measures can be based on either performance or subjective criteria. Performance measures are potentially more objective and precise, but face serious difficulties. In the context of automated human-system performance measurements in the field of training simulators, Vreuls and Obermayer [12] (1985) state that “most existing automated training performance measurement systems are so poorly designed that they are useless.” They list four types of problems with performance measurement in simulation, which seem to have some relevance to evaluating interfaces in general.

Hidden Knowledge and Embedded Performance “In most simulators the purpose of performance measurement is to infer something about the knowledge, skills, or decision processes of the human participant. Performance measures, however, depend on overt actions; internal processes that produce overt actions are not directly observable. Moreover, the results of complex processes, such as decision making, may be manifest only by simple actions or no actions at all.”

Lack of Theories of Performance “The second fundamental measurement problem is the lack of unifying theories of human performance to predict actions over a wide range of circumstances. Formal theories and models tend to define what must be measured, the relative importance of each measure, and interactions with other measures under given circumstances. In the absence of theories to guide selection of performance measures, one is driven to the alternative of measuring as much as is reasonably possible.”

Measurement Validity “Indirect measures of human performance require empirical tests to determine their validity for the intended purpose. Measures for simulator training often are derived from analytical studies and selected for use on the basis of perceived suitability; the construct, concurrent, content, and predictive validity of measures are seldom tested.”

Operational Performance Criteria “The final basic problem in simulator performance measurement is the lack of quantitative criteria for assessing the importance of performance changes. Performance changes can be measured in a simulator during training, but only infrequently can such performance changes be related to overall system or mission effectiveness.”

2 Methodologies

A number of general approaches to evaluating interfaces have been developed. Among these are usability testing, heuristic evaluation, and cognitive walkthroughs [3, 7]. Usability testing is a methodology in which real users are observed performing real tasks, and conclusions are reached based on what the users do and say. Heuristic evaluation involves having a group of interface evaluators examine an interface and look for violations of interface design principles. Cognitive walkthroughs consist of answering a set of questions about each of the decisions which an interface user must make, and rating the likelihood that the user will make an incorrect choice. Cognitive walkthroughs are intended for cases in which building a full prototype of the interface is infeasible.

3 Measures

Within the general approach taken to evaluating an interface, many types of measures may be used. The measures can be divided broadly into subjective measures, such as asking users to write down their impressions about some aspect of the interface, and performance measures, based on quantities such as response time and accuracy. The relative advantage of subjective measures is that they may be able to address more general or “cognitive” issues. The relative advantage of performance measures is that they may be more objective, and may provide the precision necessary to “fine-tune” an interface, and to make precise statements about the relative advantages of different approaches.

The subjective measures listed by Dumas and Redish [3] include: ratings of ease of learning, ease of using, etc.; preferences and reasons for preferences; predictions of behavior and reasons for the prediction (e.g., “would you buy this product?”); and spontaneous comments.

Among the performance measures listed by Dumas and Redish [3] are time to finish a task, time spent navigating menus, time spent in the online help, number of wrong menu choices, observations of frustration, etc.

Shneiderman [11] lists five measures crucial to interface evaluation: time to learn; speed of performance; rate of errors by users; retention over time; and subjective satisfaction.

Since perhaps the best-studied interfaces are those used in aviation, most of this section draws from the aviation literature. Most of the measures discussed are aimed at evaluating components of the pilot’s cognitive state, such as mental load or situational awareness, rather than the interface directly.

Bortolussi *et al.* [1] tested four widely used methods for measuring pilot workload in a simulator. These include a visual 2 and 4 choice reaction time task, time production, retrospective multi-dimensional subjective ratings, and in-flight verbal workload estimates. It is concluded that “all four techniques are capable of distinguishing between the overall levels of scenario complexity.”

Klix *et al.* [6] measure mental load using biological values, such as CNV-waves, evoked potentials, pupillography, and heart rate measurements.

Endsley [4] measures situational awareness for pilots in simulators by stopping the mission at random points and asking both directly relevant and secondary questions.

Sanders [8] points out that “the issue of performance measurement has been severely neglected and ignored in many applications of simulators.” For perceptual-motor measures, Sanders distinguishes between traditional measurement more-or-less based on speed and accuracy, and newer measurements. Examples of the latter include visual occlusion as a technique for measuring perceptual load for car-driving simulations, and dynamic windowing to study reading.

Selcon and Taylor [9] postulate that the three primary components of situational awareness are attentional demand, attentional supply, and understanding. These can be broken into ten subcategories. Under demand, instability of the

situation, variability of the situation, and complexity of the situation; under supply, arousal, spare mental capacity, concentration, and division of attention; and beneath understanding, information quantity, information quality, and familiarity. Subjective measures are used for these variables.

4 Noise Immunity

Furness [5] has suggested that the extent to which performance degrades as noise is added to an interface can be used to examine how effectively the interface conveys a mental model to the user. The idea is that if the interface has given the user an effective mental model, then the user should be able to deduce missing information resulting from noise added to the interface. This idea is intriguing because studying mental models is such a difficult and important problem.

A possible problem is that noise immunity as a measure of the conveyance of mental models may be confounded with the redundancy of information in the interface. If we are trying to determine which of two interfaces conveys a better mental model, the one which comes out better on the noise immunity measure may simply be the one with more repeated information. Since it is difficult to control for redundancy in interfaces, this may be a fundamental problem with this measure.

5 An Entropic Performance Measure

Viewed abstractly, the function of an interface is to convey information (bidirectionally) between a computer and a user. This section discusses a measure which, in principle, is a direct quantitative gauge of the flow of information between the computer and the user, and which is generalizable across tasks [13] (pp. 49–58).

The entropic measure is a trivial application of information theory [2]. (A brief introduction to information theory is given in Appendix A.) As formulated here, the idea is that for some interfaces, the entropy (roughly, a measure of complexity) of both the information flowing into the display and of the user's response to the display are well-defined. It is postulated that for a good interface the input entropy (of the information flowing into the display) and the output entropy (of the user's response to the display) will be strongly correlated. In other words, if more information becomes available, the user's response is expected to become more sophisticated if the interface is performing well.

In some ways, this is a very nice measure. Instead of simply judging human performance, as with most measures, human performance is correlated with input complexity, expressed in the same units. Furthermore, the entropic measure (unlike many of the other measures) is “non-intrusive”: the measure can be computed while the user is performing real work, without imposing distractions.

A measure of this type could potentially be used to optimize the interface to the user while the user continues to work.

However, this measure has difficulty accounting for errors, or the relative importance of errors. Furthermore, it is a “syntactic” measure: it measures observable data and actions, but says nothing about the related mental models or their effectiveness.

In addition, there are practical problems with the definition of the data input and human response output probabilities. For the entropies to be computable, it must be possible to view the information coming into the display as a set of events drawn from a known static or slowly changing probability distribution, and the same for the user’s responses to the interface. This implies that both the input and the user’s output must consist of a small or stereotyped set of possibilities occurring frequently, so that meaningful probabilities are computable.

For example, the entropy measure would not be definable if the input consists of an arbitrary picture of an environment containing many parameters which don’t repeat themselves over time. It would be definable, for instance, for a display representing sets of data taken from small ranges, such as an altimeter and a fuel gauge. Similarly, the entropic measure is not definable if the output is an essay in English, but would be definable if the output is button clicks or motions of a joystick.

It is possible that the entropic measure would be useful for “mixed displays”, which combined some of the features of both types of input and output.

Yet another problem with the entropic measure is that to be usable there must be a strong and immediate correlation between changes in the input information and the user’s response. Interfaces which emphasize perceptual-motor skills (for instance a user interface for aircraft pilots) are potential targets for this measure. Interfaces which emphasize cognitive skills (for instance a statistics package, in which the emphasis is on reaching an understanding rather than on making a rapid decision) are not potential targets for this measure.

A Information Theory

Information theory developed from Shannon’s work in the late 1940’s on communication theory [10]. The theory was originally applied to determining the degree to which data could be compressed, and to the ultimate data transmission rate for a given channel. It has since found applications in computer science (Kolmogorov Complexity), physics (thermodynamics)¹, mathematics (probability theory and statistics), economics (investment) [2], and in human performance theory [13]. Wolff [14] has pointed out that many problems in cognition and computer science can be looked at as information compression.

¹The definition of entropy given in information theory is closely related to the definition given in physics, although the relationship won’t be explored in this article. See [2] (p. 33).

Information can be defined as the reduction of uncertainty. Events which occur with high probability convey little information, since they do little to change one's knowledge of the world. Conversely, events with low probability are very informative. Formally, each of a set of events X_i which occur with probabilities $P(X_i)$ is said to have information content

$$I = \log_2(1/P(X_i))$$

Intuitively, the reason for the $1/P(X_i)$ is that less probable events should carry more information than more probable events. In essence, the log converts from number of events to number of bits needed to encode events.

The entropy $H(X)$ of a set of n events X is defined as the average information content:

$$H(X) = \sum_{i=1}^n P(X_i) \log_2(1/P(X_i))$$

The entropy gives the average number of bits necessary to encode the events in X . For instance, if X consists of the events $A, B, C,$ and $D,$ with probabilities $.5, .25, .125,$ and $.125,$ then the entropy is

$$\begin{aligned} H(X) &= .5 \log_2(1/.5) + .25 \log_2(1/.25) + .125 \log_2(1/.125) + .125 \log_2(1/.125) \\ &= .5 + .5 + .375 + .375 \\ &= 1.75 \text{ bits} \end{aligned}$$

(The maximum entropy for a given number of events occurs when all of the events occur with equal probability. In the case of four events, the entropy would be 2.0.)

The entropy of two random variables X and Y considered together is given by

$$H(X, Y) = \sum_{i=1}^n P(X_i, Y_i) \log_2(1/P(X_i, Y_i))$$

where $P(X_i, Y_i)$ is the probability of X_i and Y_i occurring together. Similarly for more than two random variables.

If X and Y are uncorrelated, then $H(X, Y) = H(X) + H(Y)$. If knowledge of X provides information about Y , then $H(X, Y) = H(X) + H(Y|X)$. $H(Y|X)$ is called the conditional entropy of Y given X , and is defined as

$$H(Y|X) = \sum_{i=1}^n P(X_i) \sum_{j=1}^n P(Y_j|X_i) \log_2(1/P(Y_j|X_i))$$

References

- [1] M. Bortolussi, B. Kantowitz, and S. Hart. Measuring pilot workload in a motion base trainer. *Applied Ergonomics*, pages 278-283, Dec 1986.

- [2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.
- [3] Joseph S. Dumas and Janice C. Redish. *A Practical to Usability Testing*. Ablex Publishing Corporation, 1993.
- [4] M. Endsley. A methodology for the objective measurement of pilot situation awareness. In *AGARD Conference Proceedings No. 478 Situational Awareness in Aerospace Operations*, pages 1–9, Oct 1989.
- [5] Thomas A. Furness. *Communicating Situation Awareness in Virtual Environments: A Multidisciplinary Proposal Submitted to the Air Force Office of Scientific Research*. Human Interface Technology Laboratory, 1992.
- [6] F. Klix, B. Krause, and R. Hagedorf. Psychological problems concerning the lay-out of human-computer interaction: A challenge to research in cognitive psychology. In *Man-Computer Interaction Research MACINTER-II*, pages 3–29. Elsevier Science Publishers, 1989.
- [7] J. Nielsen and R.L. Mack, editors. *Usability Inspection Methods*. Wiley, New York, 1994.
- [8] A. Sanders. Simulation as a tool in the measurement of human performance. In *Contemporary Ergonomics 1990: Proceedings of the Ergonomics Society 1990 Annual Conference*. Taylor and Francis Ltd., 1990.
- [9] S. Selcon and R. Taylor. Evaluation of the situational awareness rating technique (sart) as a tool for aircrew systems design. In *AGARD Conference Proceedings No. 478 - Situational Awareness in Aerospace Operations*, pages 5:1–5:19, 1989.
- [10] C.E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423,623–656, 1948.
- [11] Ben Shneiderman. *Designing the User Interface*. Addison Wesley, second edition, 1992.
- [12] D. Vreuls and R.W. Obermayer. Human system performance measurement in training simulators. *Human Factors*, 27:241–250, 1985.
- [13] Christopher D. Wickens. *Engineering Psychology and Human Performance*. Harper Collins, second edition, 1992.
- [14] J.G. Wolff. Computing, cognition and information compression. *AI Communications*, 6:107–127, June 1993.