Metrics, Metrics, Metrics: Negative Hedonicity

Robert R. Hoffman and Morris Marx, *Institute for Human and Machine Cognition*
Peter Hancock, *University of Central Florida*

Intelligent technologies such as performance support systems and decision aids represent a key aspect of modern sociotechnical systems. When new tools are introduced into the workplace, they represent hypotheses about how cognitive work is expected to change.[1,2] The tacit hypothesis is that any such change will be for the better, performance will be more efficient, and decisions will be improved—that is, they'll be made faster and on the basis of greater evidence. Experience suggests that technological interventions sometimes have the intended positive effect. However, they often result in negative effects, including unintended cascading failures and worker frustration due to "user-hostile" aspects of interfaces.[3]

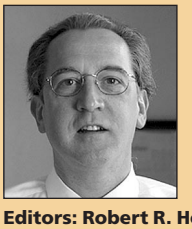
Concern is rising about the high rate of software procurement failures that are due to the inadequate consideration of human factors. Recent statistics suggest a dismal record, representing the expenditure of billions of dollars for technologies that are unusable, ineffective, and at times even defunct.[4,5] At the same time, funding for developing communication and information technologies has reached record levels (about US$500 billion in the mid-1990s).[6] Recently we've seen entire government-sponsored research programs with titles that state human–system integration as a key goal for new technologies.[7] The notorious frustrations and failures triggered by software interventions in the workplace have led to a significant concern in the software engineering community with evaluation,[8,9] including help for organizations to establish metrics for "key performance indicators."[10,11]

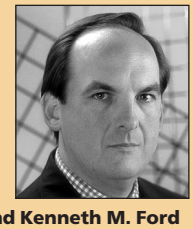
The call for "metrics"

Nearly all announcements of US government-funded research programs for developing large-scale information systems have shown a pervasive concern with "metrics." The following three paraphrased statements from recent program announcements illustrate this point:

- "[The program will] explore methodologies and technologies which achieve substantial improvement and cost reduction in software development, requirements analysis and definition, software management, complexity, and quality metrics, reuse, reengineering, maintenance."
- "Metrics are needed to determine the correct fidelity for attaining training objectives while operating within the boundaries of current technologies, human perception, schedule and cost."
- "Multidisciplinary and cross-domain approaches are highly encouraged especially if useful in the development of metrics for dynamics, complexity and usability."

The sought-for measures have to gauge efficiency, effort, accuracy, and similar reflections of a maximizing process, hearkening to John Henry versus the steam hammer.[12] This is particularly frustrating for the advocates of human-centered computing and work-centered design.[13,14] A valiant effort to think along human-centering lines was a recent report of the US National Institute for Standards and Testing.[15] This highlighted measures such as efficiency but did so with reference to dimensions including effectiveness at hypothesis generation, effectiveness at coping with massive data, and effectiveness of human-





Editors: Robert R. Hoffman, Patrick J. Hayes, and Kenneth M. Ford
Institute for Human and Machine Cognition
rhoffman@ihmc.us

1541-1672/08/$25.00 © 2008 IEEE
Published by the IEEE Computer Society

machine interaction. Measures also included confidence ratings and assessments of associated mental workload. However, in light of our knowledge of human adaptability, can we provide better methods and procedures to create measures that reflect the meaningful aspects of systems-level cognitive work and activity?

The challenge

We must take the measurement of cognitive work to entirely new levels—addressing, for example, the important trade-offs in cognitive work at the team and systems level. The latter is, after all, where we might realize the final payoff for any investment.

A significant challenge is that studies of human-computer interaction, and the measures that we take, must support the evaluation of hypotheses concerning the nature of cognitive work itself (for example, the effects of synchronous versus asynchronous communication in distance collaboration, effects due to amount of team experience, and so on). At the same time, the study design must support evaluation of the software tools themselves. In other words, new technologies must do double duty: they enable research on cognitive work by supporting cognitive work, including new work methods. We might think of them as part of the materials and procedure comprising the method of an experiment on human-machine interaction. But we must also evaluate the new technologies themselves for effectiveness as components within cognitive-work systems.

We seek a framework for creating a "fast track" for evaluating work methods and the computer technology that's an intrinsic part of our methods.

Table 1 presents the considerable variety of things we might measure.

In a previous essay in this department, we argued that workarounds and kluges were inevitable realities that we can study empirically and cannot tacitly sweep under a carpet as if they had no significance.[16] We can easily conduct ethnographic studies of sociotechnical work systems to find instances of workarounds and kluges on the basis of an ontology.[16] We can then measure such things as time to create and estimate such things as time saved when used. Kluges and similar informal processes and procedures are, we believe, only hard to specify and measure before we bother to take an empirical approach.

There's one other possibility, that of measuring "negative hedonicity" (which we'll define in a moment). This idea stems from a previous essay in this department, which presented the Pleasure Principle of human-centered computing: "Good tools provide a feeling of direct engagement. They simultaneously provide a feeling of flow and challenge."[17] Notions of "hedonomics" and related ideas have emerged in the context of human factors and industrial design.[18,19]

Negative hedonicity

Measures of "raw" performance (efficiency, accuracy, errors) hold work methods hostage to human motivation. Typically, complex cognitive systems (that is, new technologies) do the reverse, holding human motivation hostage to work methods (especially software and interface systems). Thus, it's important to study and understand vital motivational factors. This includes positive affect (for instance, the feeling of "being in the problem" versus "fighting with the technology") and increased intrinsic, goal-oriented motivation.[19] *Negative hedonicity* is the valuation of affect and motivation as negatively impacted by the work experience. This dimension is reflected in frustration, confusion, mental (or data) overload, and automation surprise.

We seek a framework for creating a "fast track" for evaluating work methods and the computer technology that's an intrinsic part of our methods.

Negative hedonic measurement is now possible using a remarkably simple device, the Hancock Switch, which consists of a prominent red button placed next to each workstation operator (see figure 1). The button's normally open circuit connects to a digital signal generator that sends a signal to the main workstation when the switch is closed. Residing on that computer is software that creates a time-stamped flag in the trace of the trial events whenever the workstation operator presses the button. We call these signals *hedonic flags*. Participants are instructed, "At any time during the study, if you feel mentally overloaded, confused, or frustrated for any reason, just press the button."

Theoretically, this causes little or no interference and doesn't change the "ordinary" course of cognitive work. The hedonic flag task leverages the natural human inclination to apply greater force to their tools (the computer keyboard, in this case) at times of frustration (for example, during inadequate feedback from the machine).[20] Although we think of posting a hedonic flag as a form of dual task, it's actually a secondary task. However, it should entail minimal entrainment of additional cognitive resources because the user is already frustrated with the primary task.[21] Posting a flag can occur at the same stage as processing the primary task (the response stage) and can involve the same modality (visual processing) and the same channel (visual focal attention). However, posting occurs only when the primary task has already been frustrated—in other words, when the primary performance has hit a momentary hiatus. Thus, rather than casting this within the dual-task interference paradigm, we see this as affect-induced redirection.

Measures and measurements

The fundamental measure of negative hedonicity would be the number of hedonic flags posted per trial or session per participant (NHF). The NHF trace would have an interesting advantage from the perspective of experimental design—that is, it would immediately enable us to incorporate a method of task reflection. As is sometimes cited in the psychology literature on introspection, Oswald Külpe and his students developed a method they called systematic postexperimental introspection.[22] Today, this would be referred to as a method of retrospection or task reflection and be referred to as the analysis of verbal reports[23] or a form of cognitive task analysis.[24] It's generally understood in cognitive psychology that meaningful and useful data on reasoning come from analyses of verbal reports, sometimes based on

Table 1. A variety of system-level measurables.

Things to increase	Things to reduce	Things to avoid
Usefulness of the technology	The gap between the "actual work" and the "true work"	Working the technology ("make-work")
Usability of the technology	Mental workload	Fighting the technology ("workarounds")
Justified trust in the technology	Time/effort	Misunderstanding the technology ("automation surprises")
Enhanced immersion ("being in the problem") or positive hedonicity	Negative affect/frustration (negative hedonicity)	
Enhanced direct perception, recognition, comprehension	Uncertainty	
Accelerated achievement of proficiency	Unjustified trust in the technology	
Enhanced intrinsic motivation	Unjustified mistrust in the technology	
Effective coping with rare or tough cases		
Rapid recovery from error		

a "think-aloud problem solving" task and sometimes based on task reflection, as in the Critical Decision Method.[25] In a number of communities of practice (sociotechnics, computer-supported collaborative work, work ethnography, ethnomethodology, and others), it's understood that rich, useful data on cognitive work in complex systems come from analyzing the content of communications and well-conducted interviews that scaffold the participant in the recall and analysis of recently encountered tough cases.[25,26]

In the Külpe method, the participant is run through the entire study a second time, this time reviewing the complete trace (using our modern technology, perhaps including video). The marks for hedonic flags would serve as memory cues, allowing for more detailed exploration into the reasons for each posting. Why was a person who performed especially well overall suddenly frustrated? Why was another person repeatedly confused? What was it about the tool that another person didn't understand at some point in a scenario? Such performance-based evaluation would go well beyond the vagaries and vicissitudes of superficial user surveys that shoehorn meanings into designers' categories, or one-off, post hoc questionnaires that leave people prone to bias from task-demand characteristics, or other forms of satisficing that merely serve to show that "some people liked it, more or less, some of the time."

There are additional possibilities. The researcher could

- look at NHFs posted over the length of a single trial or over blocks of trials;
- evaluate individual differences by examining a specific range statistic—for instance, comparing the number of flags posted by two participants, the one who posts the most flags versus the one who posts the fewest flags; or
- evaluate range statistics that are based on the principal performance measure—that is, comparing the number of hedonic flags posted by the best- and the worst-performing participants).

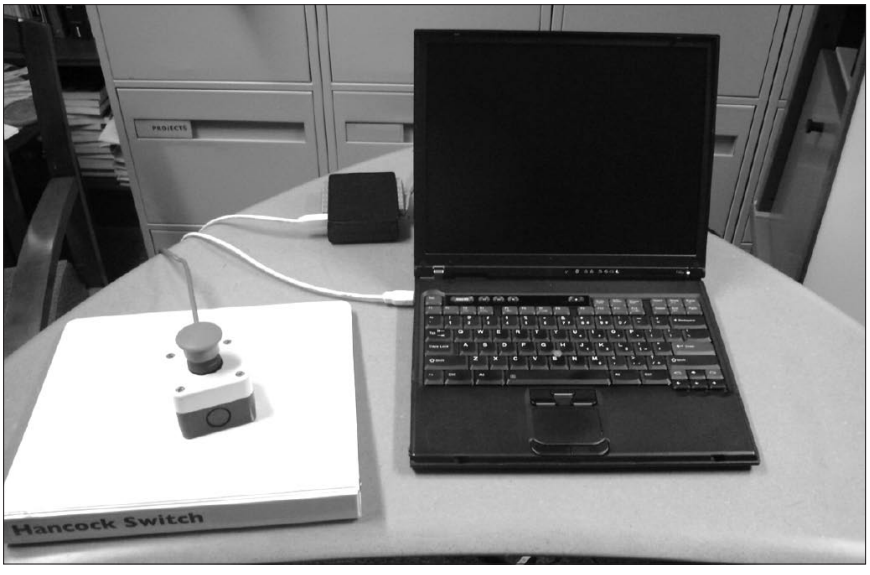


Figure 1. The Hancock Switch linked to a laptop computer.

The researcher could then use independent variables that define the main study design (for instance, easy versus difficult scenarios, or individual versus team work) to guide evaluation of the respective hedonicity measurements. For example, the study could look at the difference between

- the average number of flags posted by participants when working on the scenarios resulting in the best performance, and
- the average number of flags posted by participants when working on the scenarios resulting in the worst performance.

Alternatively, the study could look at the differences of differences, contrasting

- the best- and worst-performing participants on the scenario resulting in the best performance, with
- the best- and worst-performing participants on the scenario resulting in the worst performance.

Such studies might clarify why a tool is low in learnability or usability.

Modeling the NHF data

We postulate that the number of hedonic flags posted in a given time interval will follow a Poisson distribution—that is,

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0,\ 1,\ 2,\ \ldots$$

where λ is the rate of flags over time.

If the participants are somewhat homogeneous in their propensity to post flags, the rate parameter λ alone would enable us to compare work methods. Drawing inferences on the rate parameter of a Poisson distribution is a straightforward process.

If the participants vary in their intrinsic tendency to post hedonic flags, which of course is likely, a Bayesian approach with a distribution assigned to the rate parameter λ would be an appropriate derived measure. For example, placing a gamma distribution on λ results in a negative binomial distribution. In this case too, drawing inferences is straightforward.

Whether or not participants are homogeneous in their propensity to post flags, they'll presumably do so in an experiment's early minutes or trials because the work method will be unfamiliar and confusion more likely. This being the case, and given that the data are discrete, the distribution might take the form of the negative binomial. The most likely approach that is appropriate for data modeling would be inferences based on a cumulative probability function.

Another possibility for analyzing NHF data involves a novel variant on signal detection analysis. Generally, the main measures in traditional signal detection theory (SDT), called d′ and Beta, are intended to separate out response bias and thereby result in a cleaner measure of operator sensitivity. These are always calculated with reference to individuals' performance, making SDT of limited use in the study of sociotechnical systems. One might, however, calculate hit rates from the number of hedonic flags posted by two or more participants referenced to the same trial scenario. For instance, if a particular event resulted in three participants posting hedonic flags, and subsequent retrospections revealed that they had the same reason for that posting (for example, confusion resulting from scenario-induced mental overload), then the number 3 would be added to a sum along with numbers representing all other such consensus postings. We could compare this group hit rate to the number of nonconsensus postings for each individual (to determine false alarms) and to the number of consensus postings that a given individual didn't enter (to determine misses). From these respective calculations, we might derive sensitivity and response bias measures (d′ and Beta) with respect to aspects of the work method or scenario that are linked to operator hedonic response and yet aren't the reflection of any one operator's hedonic responsivity (or bias).

Our ideas for modeling NHF data are speculative, and experimental results will soon permit an evaluation of the utility of what we propose. We've tried in this essay to present, in a concrete and nonspeculative way, some "metrics" that relate directly to customer needs (that is, performance measures). At the same time, we hope these metrics allow meaningful evaluation of the complexity of cognitive work, one might say, sneaking system-level considerations in through the back door. ■

> We must take the measurement of cognitive work to entirely new levels—addressing, for example, the important trade-offs in cognitive work at the team and systems level.

Acknowledgments

Robert Hoffman's contribution was through participation in the Advanced Decision Architectures Collaborative Technology Alliance, sponsored by the US Army Research Laboratory under cooperative agreement DAAD19-01-2-0009.

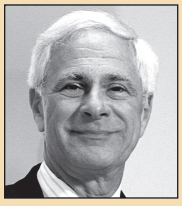
References

1. S.W.A. Dekker, J.M. Nyce, and R.R. Hoffman, "From Contextual Inquiry to Designable Futures: What Do We Need to Get There?" *IEEE Intelligent Systems*, Mar./Apr. 2003, pp. 74–77.
2. D.D. Woods, "Designs Are Hypotheses about How Artifacts Shape Cognition and Collaboration," *Ergonomics*, vol. 41, 1998, pp. 168–173.
3. E. Hollnagel and D.D. Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*, Taylor and Francis, 2006.
4. J. Goguen, "Towards a Social, Ethical Theory of Information," *Social Science Research, Technical Systems, and Cooperative Work*, G. Bowker et al., eds., Lawrence Erlbaum Associates, 1997, pp. 27–56.
5. K. Neville et al., "The Procurement Woes Revisited," *IEEE Intelligent Systems*, Jan./Feb. 2007, pp. 72–75.
6. W.W. Gibbs, "Software's Chronic Crisis," *Scientific Am.*, Sept. 1994, pp. 72–81.
7. R.W. Pew and A. Mavor, eds., *Human-System Integration in System Development: A New Look*, Nat'l Academy Press, 2007.
8. J. Grudin, "Utility and Usability: Research Issues and Development Concepts," *Interacting with Computers*, vol. 4, 1992, pp. 209–217.
9. M.B. Rosson and J.M. Carroll, *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*, Morgan Kaufmann, 2002.
10. M.J. O'Neill, *Measuring Workplace Performance*, 2nd ed., Taylor and Francis, 2007.
11. E. Schaffer, *Institutionalization of Usability*, Addison-Wesley, 2004.
12. E.J. Keats, *John Henry: An American Legend*, Pantheon Books, 1965.
13. R.R. Hoffman, P.J. Hayes, and K.M. Ford, "Human-Centered Computing: Thinking In and Outside the Box," *IEEE Intelligent Systems*, Sept./Oct. 2001, pp. 76–78.
14. R. Scott et al., "Work-Centered Support Systems: A Human-Centered Approach to System Design," *IEEE Intelligent Systems*, Mar./Apr. 2005, pp. 73–81.
15. J. Scholtz, "Metrics for Evaluation of Software Technology to Support Intelligence Analysis," *Proc. Human Factors and Ergonomics Soc. 49th Ann. Meeting*, Factors and Ergonomics Soc., 2005, pp. 918–921.
16. P. Koopman and R.R. Hoffman, "Work-Arounds, Make-Work, and Kludges," *IEEE Intelligent Systems*, Nov./Dec. 2003, pp. 70–75.
17. R.R. Hoffman and P.J. Hayes, "The Pleasure Principle," *IEEE Intelligent Systems*, Jan./Feb. 2004, pp. 86–89.
18. T. Oron-Gilad and P.A. Hancock, "The Role of Hedonomics in the Future of Industry, Service, and Product Design," *Proc. Human Factors and Ergonomics Soc. 49th Ann. Meeting*, Human Factors and Ergonomics Soc., 2005, pp. 1701–1704.
19. P.A. Hancock, A.A. Pepe, and L.L. Murphy, "Hedonomics: The Power of Positive

and Pleasurable Ergonomics," *Ergonomics in Design*, Winter 2005, pp. 8–14.
20. Y. Qi, C. Reynolds, and R.W. Picard, "The Bayes Point Machine for Computer-User Frustration Detection via Pressuremouse," *Proc. 2001 Workshop Perceptive User Interfaces* (PUI 01), vol. 15, ACM, 2001, pp. 1–5.
21. C.D. Wickens, "Multiple Resources and Performance Prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, 2002, pp. 150–177.
22. E.G. Boring, "A History of Introspection," *Psychological Bull.*, vol. 50, 1953, pp. 169–189.
23. K.A. Ericsson and H. Simon, *Protocol Analysis: Verbal Reports as Data*, 2nd ed., MIT Press, 1993.
24. B. Crandall, G. Klein, and R.R. Hoffman, *Working Minds: A Practitioner's Guide To Cognitive Task Analysis*, MIT Press, 2006.
25. R.R. Hoffman, B. Crandall, and N. Shadbolt, "A Case Study in Cognitive Task Analysis Methodology: The Critical Decision Method for the Elicitation of Expert Knowledge," *Human Factors*, vol. 40, 1998, pp. 254–276.
26. R.R. Hoffman and L. Militello, *Perspectives on Cognitive Task Analysis: Historical Origins and Modern Communities of Practice*, CRC Press, 2008.

Robert R. Hoffman is a research scientist at the Institute for Human and Machine Cognition. Contact him at rhoffman@ihmc.us.

Peter Hancock is a professor of psychology at the University of Central Florida and senior research scientist at the Institute for Human and Machine Cognition. Contact him at phancock@pegasus.cc.ucf.edu.



Morris Marx is a senior research scientist at the Institute for Human and Machine Cognition and president emeritus of the University of West Florida. Contact him at mmarx@ihmc.us.