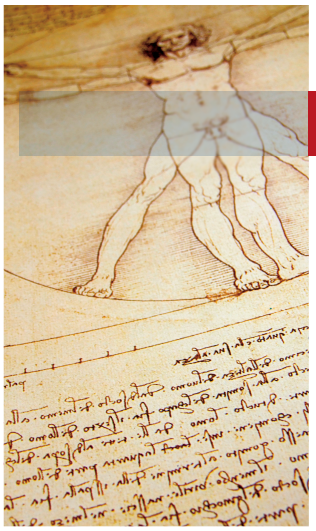

Editor: **Robert R. Hoffman, Jeffrey M. Bradshaw, and Kenneth M. Ford,**
Institute for Human and Machine Cognition, rhoffman@ihmc.us

Metrics, Metrics, Metrics, Part 2: Universal Metrics?

Robert R. Hoffman, *Institute for Human and Machine Cognition*
Peter A. Hancock, *University of Central Florida*
Jeffrey M. Bradshaw, *Institute for Human and Machine Cognition*

A previous article in this department from 2008 introduced the topic of measures and metrics.[1] The focus of that essay was on measurement of the "negative hedonics" of work—the frustrations, uncertainties, mistrust, and automation surprises caused by poorly designed technology that is not human-centered. This second article focuses on the concept of "metrics" and issues related to it. Following a discussion of relevant issues, we present an immodestly bold proposal for a set of universal metrics.

Metrics have been a salient topic of many recent government-funded research programs for developing large-scale information systems. We have counted a multitude of funding program announcements that include statements such as the following abstraction:

> The program seeks metrics quantifying the value and risk added by new information, processes, and modalities ... [The program seeks] the quantitative and qualitative metrics required by the acquisition community to use human systems integration tools and processes in the design process …

This is an expression of "The Great Hope," to be codified in mathematics in the same manner it has now been reified in the legal language of the procurement process. The doomed expectation is that if something can be measured we therefore will be able to understand it. This is a thinly veneered disguise for the reductionistic obsession to measure the success of everything by its return on investment. The sought-for measures defined in these terms have to gauge efficiency, effort, accuracy, and similar reflections of a maximizing process, hearkening back to the contest between John Henry and the steam hammer. This myopic perspective is particularly frustrating for the advocates of human-centered computing and work-centered design.

Our view is that we must measure cognitive work at the system level—addressing, for example, the important trade-offs and the wider effects of technology-induced changes in the culture of the workplace and the health of the community at large. The quantitative characterization of highly complex interactive effects, or even more problematically, subjective apperceptions of the world, presents one of the greatest challenges to advancing technology in the early 21st century.[2]

Cognitive systems engineers have called for new objective methods for evaluating the performance impact and learnability for software systems,[3–7] including the increasing number of systems requiring human-automation teamwork of a consequential sort.[8,9] In general, there has been a rising concern with the human factors of complex cognitive work, or "metrics and methodologies for evaluating technologies."[10]

So what exactly does it mean to ask for a metric?

Measures versus Metrics

To understand the foundations for a demand for metrics, we illustrate some basic ideas on measurement through the use of the simple example of intelligence testing (see Table 1).

Metrics are thresholds or decision criteria that are used in an evaluation. One has to decide, "This value differentiates." Such decisions arise from policy shaped by goals, value judgments, and other considerations. The policy that leads to the adoption of particular operational definitions (for example, "What do we want to measure?") is external to measurement. Thus, metrics do not arise either immediately or automatically from measures or measurement scales. Assuming that we

1541-1672/10/$26.00 © 2010 IEEE
Published by the IEEE Computer Society

Table 1. Some fundamental concepts of measurement.

Steps to get from a theoretical concept to a metric	Meaning	Example
Conceptual measurable	This is a concept from the subject matter of some theory. It is assumed that instances of this concept can be identified and counted.	A theory of cognitive development might assert that there are individual differences in intellectual capacity, referred to as intelligence.
Operational definition	This is a specification of a replicable, dependable procedure for counting instances or making measurements.	This is what an intelligence test does, as in the phrase, "Intelligence is what an intelligence test measures."
Numerical scale	The numerical scale expresses distinctions regarding the conceptual measurable. The distinctions could be categorical or numerical.	"Genius" is a categorical distinction. The IQ is a ratio of mental age and chronological age, which is a numerical distinction.
Measurement scale	The numerical scale values are entered into a calculation that creates a derived measurement scale.	IQ scores expressed as percentile ranks.
Measurement	A specific observation is regarded as an instance of a specific scale value.	A specific measured IQ score and its derived percentile rank is an example of a measurement.
Metric	A decision threshold is expressed as some value (minimum, maximum, or range) on the numerical scale.	If a person's IQ score measured at age 16 is at the 85th percentile or greater, they get to go to college. If a person's IQ score is 145 or greater, they are classified as genius.

have successfully gone from a conceptual measurable to one or more reasonable operational definitions, and assuming further that we have successfully derived specific measurement procedures and linked the measurements to one or more meaningful measurement scales, we cannot then assume some easy step to a metric without having some sort of policy or goal. Without some policy to specify what is desired (or good), how can we determine what a decision threshold should be? In one context, 85 percent correct might represent a useful metric. In another context, it might be misleading or indeed genuinely dangerous. In one context, 35 percent better than before might be a significant gain, whereas in another content, 35 percent might be negligible.

With this understanding as background, we argue that on certain assumptions that apply to the evaluation of computer-supported cognitive work, it might be possible to generate universal metrics that place all performance evaluations essentially on a common playing field.

The Designer's Gamble

In the standard view of hypothesis testing, real-world variability must be restricted either by passive control or more often by active manipulation. Multiple experiments are always required to peg down the determiners of human skill acquisition and performance, especially in macrocognitive work systems. Potentially, any feature of the participants (such as experience, intelligence, motivation, aptitude, and so on), test scenarios (such as interesting, rare, easy, or boring), teams (such as colocated, asynchronous, and dysfunctional), and tools (such as displays and menus) can prove relevant, as can countless other mediating and moderating variables.[7,11]

This means that if an experimental paradigm were conducted properly, the development and procurement process would take even longer than it already does and, in theory, could require a boundless sequence of tests. This would be an untenable situation at a time when the priority is to drastically reduce procurement time.[12–14] Furthermore, by the time the relevant factors have been controlled, key variables isolated, and effect sizes estimated, design requirements changed and reevaluated, and so forth, the cognitive work will almost certainly have evolved or been transformed, sometimes completely.[15]

"It is difficult to sample all the things that must be sampled to make a generalization ... the sheer number [of interacting factors] can lead to unwieldy research plans."[16] We call this the "fundamental disconnect."[17] We need to reduce the time frame required for experimentation so that its length does not preclude effective change in an evermore rapidly changing world. We need to find alternatives to both standard usability testing and standard controlled experimentation so as to expedite evaluation of the performance effects of technological interventions in macrocognitive work systems.[4,7,18]

In light of this conundrum, we think it might be fruitful to question some of the basic methodological assumptions in the standard experimental model. For example, there is the question of controlling for variables in the workplace. Let us ask the following: If all the interacting and uncontrollable factors are in effect when the actual work is being performed, why should we assume that they have to be controlled when work methods are being evaluated? Indeed, we actually need the daunting variability of the world to be represented in the evaluation of new technologies.[7,18] The traditional approach asserts

that the only path to scientific truth is to conduct an extended series of controlled factorial experiments resulting in measures of statistical significance. However, in macrocognitive work domains, we need an approach that emphasizes ecological representativeness and utility and leads to measures of practical significance.[19] Thus, we express what we call the Designer's Gamble, which can be stated as follows:

> We, the designers, believe that our new technology is good, and that good work will result from its use. Thus, we can let the daunting variability of the "real world" remain in the summary statistics and measurements, and we can conduct reasonably risky tests of usefulness and usability. We're going to gamble that the new technologies and the work methods they instill are so good that improvements in the cognitive work will be straightforwardly demonstrable despite the daunting variability of the real world.

We think that the Designer's Gamble is no mere fantasy. Just as funding program announcements sometimes appear to ask for the world, research proposals often gladly promise it. Statements of the following general type often appear in grant proposals and preproposal white papers:

> We will develop new modeling strategies leveraging previous research in dynamic networked environments. This architecture will provide near real-time interoperability and robustness and will allow the detection and modeling of information flows and actions and mitigate data overload. This will then be integrated with a suite of algorithms that will automatically reconfigure the running simulation.…

Overly confident statements such as these, relying heftily as they do on the word "will," promise more than can ever be guaranteed. Other words, such as "might," would be more appropriate. Phrases such as "we hope will" would be more honest. Organizations, teams, and individuals who seek to create information technology invariably justify their entire approach and design rationale on a tacit Designer's Gamble.

The Designer's Gamble can be an explicit assumption made during the processes of procurement (such as system development and evaluation). As such, it is a leverage point for empirical analysis and, in particular, testing hypotheses about the goodness of software tools. What follows from the Designer's Gamble is a way around the fundamental disconnect, through the explicit use of range statistics in which we look at the extremes of performance.

Range Statistics and Universal Metrics

Comparing the best and worst performers using a new technology, as opposed to their performance when using their legacy methods, informs the evaluator about both the technology's learnability and the quality of the work that results. The best performance of an individual (or team) demonstrates what is possible with the new technology, while results from the worst performing individual (or team) can draw attention to training, work variability, or selection problems. Neither extreme represents an aberration to be glossed over by calculations of averages or standard deviations. This is especially important for a statistical analysis of situations where the participants are beginners with the tools because we know that measurements taken on initial task performance are not normally distributed.[18,20–22] Thus, we can devise a

Table 2. Universal Metrics levels.

Metric level	Definition
Metric level 0 (minimal)	Range statistics are not distinguishable from those in the legacy work.
Metric level 1 (mixed)	The best performer in the new work performs better at achieving the primary task goals than the best performer using the legacy work method, but the worst performer of new work performs worse at achieving the primary task goals than the worst performer at the legacy work method. Metric level 1 is an expected outcome because it is somewhat likely that any intervention will tend to increase performance variability. At this metric level, the performance at the high end improves. Worsened performance at the low end indicates a need for either an improved job selection criterion or improved training.
Metric level 2 (improved)	The worst performer of new work performs better at achieving the primary task goals than the worst performer at the legacy work.
Metric level 3 (good)	The best performer of new work performs better at achieving the primary task goals than the best performer at the legacy work.
Metric level 4 (excellent)	The worst performer of new work performs better at achieving the primary task goals than the worst performer at the legacy work, and the best performer in the new work performs better at achieving the primary task goals than the best performer at the legacy work.
Metric level 5 (superior)	The worst performer of new work performs better at achieving the primary task goals than the best performer at the legacy work.

Figure 1. A visual explanation of the meanings of the Universal Metrics levels.

set of universal metric levels for comparing new work methods to legacy work methods, and for evaluating the learnability of work methods. Table 2 presents one such set of metrics.

Figure 1 illustrates these universal metrics levels.

The universal metrics levels presume that the new work involves the same principal task goals as the legacy work, for which data are available to form a baseline used in establishing what counts as "best" and "worst" performance. (For cases in which the work involves completely new kinds of tasks, there might not be an historical baseline and the evaluation will initially have to reference some normative model or theory of the work. An example might be the emerging forms of cyberdefense. This issue of "formative design," however, is a serious and significant topic deserving of its own separate analysis.)

Using these universal metrics levels represents a risk on the part of evaluators. The outcome of an evaluation hinges on the performance of one or two individuals (or teams). Remembering that the performance of all the others will fall between the extremes, in the "real world," operational performance likewise often hinges on the performance of one or two individuals (or teams). If the evaluation is explicitly founded on the Designer's Gamble, and if the desire is to evaluate software in ecologically valid conditions, then the logic of the approach holds firm. It is of course possible, although we hope not likely, that evaluators will try to finesse this method by eliminating poor performers post hoc (for example, on some claim of validity, such as "they were sick"). But such finagling would be transparent and counterproductive.

If, on the other hand, a software tool developer does not wish to adopt the Designer's Gamble up front, then any a priori promises about the performance gains that will result from the to-be-delivered capabilities must be expressed in a far more cautious way than we commonly see today.

Either way, the sponsor wins.

We submit that the concept of universal metrics levels can provide a framework that includes the following:

1. an approach to evaluation that emphasizes ecological representativeness and utility and escapes the constraints imposed by traditional controlled factorial experimentation;
2. a means for measuring practical significance rather than (or in addition to) statistical significance;
3. a common playing field for evaluating performance of software-supported work of all kinds;
4. a common playing field for evaluating the learnability of software-supported work methods and, by implication, the goodness of the software tools; and
5. a means of sensitizing analyses to outliers that might signal training or selection issues.

What we offer here is not a closed-end solution. Rather, it is a first step or a prospectus in a challenging journey to rethink, document, and quantify the character and capacities of large-scale interacting human-machine systems. ■

References

1. R.R. Hoffman, P. Hancock, and M. Marx, "Metrics, Metrics, Metrics: Negative Hedonicity," *IEEE Intelligent Systems,* vol. 23, no. 2, 2008, pp. 69–73.
2. P.A. Hancock, J.L. Weaver, and R. Parasuraman, "Sans Subjectivity, Ergonomics is Engineering," *Ergonomics*, vol. 45, 2002, pp. 991–994.
3. R.R. Hoffman, K.N. Neville, and J. Fowlkes, "Using Cognitive Task Analysis to Explore Issues in the Procurement of Intelligent Decision Support Systems," *Cognition, Technology, and Work*, vol. 11, 2009, pp. 57–70.
4. R. Lipshitz, "Rigor and Relevance in Naturalistic Decision Making Research: How To Study Decision Making Rigorously with Small Samples and Without Controls or Statistical Inference," to be published in *Cognitive Engineering and Decision Making,* 2010.
5. K. Neville et al., "The Procurement Woes Revisited," *IEEE Intelligent Systems*, vol. 23, no. 1, 2008, pp. 72–75.
6. W.M. Newman, "On Simulation, Measurement, and Piecewise Usability Evaluation," *Human-Computer Interaction*, vol. 13, no. 3, 1998, pp. 317–323.