Range Statistics and the Exact Modeling of Discrete Non-Gaussian Distributions on Learnability Data

Robert Hofman

Institute for Human and Machine Cognition, Pensacola, FL 32502, USA
rhoffman@ihmc.us

Abstract. A measure called i-bar is presented, which is the inverse of the mid-range derived from data on trials-to-criterion in tasks that require practice. This measure is interpreted as a conjoint measurement scale, permitting: (a) evaluation of sensitivity of the principal performance measure (which is used to set the metric for trials to criterion); (b) evaluation of the learnability of the work method (i.e. the goodness of the software tool); (c) evaluation of the resilience of the work method. It is possible to mathematically model such order statistics using negative binomial and logistic growth equations, and derive methods for generating prediction intervals. This approach involves novel ways of thinking about statistical analysis for "practical significance." The is applicable to the study of the effects of any training or intervention, including software interventions designed to improve legacy work methods and interventions that involve creating entirely new cognitive work systems.

Keywords: Range statistics, prediction intervals, rigorous usability analysis.

1 Background and Motivation

When software is introduced into the sociotechnical workplace, it represents the hypothesis that performance will be more effective and decisions will be improved (Woods, 1998). However, this hypothesis is often disconfirmed by experience. User-hostile aspects of software can induce frustration and surprise when the automation does things the worker does not understand [1]. During the development of information technology, many events result in information technology that is not human-centered—that is not usable, useful or understandable. Inadequate end-user involvement in the design process results in information technology that does not have a positive impact. Human-centering shortfalls result in a gap between the work that people must perform to achieve their primary task goals, and they work they must conduct simply because of the ways the software constrains them. User-created kluges and work-arounds are telltale signs that software is not human-centered [2].

The notorious frustrations and failures triggered by information technology interventions have led to a significant concern with evaluation and metrics in the field of software engineering and in the human factors community [3,4]. Considerable attention is being given to issues of rigor in usability analysis [e.g., 5]. Software engineers are reminded that graphical interfaces should be easy to learn, efficient to use, and

A. Marcus (Ed.): Design, User Experience, and Usability, Pt II, HCII 2011, LNCS 6770, pp. 421–430, 2011.

subjectively pleasing. Such factors are contrasted with more traditional considerations of acquisition cost, reliability, etc.

In many information technology system development projects, usability evaluation is often based on some sort of “satisficing” criterion [6]. Users are queried about their reactions to information technology, and the sole metric is user satisfaction evaluated using surveys, questionnaires, observer ratings of user activity, or "walkthrough" interviews conducted while users attempt to conduct tasks [e.g., 7]. Sometimes the sole evaluation is whether a senior member of the organization likes the new tool, perhaps after a demonstration of design features. Researchers have recently become aware of the need to empirically distinguish between the perceived usability (or "goodness") of software and the perceived attractiveness of the interface [8]. “Learnability”—that is, the speed at which one learns to achieve primary task goals—has been proposed as a measurable variable [e.g., 9], but attempts to evaluate learnability have actually measured the readability of software documentation [10]. Both software systems engineers [e.g., 11] and cognitive systems engineers have called for new objective methods for evaluating the performance impact and learnability of software systems [4,12,13].

According to the Moving Target Law of Human-Centered Computing [14], the sociotechnical workplace is constantly changing, and this entails change in cognitive constraints. In domains of importance to society, business, and government (e.g., emergency response, military command, intelligence analysis, and many others) cognitive work mixes legacy work methods and technologies with new ones, addressing old problems while struggling with emerging challenges. Thus, procurement cannot depend solely on hierarchical task decomposition [as in 15], because the work will almost certainly have changed by the time such an analysis is complete. In fact, the more detailed the task decomposition, the more likely it will be brittle in the face of resilience tests, and the more quickly it will become obsolete. This results in what we call the *fundamental disconnect:* The time frame for effective experimentation is too slow to match the pace of change in the work and technology of sociotechnical systems.

"Effective" experimentation requires studies in which variables (e.g., display designs, software features, etc.) are manipulated and controlled. Multiple experiments are always required to peg down the causal determiners of performance and skill acquisition, especially in human-computer interaction. However, "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, pp. 18-19; see also 31]. There are far too many variables to be taken into account and hence too many experimental tests to be conducted. Features of the participants (experience, intelligence, motivation, aptitude, etc.), the test scenarios (interesting, rare, easy, boring, etc.), the teams (co-located, asynchronous, dysfunctional, etc.), the tools (e.g., it is sometimes quite easy for human factors engineers to make better interfaces than ones made from a designer-centered design approach), and experimental task demand characteristics can all have causal efficacy, as can numerous mediating and moderating variables [13].

This means that an information technology development and procurement process should (by the standards of effective experimentation) hinge on a rather lengthy series of studies that go beyond measuring mere user satisfaction. Procurement would take

even longer than it already does, at a time when a main challenge is to reduce procurement time [17]. By the time the relevant factors have been controlled, key variables isolated, and effect sizes estimated, design requirements will have changed and been re-evaluated, etc., the cognitive work will almost certainly have passed on to other incarnations. Therefore, we need to escape the fundamental disconnect between the time frame for experimentation and the time frame for effective change in the world of events that are a more or less unique morass of causal influences. In sum, both standard usability testing and standard controlled experimentation have limitations in regard to evaluating performance effects of technological interventions in joint human-machine cognitive work systems [13].

2 The Range Statistics Approach

In the standard view of performance evaluation, real-world variability must be restricted either by being controlled or manipulated. However, when doing the actual work, all of the variability of the world is in play. Intelligence, capability, motivation, alertness, problem difficulty, and many other factors might influence the relation between learning and performance and all are in play in real-world work environments (e.g., a command post). One worker might have high intrinsic motivation and high intelligence; another might have insufficient experience, low aptitude, and be suffering from the flu, etc. One *wants* such daunting variability of the world to be preserved during the evaluation of new technologies. We express what we call The Designer's Gamble:

> *We, the designers, believe that our new information technology is good, and that good work will result from its use. Thus, we must let the daunting variability of the world remain in the summary statistics, and we can conduct reasonable and yet risky tests of usefulness and usability. We are going to gamble that the software/tools and new work methods are so good that the cognitive work (human-system integration, etc.) will be measurably superior despite the daunting variability of the world.*

The Designer's Gamble is no fantasy on our part: proposals often promise it. Statements of the following general type often appear in proposals, which we paraphrase: "We will develop new modeling strategies for an architecture that will provide near real-time interoperability and robustness and mitigate data overload. This will then be integrated with a suite of algorithms that will automatically reconfigure the running simulation..." Such statements are promissory notes, as shown by the reliance on the word 'will.' Organizations, companies, and teams that seek to create information technology invariably base their design rationale and methodology on the Designer's Gamble.

The Designer's Gamble is an assumption made during the processes of procurement. As such, it is a leverage point for empirical analysis and, in particular, testing hypotheses about the goodness of software tools. In other words, the Designer's Gamble suggests a way around the fundamental disconnect. Range statistics represent

a fast-track solution that can address questions concerning the goodness of the cognitive work, on the assumption of the Designer's Gamble. Researchers could still treat variability as something to be analyzed in exploration of hypotheses about the cognitive work, even using the familiar parametric statistical tests on data from experiments on human-computer interaction to probe the meaning of the variability (e.g., effects of co-located versus distributed teams). In other words, researchers can conduct the usual sorts of experiments they conduct to study human-computer interaction. However, at the same time, range statistics can be pulled out and utilized as a fast-track probe of the goodness of the software tools.

In the majority of experimental evaluations of human learning and task performance (including human-computer interaction), data are usually not collected during the practice trials that are part of the initial instruction. After all, the participants are just learning the basics of the task (i.e., the "button-ology"). Any data that might be collected would typically not be of any particular interest with regard to the major hypotheses being studied in the main trials (e.g., can people respond faster when provided with graphical rather than numeric representations in an interface?). We argue that performance on those earliest practice trials is a neglected resource, and a number of measures taken during practice trials could be informative. Furthermore, practice could involve training to a criterion level of performance, however many trials that might take, so that the experimenter has reason to believe that the participants have all learned the system to the same degree. This strategy serves to address the possible confound in the interpretation of the effects of the independent variables that are formative of the design of the main experiment. This same strategy can also be used to evaluate work methods through the use of range statistics. A generic design is illustrated in Figure 1.

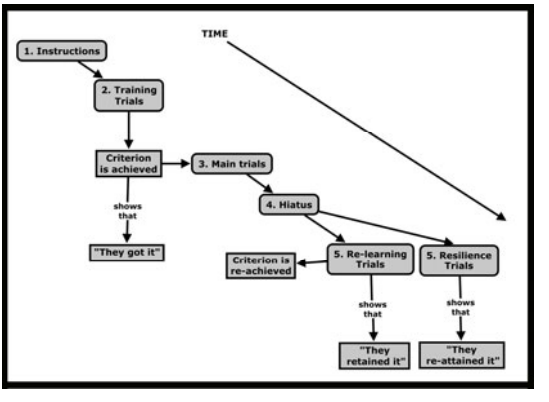

Fig. 1. A general experiment design

The derived measure of learning trials-to-criterion is familiar in experimental psychology [38]. It is used in investigations of the progress of learning, the strength of learning, and the decay of memory. It also permits control for the degree of original learning.

3 Principal Performance Measures and Range Statistics

The evaluation of performance and skill acquisition (of individuals and of teams) requires measures that reflect effectiveness and efficiency. The choice of a principal performance measure will depend on the domain, the specific job goals and other features of the cognitive work. A conceptual generic definition of a principal performance measure is *number of principal task goals successfully accomplished per unit time at work*. For example, the sensor payload operator of an unmanned aerial vehicle might have a principal performance measure of "number of targets photographed per sortie." Performance of an emergency response team in a simulated scenario might be measured by "number of victims rescued within the first hour of the response."

A principal performance measure can be used to form the criterion for training. Participants might perform as many trials as necessary to achieve some pre-specified criterion. Practice can involve training to a lax or a strict criterion. Any reasonable value can be used initially, and whether it is liberal or conservative will depend upon the work system and the nature of the individuals who are expected to operate it. For tasks having historical performance precedent, the criterion might be based on archived data, baselines, or legacy training standards. While the performance measure and the criterion are domain- and task-specific, the derived measure of trials-to-criterion suggests a general technique for usability analysis based on examination of the "novice user's experience at the initial part of the learning curve" [9, p. 28].

The literature on the acquisition and retention of skill (both motor and cognitive tasks) encompasses hundreds of studies and review articles, and in the majority of tasks that have been studied, the criterion of "proficiency" or "minimal mastery" is typically defined as one to three errorless trials [19]. Typically, we expect a steep incline and the achievement of a reasonable level of proficiency within a short time for highly learnable systems. Counts of trials to achieve (or re-achieve) criterion are highly unlikely to have a Gaussian distribution [18]. Trials-to-criterion is an instance of a process where values are constrained by some sort of stopping rule. One would expect relatively few participants to achieve criterion on the first, second, or even perhaps third trial; if they did succeed, then one would have to conclude that the cognitive work was trivial. However, one would expect to see many participants achieve criterion after more than a few trials. Such a distribution of small numbers of small numbers typically would be highly skewed and have a "fat tail." This characterizes distributions such as the negative binomial, illustrated in Figure 2. For such cases, an order statistic (range or median) is preferred because the average will be misleading and unrepresentative (i.e., there is a considerable difference between the mean and the median)[20]. Furthermore, we are not interested in averages; we are interested in the form of entire distributions, and especially the extremes.

We want to focus on the use of range statistics to evaluate the work method. Range statistics allow the "daunting variability of the world" to become more readily understandable through analysis.

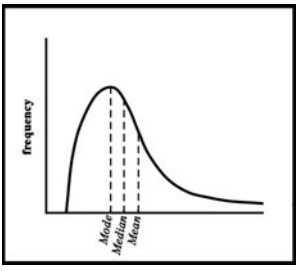

Fig. 2. A stylized distribution for trials to achieve or re-achieve criterion

4 Learnability Evaluation

Letting B stand for trials-to-criterion (or trials to re-achieve criterion) for the *best* performing participant and W stand for trials-to-criterion (or trials to re-achieve criterion) for the *worst* performing participant, the mid-range is (B + W) /2. In this case of only two numbers, the mid-range is a form of average, but we are not interested in the mid-range because it may carry with it some of the properties of the mathematical average, as interesting and useful as those properties might be. Rather, we are interested in it because of the way it conserves the performance span. The particular function of the mid-range that is of interest is the inverse of the mid-range, a new statistic that we denote with the symbol $\bar{i}$ (pronounced "i-bar"), which equals 2/(B + W). We choose this function of B and W because it is a simple transformation of the mid-range and results in numbers that fall between zero and one. The i-bar statistic is one of a class of derived measures that capitalize on the ordered nature of data. Thus, statistics in this class are also referred to as Order Statistics.

The $\bar{i}$ numerical scale can be interpreted as a conjoint measurement scale, that is, it is a measure of more than one thing. If $\bar{i}$ is quite close to 1.00, both the best and worst performing participants “got it” within just a few trials. Either the cognitive work is trivial or the criterion was set too low. If $\bar{i}$ is very close to zero, then one would conclude that the cognitive work is very difficult or the criterion was set too high. Thus, the $\bar{i}$ scale can serve as a tool for fine-tuning the criterion, or guiding the selection of the learning trials cases (or problem tasks) of an appropriate degree of difficulty.

Once one has reason to believe that the criterion is appropriate, $\bar{i}$ can be interpreted as a scale of the learnability of a work method. In this situation, if $\bar{i}$ is high, the Designer won the gamble. If $\bar{i}$ is low, the Designer lost. Details of this interpretation of $\bar{i}$ are presented in Table 1. This is just one interpretation, and is based primarily on our experience in the experimental psychology laboratory. The precise $\bar{i}$ values for the intervals are likely to be domain-specific; however, the qualitative scale and the suitability of $\bar{i}$ as a measure should apply widely.

In the Range of Stretch, the cognitive work might be extremely difficult, the work method might be very low in learnability, the criterion might have been set too high, or some combination of these may be the case. In the case of extremely difficult cognitive work, differences in $\bar{i}$ at the second decimal place might be meaningful. An example might be helicopter training, where trainees receive hours of practice at the task of hovering a helicopter, taking an average of about 20 hours to receive approval

to attempt solo flight [21]. Often when people learn to fly an aircraft simulator, only after the first dozen or so practice trials, which usually result in crashes, does one begin to see trials where speed, altitude, heading, etc. are successfully maintained, even for "easy" flights.

Table 1. An interpretation of the $\bar{i}$ scale

$\bar{i}$ Scale Ranges	Values of (B, W) $\bar{i}$	Desired Discrimination
↑ Range of Trivial Cognitive Work	(1,1) 1.00 (1,2) 0.66	$\bar{i}$ between 1.00 and 0.66 suggests that the cognitive work may be trivial or that the performance criterion needs to be raised.
↓ Range of Non-trivial Cognitive Work	(1,3) 0.50	Edge of the range. Criterion may still be set too low.
↓ Range of (re)learnability	(1,4) 0.40 (2,3) 0.40 (2,4) 0.33 (2,5) 0.29 (1,6) 0.29 (3,5) 0.25 (3,6) 0.22 (4,6) 0.20	Fine discriminability is desired.
↓ Range of Stretch	(2,9) 0.18 (3,8) 0.18 (5,7) 0.16 (4,9) 0.15 ↓	Finest discriminabilty is desired.

Thus, $\bar{i}$ can be interpreted as a conjoint measure: reasonableness of the criterion, learnability, re-learnability, and stretch.

We have proposed that it is possible to augment the use of the statistic $\bar{i}$ by finding its probability distribution. This distribution depends on the choice of a probability density function to model the data (trials to achieve or re-achieve criterion). Once that is determined, the joint cumulative probability distribution function for B and W can be derived. From that, the convolution yielding the density of B + W follows and the one-to-one transformation from the density of B + W to that of $\bar{i} = 2/(B + W)$ is straightforward.

With such exact modeling, one can ask, for instance, what is the probability of an $\bar{i}$ of some value for trials to re-achieve criterion given that trials-to-criterion form a distribution of the assumed type? If that probability is low, one can conclude that the participants did not retain the original learning, and the Designer lost the gamble. If that probability is high one can conclude that the participants not only “got it” but that they also retained it, and therefore conclude that the Designer won the gamble.

5 Resilience and Team Measures

Resilience is the ability to recognize and adapt to unanticipated perturbations that might stretch the workers' competence and demand a shift of processes and strategies

[22]. For the study of resilience we can adapt a method used in clinical trials. Once there had accrued sufficient data to warrant conclusions about the learnability of a work method, there could be a session of resilience trials (rather than re-learning trials; see Figure 1) in which "the system" is stretched. This can be achieved in a variety of ways. One might simulate a communication loss, or a loss of team functionality. Performance could be evaluated by trials to re-achieve-criterion. One assumes some reasonable level of learnability but now interprets the range as a reflection of the resilience of the work system, and likewise interprets the $\bar{i}$ numerical scale as a measure of resilience.

The measure that we have described can also be applied in the analysis of work methods and technologies for teams and team cognitive work. For instance, rather than evaluating trials to achieve criterion on the part of the best (and worst) performing participants, one can evaluate $\bar{i}$ for the best (and worst) performing team. Measures can be of learnability, re-learnability and resilience with respect to the team cognitive work.

The approach we present does not seek a measure that collapses an entire set of data into a single measure of central tendency or into a single measure of variability. We are interested in testing hypotheses about extremes. "In such studies, statistical tests addressed to differences in central tendency will shield rather than reveal group differences" [23]. We remain confident that range statistics are appropriate for the analysis of the learnability of cognitive work methods. There is nothing unusual about applying extreme value statistics to practical problems, such as ranked set sampling [24].

6 Exact Mathematical Modeling of Discrete Non-Gaussian Distributions

A theoretical foundation for Range Statistics will involve exact mathematical modeling of discrete non-Gaussian distributions and some method for deriving probabilities and testing hypotheses. Both theory and empirical experience imply that the distributions of i-bar will be highly non-Gaussian. Therefore, a premise in the investigation of range statistics is that it is not appropriate to rely on the mean and variance and the analysis of variability by averaging procedures. The analysis of range statistics looks more closely at the median and mode in relation to the range. In such contexts as training, the range is especially interesting because best performance is a benchmark and worst performance represents a potential training problem. Thus, the range itself is of inherent interest.

Close-to-exact fits would be possible and would be preferred to approximations for purposes of statistical analysis. Although the standard approach in mathematical psychology is to build formulae based on a theory of learning, much can be learned from using an empirically determined probability model for the data. Once this is set, a probabilistic structure can be created. Very little research exists on the possible probability density functions of trials-to-criteria data. That question is by itself of interest, and experiments generating such data would be quite useful. What we do know is that the probability density function should be steep on the left and have a fat tail on the right. The natural occurrence of distributions that take the form of the negative binomial, and the applicability of modeling the negative binomial, have been pointed out

in the psychometrics literature [25]. Other probability density functions incorporating the degree of learning achieved from trial to trial show promise as well.

There is reason to believe that trials-to-criterion will take the form of the negative binomial [26]. A second family of distributions that might fit to trials-to-criterion data is based on the logistic growth model. This is a cumulative distribution function suited to describing processes in which there is growth up to some limit, that is, the S-curve growth of some set. A third model, specified conditional probabilities, would assign conditional probabilities describing the probability of success, *given* that successive failures have occurred.

Regardless of the method of determining the probability density function for the data, the probability density function for i-bar can be calculated. From that it is possible to derive what in classical statistics are called *confidence or prediction intervals,* the probability that i-bar will fall within certain ranges of values. From these intervals statistical inferences can be made. For example, one can ask, "Given the observed data on original learning (trials-to-criterion) of B = 3 and mid-range = 2, what is the probability that a team would re-achieve criterion on 14 trials as the worst performance in the resilience test?" From that probability, and the $\bar{i}$ result, one might derive conclusions concerning the importance of adding into the work method some means for coping with the particular resilience factors that were examined in the resilience trials.

Our progress to date shows that questions about the likelihoods of values of range statistics on discrete non-Gaussian distributions can be framed mathematically and answered. A worked-out statistical theory for the derived measure of trials-to-criterion does not exist. Such a theory, or an appropriate numerical substitute, is within our grasp; as is a software tool to support modeling, range statistics analysis, and the generation of prediction intervals.

References

1. Hoffman, R.R., Marx, M., Hancock, P.A.: Metrics, metrics, metrics: Negative hedonicity. IEEE: Intelligent Systems, 69–73 (March-April 2008)
2. Koopman, P., & Hoffman, R.R.: Work-Arounds, Make-Work, and Kludges. IEEE: Intelligent Systems, 70–75 (November/December 2003)
3. Goguen, J.: Towards a social, ethical theory of information. In: Bowker, G., Gasser, L., Star, L., Turner, W. (eds.) Social Science Research, Technical Systems, and Cooperative Work, pp. 27–56. Lawrence Erlbaum Associates, Hillsdale (1997)
4. Neville, K., Hoffman, R.R., Linde, C., Elm, W.C., Fowlkes, J.: The procurement woes revisited. IEEE Intelligent Systems, 72–75 (January/February 2008)
5. Rosson, M.B., Carroll, J.M.: Usability engineering: Scenario-based development of human computer interaction. Morgan Kaufman, San Francisco (2002)
6. Gillan, D.J., Bias, R.G.: Usability science I: Foundations. International Journal of Human-Computer Interaction 13, 351–372 (2002)
7. Sears, A.: Heuristic walkthroughs: finding the problems without noise. International Journal of Human-Computer Interaction 9, 213–234 (1997)
8. Hassenzahl, M., Monk, A.: The inference of perceived usability from beauty. Human-Computer Interaction 25, 235–260 (2010)
9. Nielsen, J.: Usability engineering. Academic Press, San Diego (1993)

10. Kantola, N., Jokela, T.: Determining high-level quantitative usability requirements: A case study. Springer, Berlin (2007)
11. Rubin, J., Chisnell, D.: Handbook of usability testing, 2nd edn. Wiley, Indianapolis (2008)
12. Hoffman, R.R., Neville, K.N., Fowlkes, J.: Using cognitive task analysis to explore issues in the procurement of intelligent decision support systems. Cognition, Technology, and Work 11, 57–70 (2009)
13. Roth, E.N., Eggleston, R.G.: Forging new evaluation paradigms: Beyond statistical generalization. In: Patterson, E., Miler, J. (eds.) Macrocognition Metrics and Scenarios, pp. 204–219. Ashgate, London (2010)
14. Ballas, J.A.: Human centered computing for tactical weather forecasting: An example of the Moving Target Rule. In: Hoffman, R.R. (ed.) Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making, pp. 317–326. Erlbaum, Mahwah (2007)
15. Shepherd, A.: Hierarchical task analysis. Taylor and Francis, London (2001)
16. Firestone, W.A.: Alternative arguments for generalizing from data as applied to qualitative research. Educational Researcher 22, 16–23 (1993)
17. Public Law 111-23-Weapon Systems Acquisition Reform Act of (2009)
18. Woodworth, R.S.: Experimental psychology. Henry Holt and Company, New York (1938)
19. Hoffman, R.R.: Accelerated Proficiency and Facilitated Retention: Recommendations Based on An Integration of Research and Findings from a Working Meeting. Report on Grant FA8650-09-2-6033, U.S. Air Force Research Laboratory, Wright-Patterson AFB OH (2010)
20. Newell, K.M., Hancock, P.A.: Forgotten moments: Skewness and kurtosis are influential factors in inferences extrapolated from response distributions. Journal of Motor Behavior 16, 320–335 (1984)
21. Still, D.L., Temme, L.A.: Configuring desktop helicopter simulation for research. Aviation Space and Environmental Medicine 77, 323 (2006)
22. Woods, D.D.: Engineering organizational resilience to enhance safety. Presentation at the Eighth International Conference on Naturalistic Decision Making, Asilomar Conference Center, Monterey, CA (June 2007)
23. Siegel, S.: Non-parametric statistics for the behavioral sciences. McGraw Hill Book Company, New York (1956)
24. Amin, R.W., Li, K.: The EWMA maxmin tolerance limits. The International Journal of Quality and Reliability Management 17, 27–41 (2000)
25. Patil, G.P.: On the evaluation of the negative binomial distribution with examples. Technometrics 2, 501–505 (1960)
26. Hill, L.B., Rejall, A.E., Thorndike, E.L.: Practice in the case of typewriting. Pedagogical Seminary 20, 516–529 (1913)