INL/CON-07-12780
PREPRINT

The Measure of Human Error: Direct and Indirect Performance Shaping Factors

8th IEEE Conference on Human Factors and Power Plants and 13th Conference on Human Performance, Root Cause and Trending (IEEE HFPP & HPRCT)

Ronald L. Boring
Candice D. Griffith
Jeffrey C. Joe

August 2007

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

The Measure of Human Error: Direct and Indirect Performance Shaping Factors

Ronald L. Boring[1], Candice D. Griffith[2], Jeffrey C. Joe[1]

[1] Human Factors, Instrumentation and Control Systems Department, Idaho National Laboratory, Idaho Falls, Idaho, USA
{ronald.boring, jeffrey.joe}@inl.gov
[2] Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, Tennessee, USA
candice.d.griffith@vanderbilt.edu

***Abstract*—The goal of performance shaping factors (PSFs) is to provide measures to account for human performance. PSFs fall into two categories—direct and indirect measures of human performance. While some PSFs such as "time to complete a task" are directly measurable, other PSFs, such as "fitness for duty," can only be measured indirectly through other measures and PSFs, such as through fatigue measures. This paper explores the role of direct and indirect measures in human reliability analysis (HRA) and the implications that measurement theory has on analyses and applications using PSFs. The paper concludes with suggestions for maximizing the reliability and validity of PSFs.**

I. INTRODUCTION

Performance shaping factors (PSFs) encompass those influences that enhance or degrade human performance. PSFs are used within human reliability analysis (HRA) methods to identify contributors to human errors and to provide a basis for quantifying those contributors systematically. While completing an HRA, an analyst may review a list of possible PSFs to identify possible sources of human error. The analyst may subsequently use predefined error rates associated with specific PSFs to determine a human error probability for a given task or situation.

Within HRA, PSFs are often categorized as internal or external, corresponding to the individual vs. situational or environmental circumstances, respectively, that bring to bear on performance. To date, the research literature has not addressed the consideration that PSFs fall into two categories—*direct* and *indirect* measures of human performance. While some popular PSFs such as "time needed to complete a task" are directly measurable, other PSFs, such as "fitness for duty," can primarily be measured indirectly through other measures and PSFs, such as through fatigue measures. This paper explores the role of direct and indirect PSFs in HRA and the implications that measurement theory has on analyses and applications using PSFs. The paper draws analogs to measurement as used in the physical sciences. It concludes with a discussion of the implications of direct and indirect PSFs on reliability and validity, using Fitness for Duty as a case study, and provides specific guidance to enhance reliability and validity when using direct and indirect PSFs.

II. A REVIEW OF DIRECT AND INDIRECT MEASURES

PSFs normatively measure the degree or magnitude of an effect on performance. Broadly speaking, magnitude is the measurable, countable, or comparative quality of something [1]. Magnitude reflects a continuous quantum rather than discrete, categorical membership.[1] It also serves as the basis of the empirical physical sciences, which have developed sophisticated methods to quantify physical magnitudes through measurement [3]. Galileo set the early stage for the importance of measurement by declaring that the goal of science was to "measure what is measurable and to try to render measurable what is not yet so" (cited in [4], p. 181).

Berka [4] accords the following characteristics to the materialistic measurement used in the physical sciences (p. 182-3):

1. Measurement is ontologically committed (i.e., rooted in and, hence, grounded by objective reality.
2. Magnitudes are historically and theoretically determined reflections of quantitative aspects of objectively existing entities and not merely outcome of metricization or measuring procedures.
3. The object of measurement exists prior to metricization or measuring procedures.
4. In agreement with the historical determination of every phenomenon, a transfer of methods from one universe of discourse into another one is adequate only on the objective condition that certain structural similarities hold between the domains in question.

These points are derived from Berka's attempt to define measurement as currently used in the physical sciences, not from an a priori historical formalism that has guided measurement in practice. These four axioms roughly translate to mean that measurement is based on physical magnitude

[1] A measurement scale may not always reflect this continuum with pinpoint precision. Underdeveloped or difficult-to-quantify measurement scales may follow an ordinal scale, while well established measurement scales tend to provide interval or ratio scaling [2].

dimensions and that measurements should be a type of natural reflection of physical magnitudes, not a contrived formulation only made possible by complex measurement instruments.

Historically, the *conservative conception of measurement* [5] first espoused by Helmholtz [6], considers measurement as the one-to-one correspondence of a physical property to a real number. For example, a metal rod of a given length might be used as one unit of length measurement. This rod would correspond to a numeric value of 1 in terms of measurement. A second rod of equal length placed adjoining the first rod would equal two units of that measurement. When presented with a novel object, the metal rods may be used to measure the length of the new object. In the conservative conception of measurement, there is always a direct relationship between the magnitude dimension of an object and a real object that represents a numerical quantity.

The conservative conception of measurement encompasses physical magnitudes such as length, weight, angle, and the like [5]. It, however, fails to account for certain magnitude dimensions that cannot be directly linked to a physical object. For example, temperature remains an elusive magnitude to measure directly. Instead, temperature must be measured indirectly. Since it is known that objects expand and contract relative to temperature, it is possible to use this expansion and contraction in a lawlike manner to measure the effect of temperature on an object. Nonetheless, it is not possible to measure temperature directly. A conventional thermometer actually measures the height of a temperature-sensitive fluid like mercury contained in a thin, long, translucent tube. As such, conventional measures of temperature are simply measures of the length of a fluid. The measurement relationship between temperature and underlying physical units remains indirect.

In order to account for the necessity of indirect measurements of physical magnitudes, more recent formulations of measurement theory use a *liberal conception of measurement* [5]. In this formulation, numbers follow a specified functional relationship to magnitudes. The liberal conception of measurement affords a more flexible view that accommodates the necessity to measure certain magnitude dimensions according to other magnitude dimensions. Conservative and liberal—direct and indirect—measures are summarized in Table 1.

TABLE I. DEFINITION OF DIRECT AND INDIRECT MEASURES.

Direct Measures	Indirect Measures
One-to-one relationship between magnitude of something and its physical, measurable properties	Relationship between magnitude of something and its physical, measurable properties can only be determined by its effects on something else

An important concept in measurement in the physical sciences centers on the multidimensionality of measurement for any given object. As Kyburg [7], p. 17, notes:

> Measurement is often characterized as the assignment of numbers to objects (or processes). Thus we may assign one number to a steel rod to reflect its length, another to indicate its mass, yet another to correspond to its electrical resistance, and so on. It is thus natural to view a quantity as a function whose domain is the set of things that quantity may characterize, and whose range is included in the set of real numbers.

Any given object has a multitude of magnitude dimensions in which it may be measured. While in many cases these magnitude dimensions may be orthogonal, they are often interrelated. For example, the 1889 definition of the magnitude of a meter was defined by the International Bureau of Weights and Measures to be equivalent to a graduated platinum-iridium cross section at 0º C [8]. Note that the fidelity of the measurement depended on temperature, another magnitude dimension. More recently, the 1983 definition of a meter is "the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second," where the speed of light is 299,792,458 m/s and the light is defined as a helium-neon laser with a wavelength equal to 632.99139822 nm (cited in [8]). The current definition of the length of a meter is thus measured in terms of precisely defined magnitude measurements of time and light wavelength.

The physical sciences exercise a seemingly increasing enthusiasm for measuring physical magnitudes according to interrelated dimensions. The intention of the increasing multidimensionality of standardized measurements is not to obfuscate or to walk a precariously close line to recursion. Rather, these multidimensional measurements serve to minimize the variability in measurement. Whereas a physical object such as a rod made out of platinum-iridium might be subject to fluctuations beyond those accounted for by temperature, a wavelength of a burst of light measured in time brings a higher constancy to the measurement standard. Increasing the constancy of the standard ensures that measures made on physical magnitudes accurately reflect the characteristics of those magnitudes. The precision of empirical laws is necessarily limited by the noisiness of magnitude measurements. Hence, the goal of science is to achieve the highest measurement constancy and fidelity—also known as reliability and validity—that are possible. Accurate measurement criteria—even if they are indirect and multidimensional—are the most parsimonious.

TABLE II. DEFINITION OF DIRECT AND INDIRECT PSFS.

Direct PSFs	Indirect PSFs
Those PSFs that can be measured directly, whereby there is a one-to-one relationship between the magnitude of the PSF and that which is measured	Those PSFs that cannot be measured directly, whereby the magnitude of the PSF can only be determined multivariately or subjectively

III. DIRECT AND INDIRECT PSFs IN HRA

Table 2 expands the definitions of direct and indirect measures to encompass direct and indirect PSFs. Current HRA erroneously treats all PSFs as direct measures of human performance or fails to consider the implications of direct and indirect measures for HRA identification and quantification. It is appropriate and necessary to reconsider PSFs in light of

indirect measurement and to treat resultant data in accord with the conservative and liberal conceptions of measurement and their corresponding limitations and benefits. Equally importantly, it is necessary to codify the relationship of indirect PSFs to measurable properties of human performance, rather than to maintain a loose, informal definition for such PSFs.

Put differently, there are numerous sources of measurement error associated with PSFs, and we hypothesize that the primary causes for these errors is in not distinguishing between direct and indirect measures and not understanding what the implications and limitations are for using either type of measure. For example, there is the potential for a direct measure (e.g. reaction time) to be misapplied or over-generalized, and the potential for indirect measures (e.g. temperature) to inaccurately measure the relationship of a physical property to a real number. There is also the potential confounding of interacting and overlapping constructs within and between PSFs (e.g., time required may change due to stress levels). The remainder of this paper will focus on measurement errors in PSFs that result from the problems inherent in using indirect and direct measures.

Table 3 provides an expert classification of *HRA Good Practices* [9] PSFs as direct or indirect. Two issues are apparent in this classification:

1. In several cases, the assignment of an indirect PSF level requires making subjective judgments. Subjective judgments are commonly multivariate, drawing on other direct or indirect measures synthesized through cognitive processes that may not be transparent even to the person making the judgment. Without clear criteria for making such judgments, one person's judgment may vary considerably from that another or even from his or her own judgment on another occasion [10].
2. Several of the direct PSFs feature Boolean levels of assignment. While this categorization facilitates ready assessment of PSFs in many HRA methods, it fails to capture the continuous quantum essential to magnitude measurement. The absence of measurement grades for these PSFs can, in many cases, be resolved through the development of new, more nuanced measures, including fine tuned indirect measures.

These points are not seen as an indictment on the PSFs used in HRA. We acknowledge that PSFs as currently used serve as effective tools for identifying, quantifying, and ultimately mitigating contributions to risk. However, as new HRA methods are developed and existing HRA methods are refined, it is useful to consider possible sources of measurement error. Measurement errors can stem from a failure to properly consider the limitations of direct vs. indirect measures.

IV. FITNESS FOR DUTY AS A CASE STUDY

Fitness for Duty (FFD) is a PSF that has the potential for numerous kinds of measurement error. FFD is defined as [11]:

TABLE III. COMMON PSFs CLASSIFIED AS DIRECT OR INDIRECT.

PSF	Direct/Indirect?
Training and experience	*Direct.* Training levels can be directly measured (e.g., number of hours in simulator), as can experience (e.g., years worked on reactor). There is an implied relationship between training/experience and competence. While training and experience can be measured directly, their effect on competence is indirect.
Procedures	*Direct.* On a Boolean (true/false) level, one could discuss whether the procedure addressed the required steps, but this does not resolve the quality of procedures, which involves subjective judgment. *Indirect.* There is no direct scale for the quality of procedures. Even if a scale were devised by which to grade the quality of procedures objectively, it would be a reflection of indirect aspects that team together to create quality criteria.
Availability of instrumentation	*Direct.* This can be measured directly on a Boolean scale.
Time available	*Direct.* There are plant models and regulatory requirements that prescribe how long a plant condition, if left unchecked, can proceed before it results in core damage or other undesirable states. Similar criteria exist in other safety critical industries.
Complexity	*Indirect.* There are various complexity models and scales, but they determine complexity through a multivariate concatenation of other factors.
Workload, time pressure, stress	*Indirect.* Workload is usually measured by the number of simultaneous tasks and is sometimes coupled with complexity. Time pressure is often coupled with stress or even with time available. (While time available can be measured directly, the pressure that the individual feels as a result of time limits is only indirectly measurable.) Stress has good direct measures such as physiological measures.
Team/crew dynamics	*Indirect.* As with complexity, there are several scales to measure team dynamics, but these are of the indirect variety.
Available staffing and resources	*Direct*. It is possible in Boolean fashion or as a ratio to determine the number of required people for a task vs. the number of people who are on the job.
Ergonomic quality of Human-System Interface	*Indirect.* A number of ergonomic standards exist, but these are not typically designed to provide a quality level [12], or, if so, the level is multivariate and indirect.
Environment	*Indirect*. There is no single measure of the quality of the environment. It is a composite of a number of directly measurable factors such as temperature and noise level. Considered in isolation, these would be direct measures.
Accessibility and operability of equipment	*Direct.* It is possible to make a Boolean judgment on the availability of equipment. If multiple equipment or quality of the equipment is considered, however, the PSF becomes indirect.
Need for special tools	*Direct.* For example, the need to put on protective clothing. This is a Boolean judgment on the requirement for such tools.
Communications	*Indirect.* There are numerous communications measures, but they build on a number of indirect factors.
Special Fitness needs	*Indirect.* Commonly referred to as "Fitness for Duty." It is multivariate. Some individual measures like blood alcohol level are direct if considered in isolation of other factors.
Consideration of realistic accident sequence diversions and deviations	*Indirect.* This PSF refers to the build up of expectations for how the situation will proceed and is related to the operator's experience and how directly that maps to what the current situation is doing. This PSF requires considerable expertise and subjective judgment to assess.

> …whether or not the individual performing the task is physically and mentally fit to perform the task at the time. Things that may affect fitness include fatigue, sickness, drug use (legal or illegal), overconfidence, personal problems, and distractions. Fitness for duty includes factors associated with the individual, but not related to training, experience, or stress.

A further decomposition of FFD groups its effects as psychological (e.g., mental fatigue or personal problems) and physiological (e.g., muscle fatigue or sickness). This duplicity adds to the complexity of measuring this PSF.

FFD is fundamentally a challenging PSF to measure because it has both direct and indirect measures for the constructs and components included in its definition. For the fatigue construct in FFD, there are direct and indirect measures. For example, when examining fatigue's psychological performance effects, an example of an indirect measure of fatigue is reaction time (i.e., psychological fatigue is assumed to slow reaction time). There is no yet established direct measure of psychological fatigue. When the effect of fatigue is considered in the physiological domain, the direct measures are physical abilities while the indirect measures are subjective self reporting and expert judgment. For the sickness construct, the direct measures are physical tests (e.g. body temperature), and the indirect measures are self reporting and expert judgment. For drug use, the direct measures are drug tests, and the indirect measures are, again, self reporting and expert judgment.

The direct and indirect measurement problems with each of the constructs in FFD are as follows:

- Fatigue could be measured through the direct measure of hours worked, but this may prove a problematic measure, even though it is direct and objective. While it is easy to measure how many hours have lapsed between work shifts, this measure does not guarantee that a worker will not be fatigued. Such a measure does not, for example, measure what that individual did during his or her time off.
- Fatigue has the indirect measure of reaction time (among others such as vigilance), which may also prove problematic. Even when measuring a phenomenon with a direct measure such as time, it does not mean that the phenomenon is directly measurable. Reaction time is a direct measure in that it is simply a measure of how much time it takes to generate a behavioral response to a stimulus, but it is not possible to map reaction time in a one-to-one relationship with cognitive ability. Rather, inferences from people's reaction time are generalized to broader concepts such as cognitive function. Numerous psychological experiments, however, have consistently demonstrated that reaction time is related to stable psychological attributes (i.e., memory, perception, attitudes) such that it is possible to infer the effects of experimental variables on more complex cognitive constructs.

  For example, increases in reaction time are typically indicative of increased mental fatigue [13], but there are many other factors that affect a person's reaction time. These extraneous factors confound the relationship between reaction time and performance by introducing measurement noise or uncertainty. Such a lack of direct measurability is a significant source of error when using the FFD PSF to quantify performance decrements.
- The sickness component of FFD has direct measures found in various medical tests, but these are problematic, because only specific conditions can be tested. Typically it is necessary to look for these conditions ahead of time, making it difficult to use sickness in a retrospective analysis.
- Sickness may be judged indirectly by experts, but such a measure is problematic because it is inherently subjective and may reflect considerable variability between experts [14]. Also, self reporting could be considered a possible measure of sickness, but such a measure is subject to underreporting or optimistic self-assessment of one's abilities in the face of diminished health.
- Another component of FFD—drug use—has the direct measure of the drug level in blood tests. However, such tests may not always represent a perfect mapping between the level of the intoxicant and the degree of degraded performance. For example, there is widespread acceptance that blood alcohol content (BAC) tests measure the amount of alcohol present in the blood stream of the individual and these results correspond to degraded performance (i.e., the BAC test is measuring a physically understandable and directly quantifiable attribute of performance). The exact correlation between BAC and performance remains a topic for legal debate, but there is little debate that alcohol affects a person both physiologically and psychologically but each person slightly differently.
- Drug use may be judged indirectly by experts, the shortcomings of which are documented above under the sickness component of FFD.
- The last three constructs of FFD are overconfidence, personal problems, and distractions. These constructs have no real direct measures and must be measured through indirect measures such as self-reporting and expert judgment.

We recognize that FFD also has potential confounds in the constructs that make up its definition. For example, it is possible that personal problems (e.g., going through a divorce) are related to or are causing drug use (e.g., excessive alcohol or drug use). This lack of orthogonality between definitional constructs in FFD may contribute to additional sources of measurement errors that are beyond the scope of this paper. For a discussion of issues endemic to confounded definitions in PSFs, see [15].

V. CONSEQUENCES OF DIRECT AND INDIRECT PSFS IN HRA

The consequence or effect of not distinguishing between direct and indirect PSFs in HRA is that the measures that are developed as proxies (e.g., reaction time) for the theoretical

constructs of interest (e.g., fatigue) tend to have poor validity and low reliability. Validity refers to the degree to which inferences can legitimately be made from proxy measures (e.g., number hours off between work shifts) to the theoretical constructs on which those proxy measures were based (e.g., fatigue). Validity is a performance criterion that assesses how well one can generalize from their proxy measures to the concept or concepts underlying the measures. For example, when an HRA method develops a measure for the "Training and experience" PSF, (e.g., a passing grade on a final exam), how can one be assured that the proxy measure used is really measuring whether the individual is adequately trained (i.e., how do we know cheating did not occur)? Similarly, just because an operator has spent a certain number of qualifying hours in a training simulator does not necessarily mean he or she is equally trained as another operator, who has spent the same number of hours in the simulator. A multitude of possible factors may have affected the effectiveness of the training. For example, the first operator may have been mentally fatigued during his or her simulator run, while the second operator was not fatigued.

As a result of poor construct validity, the reliability of the measure may also prove to be an issue. Reliability in the most general sense is, "The degree to which test scores [proxy measures] are free from errors of measurement" ([16], p. 19) but is more specifically related to the consistency, repeatability, and stability of the proxy measures over time and across different applications and contexts. In other words, the measure for the PSF becomes unreliable because there are systematic and unsystematic uncertainties (i.e., sources of measurement error) associated with how that measure is used.

VI. Ensuring Reliability and Validtiy

There are numerous ways to ensure reliability and validity of proxy measures of constructs or components that make up PSFs like FFD. A straightforward way is simply to test and then improve the proxy measure's ability to predict future occurrences of the construct to which it is related. Acceptance criteria need to be established for what probability above chance the proxy measure needs to achieve in order to be considered an accurate predictor of future instances of the construct of interest. The extent to which the proxy measure explains or predicts future occurrences of the construct of interest consistently over time, across domains, and individuals is the extent to which it can be argued that it is valid and reliable. It is not clear to what extent this has been done for any of the proxy measures for PSFs.

A more involved form of improving validity is formally called internal-structure analysis. Internal-structure analysis is a process of triangulation of multiple proxy measures for a single construct. For a construct like fatigue, where there are multiple possible proxy measures (e.g., reaction time and percent correct on a vigilance task [17]), one way of determining if those measures are valid and reliable is if they "hang together" over time, across domains, and individuals. For example, when an individual is given a rigorous physical routine and asked to perform a number of difficult mental calculations, his or her measured reaction time should increase and percent correct on the vigilance task should decrease. This inverse relationship between reaction time and percent correct needs to be stable over time, across domains, and individuals in order for the proxy measures to be considered valid and reliable. Formal assessment of this is accomplished through the common statistical technique called factor analysis [18].

There is also a validation process called cross-structural validation. Cross-structural validation is concerned with determining whether a proxy measure is unrelated to constructs that are considered to be theoretically different. For example, BAC shows good cross-structural validity because it is a good proxy measure for intoxication, and is unrelated to (i.e., not a good predictor of) the theoretically different construct of fatigue. Cross-structural validation of proxy measures is important in the context of the previous discussion of one-to-one correspondence of direct measures. While it is easy to become preoccupied with whether or not a proxy measure is actually measuring the construct it claims to be measuring (i.e., convergent validity), an equally problematic issue is when the proxy measure shows a correspondence to more than one construct. When a proxy measure has a one-to-many correspondence, it calls into question what construct the proxy measure is truly measuring. For example, reaction time is a proxy measure typically associated with measuring mental fatigue, but it is also be associated with intoxication (e.g., greater intoxication leads to slower reaction times), which is another component or construct of FFD. Fortunately for reaction time, the context in which it is used as a proxy measure usually establishes what construct it is meant to represent. When reaction time is used in a pre-post experimental design, whereby the individual is asked to perform a number of difficult mental calculations, it is evident that it is a proxy measure for mental fatigue. When reaction time is used in a pre-post experimental design where the individual is asked to consume copious amounts of alcohol, it is evident that it is a proxy measure for intoxication [19]. This delineation of the intent of the measure by the experimental design, however, is not always possible. Obviously, if one wanted to study the effects of intoxication and different types of difficult mental calculations on human performance, reaction time would not be a good proxy measure. Thus, cross-structural validation should not be ignored when developing measures of PSFs or constructs within PSFs. Cross-structural validation is also important in the context of addressing a potential fallacy in assuming constructs are different from each other just because they have different names or semantic definitions, when a proxy measure may indicate that they are more similar than originally believed. More information on cross-structural analysis can be found in [20].

VII. Good Practices for Direct and Indirect PSFs

In light of the previous discussions, we conclude with a preliminary set of good practices for conducting analyses with direct and indirect PSFs. Note that these considerations are not

exhaustive. Further good practices for the use of direct and indirect PSFs will follow as greater experience at distinguishing and utilizing these measures is gathered by methods developers and practitioners.

- *Utilize PSFs that are compatible with good measurement practices.* It is beneficial within an HRA analysis to incorporate PSFs that have clear definitions, offer a tractable corollary to that which is measured—human performance (including PSFs that not only account for deleterious effects on performance but also those that enhance performance), and offer a measure of continuous quantum.
- *Pick the best available PSF, whether direct or indirect.* When there is a choice of direct or indirect PSFs, it is important to remember that a direct PSF is not inherently preferable to an indirect PSF. A valid and reliable indirect PSF that offers a continuous scale is, however, generally preferable to a direct PSF that only offers Boolean categorization.
- *Ensure the orthogonality of the definitional constructs.* Especially in the case of multidimensional indirect constructs of a single PSF such as FFD, the analyst should utilize measures that do not overlap, which would introduce the possibility of double-counting effects. Where possible, the singularity of a construct effect should be factored out in the development and use of the PSF. Likewise, it is important to ensure that the individual constructs offer a reasonably complete account of the phenomenon under investigation so as to eliminate unaccounted for spurious effects and corresponding measurement uncertainty.
- *Verify the validity of the PSF.* The analyst should ensure, informally or through formal structure analysis techniques, that the PSF measures what it purports to measure. Specifically, it is crucial that the PSF as measured corroborates the performance effect. The danger of invalid measures is greatest for indirect PSFs that rely solely on expert judgment or that incorporate complex, multifaceted constructs of a PSF.
- *Verify the reliability of the PSF.* The stability and generalizability of PSFs need to be carefully considered in an analysis of human reliability. In the case where expert judgment is used to select the appropriate level of the PSF, a PSF that does not adequately consider human decision making processes and biases could result in inconsistency in PSF assignment between analysts or even by the same analyst on a different occasion. In terms of generalizability, a PSF that is designed for a particular domain (e.g., nuclear power plant control room operations) may not generalize to another domain (e.g., aircraft piloting) in a manner that allows the analyst to use the PSF scale reliably.

VIII. Discussion

This paper has reviewed direct and indirect measures and their connection to PSFs in HRA. Treating direct and indirect PSFs interchangeably risks overlooking common sources of measurement error, namely by using the PSFs in an invalid or unreliable manner. By carefully considering direct and indirect PSFs and their respective strengths and weaknesses, the analyst maximizes his or her measurement prowess, ensuring that he or she is measuring what is intended and doing so in the most consistent manner possible. This process, in turn, has the potential to minimize those sources of measurement error that may introduce uncertainty into HRA.

Disclaimer

This work was supported by the US Department of Energy (DOE) under DOE Idaho Operations Contract DE-AC07-05ID14517. This article was prepared as an account of work sponsored by an agency of the US Government. The opinions expressed in this paper are those of the authors and not of an agency of the US Government. Neither the US Government nor any agency thereof, nor any employee, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product, or process disclosed in this publication, or represents that its use by such third party would not infringe privately owned rights.

References

[1] R.L. Boring and R.L. West, "Mind as magnitude: Reconsidering information processing in cognitive engineering," Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society, pp. 1826-1830, 2005.

[2] S.S. Stevens, Psychophysics. Introduction to its Perceptual, Neural, and Social Prospects, New York: Wiley, 1975.

[3] B. Ellis, Basic Concepts of Measurement, Cambridge, UK: Cambridge University Press, 1968.

[4] K. Berka, "Are there objective grounds for measurement procedures?" C.W.Savage and P. Ehrlich (Eds.), Philosophical and Foundational Issues in Measurement Theory, Hillsdale, NJ: Lawrence Erlbaum, pp. 181-194, 1992.

[5] C.W. Savage and P. Ehrlich, "A brief introduction to measurement theory and to the essays," C.W. Savage and P. Ehrlich (Eds.), Philosophical and Foundational Issues in Measurement Theory, Hillsdale, NJ: Lawrence Erlbaum, pp. 1-14, 1992.

[6] H. Helmholtz, Die Tatsachen in der Wahrnehmung: Zählen und Messen erkenntnis-theoretisch brachtet, Darmstadt: Wissenschaftliche Buchgesellschaft, 1887/1959.

[7] H.E. Kyburg, Theory and Measurement, Cambridge, UK: Cambridge University Press, 1984.

[8] W.B. Penzes, Time Line for the Definition of the Meter. Retrieved October 16, 2002, from the National Institute of Standards and Technology Web: http://www.mel.nist.gov/div821/museum/timeline.htm

[9] A. Kolaczkowski, J. Forester, E. Lois, and S. Cooper, Good Practices for Implementing Human Reliability Analysis (HRA), NUREG-1792, Washington, DC: US Nuclear Regulatory Commission, April 2005.

[10] R.L. Boring, "Improving human scaling reliability," Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, pp. 1820-1824, 2003.

[11] D. Gertman, H. Blackman, J. Byers, L. Haney, C. Smith, and J. Marble, The SPAR-H Method, NUREG/CR-6883, Washington, DC: US Nuclear Regulatory Commission, August 2005.

[12] R.L. Boring, T.Q. Tran, D.I. Gertman, and A. Ragsdale, "A human reliability based usability evaluation method for safety-critical software," Proceedings of the 5th International Topical Meeting on Nuclear Plant Instrumentation, Controls, and Human Machine Interface Technology (NPIC&HMIT), pp. 1275-1279, 2006.

[13] W. Choo, W. Lee, V. Venkatratraman, F. Sheu, and M. Chee, "Dissociation of cortical regions modulated by both working memory load and sleep deprivation and by sleep deprivation alone," NeuroImage, vol. 25, pp. 579-587, 2005.

[14] R.L. Boring, D.I. Gertman, J.C. Joe, L.G. Blackwood, H.S. Blackman, and B.M. Brady, "A simplified expert elicitation guideline," Proceedings of the 8th International Conference on Probabilistic Safety Assessment and Management," Paper PSAM-0089, pp. 1-9, 2006.

[15] W.J. Galyean, "Orthogonal PSF taxonomy for human reliabilty analyses," Proceedings of the 8th International Conference on Probabilistic Safety Assessment and Management," Paper PSAM-0281, pp. 1-5, 2006.

[16] American Psychological Association, Standards for Educational and Psychological Testing, Washington, DC: American Psychological Association, 1985.

[17] M. Thomas, H. Sing, G. Belenky, H. Holcomb, H. Mayberg, R, Dannals, H. Wagner Jr., D. Thorne, K. Popp, L. Rowland, A. Welsh, S. Balwinski, and D. Redmond, "Neural basis of alertness and cognitive performance impairments during sleepiness. I. Effects of 24 h of sleep deprivation on waking human regional brain activity," Journal of Sleep Research, vol. 9, pp. 335-352, 2000.

[18] J. Nunnally, Psychometric Theory (2nd Ed.), New York: McGraw-Hill, 1978.

[19] A., Williamson, and A. Feyer, "Moderate sleep deprivation produces impairments on cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication," *Occupational Environmental Medicine*, 57, pp. 649–655, 2000

[20] L. Cronbach and P. Meehl, "Construct validity in psychological tests," Psychological Bulletin, vol. 52, 281-302, 1955.

BIOGRAPHIES

Ronald L. Boring has an MA in Human Factors and Experimental Psychology from New Mexico State University and a PhD in Cognitive Science from Carleton University. He has published in a wide variety of human reliability, human factors, and human-computer interaction forums. His primary research emphasis since joining the Idaho National Laboratory has been in human reliability analysis as researcher and project manager on projects for the US Nuclear Regulatory Commission, the National Aeronautics and Space Administration, and the US Department of Energy. Prior to joining the Idaho National Laboratory, he worked as a usability engineer for Microsoft Corporation and Expedia Corporation, also working as a guest researcher in human-computer interaction at the National Research Council of Canada. He is currently on temporary assignment as a visiting scientist at Halden Reactor Project in Norway.

Candice D. Griffith holds an MS from Vanderbilt University and is currently working on her PhD at Vanderbilt University under Dr. Sankaran Mahadevan in the field of Human Reliability Analysis within the National Science Foundation Integrative Graduate Education, Research, and Training (IGERT) program on Risk and Reliability Management. Her research has been mentored by the US Nuclear Regulator Ccommission and Idaho National Laboratory. She has been involved in Human Factors projects for US Nuclear Regulatory Commission and the National Aeronautics and Space Administration.

Jeffrey C. Joe holds in MS in Social Psychology from the University of Utah. He has been a Human Factors research scientist at the Idaho National Laboratory for the last seven years. His research interests are in the general areas of human factors, social, and organizational psychology. Specific research areas of interest include: human performance, organizational influences on human performance, attitudes and attitude change, human reliability analysis, and decision-making. He has recently been the project manager for Human Factors research activities with the US Nuclear Regulatory Commission and the National Aeronautics and Space Administration.