INL/CON-07-12742
PREPRINT

Extracting and Converting Quantitative Data Into Human Error Probabilities

Joint 8th IEEE HFPP / 13th HPRCT

Tuan Q. Tran
Ronald L. Boring
Jeffrey C. Joe
Candice D. Griffith

August 2007

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

Extracting and Converting Quantitative Data Into Human Error Probabilities

Tuan Q. Tran[(1)], Ronald L. Boring[(1)], Jeffrey C. Joe[(1)], and Candice D. Griffith[(2)],

[(1)] Human Factors & I&C Systems Dept., Idaho National Laboratory, Idaho Falls, ID, USA.
[(2)] Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA.
Email: {tuan.tran, ronald.boring, jeffrey.joe }@inl.gov; candice.d.griffith@vanderbilt.edu

Abstract - **This paper discusses a proposed method using a combination of advanced statistical approaches (e.g., meta-analysis, regression, structural equation modeling) that will not only convert different empirical results into a common metric for scaling individual PSFs effects, but will also examine the complex interrelationships among PSFs. Furthermore, the paper discusses how the derived statistical estimates (i.e., effect sizes) can be mapped onto a HRA method (e.g. SPAR-H) to generate HEPs that can then be use in probabilistic risk assessment (PRA). The paper concludes with a discussion of the benefits of using academic literature in assisting HRA analysts in generating sound HEPs and HRA developers in validating current HRA models and formulating new HRA models.**

I. INTRODUCTION

In many current HRA methods, the quantification process begins with the user identifying the task of interest as well as determining the task's nominal human error probability (NHEP). The NHEP is simply the base-rate for an error to occur under normal operating conditions. For example, the Standardized Plant Analysis Risk HRA (SPAR-H) method places a NHEP of 0.01 (or 1E-2) for cognitively engaging tasks and NHEP of 0.001 (or 1E-3) for action-related tasks [1]. Thus according to SPAR-H, under normal operating conditions an individual has a 1 in 100 chance of committing a cognitive error while a 1 in 1000 chance of committing an action-oriented error. After determining the task NHEP, the user, typically a human reliability analyst, then identifies PSFs that are believed to affect the task performance. The PSFs are then quantified and used to modify (i.e., enhance or degrade) the calculated NHEPs to produce an HEP. Thus, identifying and quantifying PSFs are critical steps in the quantification process.

Thus, identifying and quantifying performance-shaping factors (PSFs) are critical steps in many human reliability analysis (HRA) methods. However, the lack of empirical data in a form that these HRA methods can use to base PSF values on has been challenging in terms of helping the HRA analyst derive sound human error probability (HEP) estimates and supporting HRA method developers in formulating and validating HRA models. *The lack of data* in a readily usable form for PSF estimation stems from two challenges: 1) scaling individual PSF effects from different empirical sources, and 2) partitioning the inter-relational effects between the different types of PSFs (as illustrated in Fig. 1 for SPAR-H).

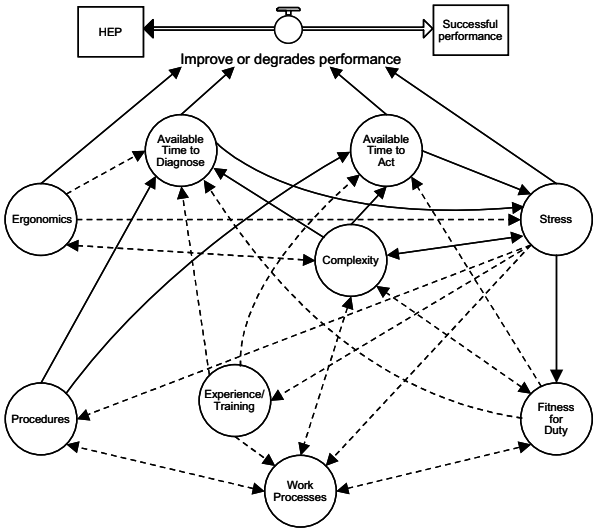

Figure 1. Relationships between PSFs in SPAR-H, where solid lines indicate a strong relationship and dotted lines indicate a moderate relationship (from Reference [1])

Scaling the magnitude of an individual PSF to a generic PSF is very difficult because it requires a well-controlled experiment in which the researcher manipulates the effect of an individual PSF on performance while holding all other PSFs constant (thus, preventing spurious effects and/or confounds). Only in an experimental design can one reveal the relative effect of a PSF on performance. While such requirements may be outside the practical realm of industry control, the academic literature is full of well-controlled experiments examining PSFs on a wide-variety of tasks. Unfortunately, the rich data in the academic literature has not been fully captured by HRA analysts and methods developers because most of the experimental results are not reported in terms of human error probabilities. More importantly, these experimental results tend to be reported in different scaling metrics (e.g., some results are reported in terms of means, chi-squares, proportions, correlations, t-statistics, F-statistics, etc.), thus making it difficult for an analyst to combine or average the results across different studies. Because of this, there is no standard outside expert estimation for an analyst to use in estimating the magnitude of PSF effects. Consequently, this can lead to inconsistent estimates among different users or methods. To further complicate the scaling issue, many PSFs are assumed to co-occur and in some cases, can exert influence on other PSFs in moderating human performance. To

date, the exact interrelationship between PSFs is unknown.

One solution to the scaling issue consists of compiling the database with relevant empirical literature on PSFs as well as scoping the statistical literature for an approach to convert different scaling metrics onto a common scale. At present, a database created by the Idaho National Laboratory (INL) has successfully documented a large number of empirical studies on PSFs. This allows the user to search the database for relevant PSFs that he/she is interested in exploring.

One statistical approach called *meta-analysis* (specifically, meta-analysis' statistics of effect size), not only allows users to convert different empirical results onto a common metric in scaling individual PSFs effects but also allows users to examine the complex interrelationships among PSFs. The remaining portion of this paper describes what meta-analysis is and how meta-analysis can assist human reliability analysts. It will also describe how meta-analysis results can be mapped onto the example HRA method of SPAR-H to generate HEPs that can then be used in PRA. Finally, a step-by-step example is provided to assist users in conducting a meta-analysis.

II. META-ANALYSIS AND ITS IMPLICATIONS TO SCALING PERFORMANCE SHAPING FACTORS

Meta-analysis is a method used to review research literature. Instead of conducting original research studies, meta-analysis uses the information that has already been collected in the literature by pooling and converting those different reported statistics onto a common metric (i.e., effect size), so that more generalizable results can be interpreted [2]. Like basic experimental designs, a meta-analysis has independent and dependent variables. The independent variables in meta-analysis are the reviewed studies that contain the variables of interest (in our case, PSFs) while the dependent measure(s) are the reviewed studies' outcome measures (e.g., response times or error rates). Importantly, meta-analysis can convert different reported statistics onto a common metric known as effect size (typically Cohen's *d*). The conceptual basis of effect size is analogous to Fischer's standardizeD (*Z*) scores, in that results are transformed into a standardized (*d*) metric scale [3]. That is, effect size is the term given to the degree of change in the variable under study. The degree of change relates to the phenomenon under investigation (e.g., stress), the degree that it is present in the population, and its expressed degree of difference from the null hypothesis, where the null hypothesis is that there is no effect.

Typically, effect size is calculated by subtracting the mean of the control group from the mean of the treatment group and divided by the pooled standard deviation of the two groups [2]. Thus, using Table. 1, an effect size of 0.7 between studies examining stress and no stress means that, on average, individuals in the stress studies performed better than 76% of the individuals in the no stress condition. Conversely, an effect size of -0.8 on studies examining stress and no stress means that, on average, individuals in the no stress condition performed better than 79% of the individuals in the stress condition. A more conventional and easy method to interpret effect size is to use Cohen's metric of 0.2 as "small effect size", 0.5 as "medium effect size", and 0.8 as "large effect size" [3]. Either way, effect size allows users to aggregate numerous studies independent of the original reported statistics (e.g., chi-square, F-test, t-test, means). For simplicity, this document focuses on Cohen's metrics.

TABLE 1. INTERPRETING EFFECT SIZE (TAKEN FROM REFERENCE[3]).

Effect Size (*d*)	%ile	Effect Size (*d*)	%ile	Effect Size (*d*)	%ile
0.0	50%	0.6	73%	1.4	92%
0.1	54%	0.7	76%	1.6	95%
0.2	58%	0.8	79%	1.8	96%
0.3	62%	0.9	82%	2.0	98%
0.4	66%	1.0	84%	2.5	99%
0.5	69%	1.2	88%	2.8	99.9%

The main strength of meta-analysis comes from its ability to summarize research findings across different studies. This procedure makes it possible to convert different statistical results onto a common scale as well as examining for underlying relationships between variables, such as moderator variables. Other strong points of meta-analysis are that it does not rely on sample size, can handle large numbers of studies, and protects against over-interpreting differences across studies. The next sections of this paper will focus on how to calculate effect sizes to determine relative effects of individual PSFs as well as guidance in examining the interrelationship between PSFs, and, finally, how Cohen's *d* can be mapped onto SPAR-H to produce HEPs.

III. CALCULATING EFFECT SIZE (d) TO DETERMINE RELATIVE EFFECTS OF INDIVIDUAL PSFs IN EMPIRICAL LITERATURE

Eq. 1 below is commonly used to convert means and standard deviations into effect sizes and uses the "pooled standard deviation" in Eq. 2. The reason for using the pooled standard deviation instead of the more familiar control group standard deviation is because many studies may lack a true control group or the control group sample size is relatively small. In such cases, the pooled standard deviation reflects the best estimate of the 'population' standard deviation [3].

$$ES_{sm} = \frac{(X_1 - X_2)}{s_{pooled}} \quad (1)$$

Where,
ES_{sm} = *standardized mean difference effect size*
X_1 = *mean of control*
X_2 = *mean of test*
s_{pooled} = *pooled standard deviation*

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 1}} \qquad (2)$$

Where,
ES_{sm} = standardized mean difference effect size
s_1^2 = variance of sample 1
s_2^2 = variance of sample 2
n_1 = control sample size
n_2 = test sample size
s_{pooled} = pooled sample deviation

There are cases in the database when both the means and standard deviations are not reported. Instead, the reported statistics are simply *F*-statistics or *t*-statistics. In such cases, it is recommended that the user use Eq. 3 and Eq. 4. When the reported statistics is chi-square, the analyst should use Eq. 5.

$$ES_{sm} = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}} \qquad (3)$$

Where,
ES_{sm} = standardized mean difference effect size
n_1 = control sample size
n_2 = test sample size
F = F-test value

$$ES_{sm} = t\sqrt{\frac{(n_1 + n_2)}{n_1 n_2}} \qquad (4)$$

Where,
ES_{sm} = standardized mean difference effect size
n_1 = control sample size
n_2 = test sample size
t = t-test value

$$ES_{sm} = 2\sqrt{\frac{\chi^2}{N - \chi^2}} \qquad (5)$$

Where,
ES_{sm} = standardized mean difference effect size
χ^2 = Chi-Square Value
N = total sample size

There may be cases when the user may want to aggregate two different statistics such as an *F*-statistics (i.e., analysis between dichotomous variables) and a correlation (i.e., analysis between two continuous variables). In such cases, the user is directed to follow the r to d transformation of Eq. 6.

$$d = \left[\frac{[2(r)]}{[(1 - r^2)^{.5}}\right] \qquad (6)$$

Where,
r = correlation coefficient

Finally, the user should be aware that in some cases, there may be only a small number of relevant PSF studies available in the database that the analyst is interested in, for example, organizational factors. When using only a small number of studies in a meta-analysis, the user must be aware of sampling error in that the pooled standard deviation is only an estimate and does not reflect the true population standard deviation [3]. To correct for sampling error, the user should use an unbiased estimate of Cohen's *d*. Specifically, the unbiased estimate of *d* is approximately equal to the calculated value of:

$$d \times \left(1 - \frac{3}{\{4(N_E + N_C) - 9\}}\right) \qquad (7)$$

Where,
N_E = test sample size
N_C = control sample size

IV. RELATIONSHIP BETWEEN PSFs

As discussed earlier, some PSFs are considered to be non-orthogonal. Because of this, users should be aware of studies in the database that examine two or more PSFs within a single experiment, since the relative effect of a single PSF on performance may be modulated by the other PSF(s). These indirect influences by other PSFs can be described as *moderators* and *mediators*. A moderator is a third variable that can affect the direction and/or strength between the primary variable of interest and the outcome measures. Thus, in terms of statistics, moderators lead to interactions [4]. For example, stress may have a relatively strong effect on performance during high workload but relatively weak effect during low workload. Whereas, mediation is where the primary variable affects a third variable directly, which then affects the outcome measure. For example, stress affects workload that then affects performance (see Fig. 2). Thus, when an analyst includes a study that consists of more than one PSF, a moderator/mediator analysis is recommended.

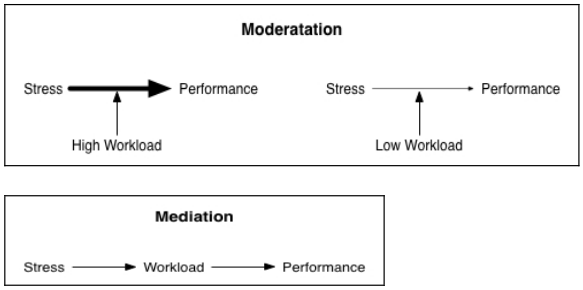

Figure 2. Examples of moderator and mediator effects

In examining moderation effects, the user is looking for a statistical interaction. Thus, a user can conduct numerous tests in examining statistical interactions depending on the PSF variable types:

- If the PSFs are all categorical, then use a chi-square statistic;
- If one PSF is categorical and another is continuous, then use an Analysis of Variance (ANOVA);
- If the PSFs are all continuous, then use regression analysis.

On the other hand, in mediation analysis, it has been recommended to use structural equation modeling. In fact, although moderation analysis can be examined with the above statistics, researchers have supported the structural equation modeling approach for both moderators and mediators [4]. Nevertheless, for simplicity, chi-square, ANOVA, and regression analysis are suitable statistics in examining moderators.

V. MAPPING COHEN'S d TO SPAR-H TO PRODUCE HEPS

In HRA methods (e.g., SPAR-H), the analysts are required to evaluate the magnitude of PSF effects. Because these PSF evaluations are used to modify NHEPs to obtain HEPs, it is critical that the reliability between analysts be high to ensure the reproducibility of results. The concern of low inter-rater reliability is always present when the evaluation is subjective in nature. The effect size approach can be useful here in aiding the analyst in evaluating the magnitude of the PSFs. As mentioned earlier, a conventional method to interpret effect size is to use Cohen's effect size metric of 0.2 as "small effect size," 0.5 as "medium effect size," and 0.8 as "large effect size." Analysts can use Cohen's effect size metric in scaling their PSF ratings. In some instances, Cohen's effect size metric can be directly mapped to a HRA PSF rating scales as shown in Table 2.

In other cases, analysts can use Cohen's effect size metric as a mean to gauge the appropriate level of PSFs. Thus, Cohen's effect size can be useful in increasing analysts' inter-rater reliability by bringing some standardization in the HRA methods of evaluating PSFs.

TABLE 2. EXAMPLE OF SCALING COHEN'S EFFECT SIZE METRIC TO SPAR-H PSF RATINGS

PSF	Level/Multiplier	Effect Size (*d*)
Stress/ Stressors	Extreme	0.8
	High	0.5
	Nominal	0.2
	Insufficient Information	
Complexity	Highly complex	0.8
	Moderately complex	0.5
	Nominal	0.2
	Obvious diagnosis	
	Insufficient Information	
Experience/ Training	Low	0.5
	Nominal	0.2
	High	0.8
	Insufficient Information	

In summary, this section has illustrated how an analyst can use the database to extract quantitative information on PSFs of interest, calculate effect sizes, and use Cohen's effect size metric to gauge appropriate PSF ratings within HRA methods to obtain HEPs. The next section illustrates a practical example how an analyst can use this method for HRA.

VI. ILLUSTRATION OF A META-ANALYSIS

Steps in conducting a meta-analysis for the purpose of obtaining HEPs are as follows. First, the user must select a PSF or PSFs he or she wants to examine in the literature. Then, the user then extracts the necessary quantitative information (e.g., means, *F*-statistics) from the literature to calculate effect sizes. Next, depending on whether the study examines two or more PSFs, the user should perform a moderator or mediator analysis to be certain that the calculated effect size value is valid. Finally, using Cohen's effect size metric, the user can gauge the appropriate scaling level of PSFs in a HRA method.

To give a practical example, (adapted from Reference [5]), one concern in spaceflight could be that astronauts do not sleep well, given their environment (e.g., environmental noises from instruments, weightlessness). Thus, an analyst may be interested in examining whether an individual's fatigue due to sleep loss can affect astronaut's performance and to what degree. Next, the user searches through the database using "fatigue" and "sleep" as guiding keywords and extracts relevant quantitative information to use in effect size calculation. Also during this stage, the analyst should make note of any other PSFs (e.g., workload) or other variables (e.g., gender) that were present in the selected studies, which he or she believes may mediate and/or moderate the PSF effect. To do this, it is recommended that the user construct a "data coding log" to organize different factors that he or she may want to use as a moderator or mediator variable. Figure 2 provides a simple example of a data-coding log.

After obtaining and logging the PSF quantitative information from the database, the user can use Eq. 1 through Eq. 7 to convert the quantitative information into effect sizes. Appendix A illustrates the results of the effect size calculation.

TABLE 3. EXAMPLE OF A DATA CODING LOG FOR AN INDIVIDUAL EMPIRICAL STUDY ENTRY

Study Quantitative Information	Analyst Input
Length of Sleep Deprivation	
Output Measure: (simple reaction time, error rates, other performance measures)	
Performance direction after test condition (e.g., better (+), worse (-))	
Sample size of control group and test group	
Type of stat used (means & Stdev, proportions or frequencies, significance test (e.g., t-test, F-value, chi-squared, p-value) quantitative results cannot be calculated)	

TABLE 3. EXAMPLE OF A DATA CODING LOG FOR AN INDIVIDUAL EMPIRICAL STUDY ENTRY (CON'T)

Study Quantitative Information	Analyst Input
Specialty field of sample source (e.g., astronauts, pilots, military, NPP workers, others)	
Average Hours of Sleep	
Experimental Design: (e.g., within vs. between group)	
Types of Experimental Conditions (e.g., kept active, deprived of caffeine, kept at lab, other)	
Other PSF presence (e.g., available time, stress/stressors, complexity, experience/training, procedures, ergonomics/HMI, work process)	

The meta-analysis performed on these data resulted in a overall effect size of -0.6341. Using Table 1 to determine effect size shows that the average individual in the non-sleep deprived condition performed better than 73% of the individuals in the sleep deprived condition. In terms of Cohen's effect size metric, sleep deprivation has a "medium effect" on performances.

Moderator and mediator analyses were not performed in this example because this meta-analysis was performed on a small sample of research articles on fatigue, and doing these analyses at this point would have been premature. Ideally, all relevant articles should be included in the meta-analysis before these ancillary tests are performed. Finally, the analyst can use Cohen's effect size metric to produce HEPs. For example, using SPAR-H, fatigue is a component of the "Fitness for Duty" PSF. As Table 4 shows, because SPAR-H uses only two levels of degraded performance for "Fitness for Duty," an analyst would map Cohen's "large" effect size to "Unfit," thus obtaining an HEP of 1.0 (failure). A "medium" effect size should be mapped to "Degraded Fitness"; thus, obtaining a multiplier of 5.0 for this PSF. If the calculated effect size was "low" on Cohen's metric, the analyst should rate "Fitness for Duty" as nominal and obtain a PSF multiplier of 1.0.

TABLE 4. FITNESS FOR DUTY PSF LEVELS IN SPAR-H

Fitness for Duty	Unfit	p(failure) = 1.0
	Degraded Fitness	5
	Nominal	1
	Insufficient Information	1

VII. CONCLUSION AND NEXT STEPS

Meta-analytic techniques may be use to aggregate or formulate new HEP values based on existing data. Cohen's effect size metric provides a systematic guide for analyst to use to evaluate the severity of PSFs on performance, thus leading to higher inter-rater reliability as well as reproducibility of results. Methods developers can examine the interrelationship between PSFs by setting each PSF as moderating or mediating variables. After calculating each PSF's effect size, a methods developer can perform a structural equation modeling analysis to examine the many interrelationships (i.e., moderating and mediating relationships) among PSFs.

VII. DISCLAIMER

IX. REFERENCES

[1] Gertman, D., Blackman, H., Marble, J., Byers, J., & Smith, C.. *The SPAR-H Human Reliability Analysis Method, NUREG/CR-6883.* Washington, DC: US Nuclear Regulatory Commission, 2005.

[2] Durlak, J.A. Understanding meta-analysis. In L.G. Grimm & P.R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics*. Washington, DC: American Psychological Association, 2005.

[3] Coe, R. *What is an "Effect Size?": A Guide for Users.* Available online at: http://www.cemcentre.org/renderpage.asp?linkID=30325016

[4] Shadish, W.R.J., & Sweeney, R.B. Mediators and moderators in meta-analysis: There's a reason why we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology, 59*(6), 1991, pp. 883-8.

[5] Griffith, C. & Mahadevan, S. Sleep deprivation effects of human Performance: A meta-analysis Approach. *Proceedings from the Eighth International Conference on Probabilistic Safety Assessment and Management (PSAM 8), 2006.*

Study Ref	Variable	Stat info	$M_{Control}$ (SD)	M_{Test} (SD)	Test Stat Info	$n_{control}$	n_{test}		ES d-index	Direction	Sign
	Speed										
Nilsson	Speed (ms)	means & SD	225 (43.44)	265 (62.35)	-	11	11	2GC[2]	-0.7444	Worse	-1
Thomas	Speed (response/min)	means & SD	61.4 (24.6)	71 (27.2)	-	17	17	RM[3]	-0.3702	Worse	-1
Choo	Speed (ms)	means & SD	552 (149)	668 (182)	-	12	12	RM	-0.6974	Worse	-1
Choo	Speed (ms)	means & SD	588 (162)	746 (271)	-	12	12	RM	-0.7077	Worse	-1
Choo	Speed (ms)	means & SD	617 (233)	718 (245)	-	12	12	RM	-0.4225	Worse	-1
Williamson	Speed (ms)	means no SD	489	540	-	39	39	RM	NA	Worse	-1
Chee	RT[1]- variability	t-test	-	-	t(13) = 2.2, p<.05	13	13	RM	-0.8629	Worse	-1
Kobbeltvedt	planning time	F-test	-	-	F(1,89) =71.12, p<.01	21	69	2GC	-2.1018	Worse	-1
	Accuracy										
Thomas	% correct	means & SD	92.3 (6.4)	95.5 (5.2)	-	17	17	RM	-0.5488	Worse	-1
Thomas	# correct/min	means & SD	57.5 (25.2)	68.3 (27.3)	-	17	17	RM	-0.4111	Worse	-1
Williamson	# misses	means no SD	0.36	3.1	-	39	39	RM	NA	Worse	-1
Angus	# correct vs. incorrect	F-test	-	-	F(8,40) = 37.55, p<.001	-	-	-	NA	Worse	-1
Angus	# correct/min	F-test	-	-	F(8,40) = 35.03, p<.001	-	-	-	NA	Worse	-1
Kobbeltvedt	procedural errors	F-test	-	-	F(1,88) = 5.34, p<.05	21	69	2GC	0.5758	*Better*	1
Kobbeltvedt	critical EoO[4]	Chi-Square	-	-	$\chi^2(1,89) = 9.43$, p<.01	21	69	2GC	-0.6842	Worse	-1
								Mean ES	*-0.6341*		

Figure 2. Summary of effect size calculations.