Cogn Tech Work (2014) 16:289–297
DOI 10.1007/s10111-014-0275-1

ORIGINAL ARTICLE

Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective

J. C. F. de Winter

Received: 23 January 2013 / Accepted: 1 April 2014 / Published online: 20 May 2014

Abstract Situation awareness and workload are popular constructs in human factors science. It has been hotly debated whether these constructs are scientifically credible, or whether they should merely be seen as folk models. Reflecting on the works of psychophysicist Stanley Smith Stevens and of measurement theorist David Hand, we suggest a resolution to this debate, namely that human factors constructs are situated towards the operational end of a representational–operational continuum. From an operational perspective, human factors constructs do not reflect an empirical reality, but they aim to predict. For operationalism to be successful, however, it is important to have suitable measurement procedures available. To explore how human factors constructs are measured, we focused on (mental) workload and its measurement by questionnaires and applied a culturomic analysis to investigate secular trends in word use. The results reveal an explosive use of the NASA Task Load Index (TLX). Other questionnaires, such as the Cooper Harper rating scale and the Subjective Workload Assessment Technique, show a modest increase, whereas many others appear short lived. We found no indication that the TLX is improved by iterative self-correction towards optimal validity, and we argue that usage of the NASA-TLX has become dominant through a Matthew effect. Recommendations for improving the quality of human factors research are provided.

This original paper is discussed in the commentaries available at: doi:10.1007/s10111-014-0276-0; doi:10.1007/s10111-014-0277-z; doi:10.1007/s10111-014-0278-y.

J. C. F. de Winter (✉)
Department BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
e-mail: j.c.f.dewinter@tudelft.nl

Keywords Human factors constructs · Situation awareness · Workload · Measurement theory · Representationalism · Operationalism

1 Controversy about human factors constructs

A notorious controversy in human factors science is whether constructs such as mental workload and situation awareness are scientifically credible. In several articles, Dekker and Woods (2002), Dekker and Hollnagel (2004), and Dekker et al. (2010) point out that human factors constructs do not live up to the ideal of natural sciences and "don't so much reflect but rather create a particular empirical world, which would not even exist without those constructs" (Dekker et al. 2010, p. 27). These articles argue that human factors constructs are folk models, merely substituting one label for another, and immune to falsification. Similarly, Flach (1995) contends that situation awareness is an all too "convenient explanation that the general population can easily grasp and embrace" (p. 155) and warns that one should not fall in the dangerous trap of regarding situation awareness as a causal agent that exists in the mind of the human operator.

In a response article, Parasuraman et al. (2008) argue that the criticism by Dekker and colleagues is unjust. Parasuraman et al. maintain that human factors constructs are not part of an empirical reality, and a construct such as workload, "is not a statement of fact" (p. 155) and therefore "falsifiability of the construct itself is a meaningless notion" (p. 155). However, they stress that human factors constructs are "viable" (p. 140), "quantifiable" (p. 141), can be "operationalized" (p. 143), and have "usefulness in prediction" (p. 155). Measurement of workload, for example, can provide useful information on operator skill

and strategy (see Table 5 for an illustration taken from Hancock 1996).

In this article, we show that the debate between Dekker and Parasuraman closely resembles earlier debates on the measurement of psychological concepts. For example, when reviewing the history of psychological measurement, Hand (2004) quotes Dawes and Smith (1985): "It is not uncommon for psychologists and other social scientists to investigate a phenomenon at great length without knowing what they are talking about. So it is with attitude. While 20,209 articles and books are listed under the rubric 'attitude' in the Psychological Abstracts from 1970 to 1979, there is little agreement about the definition of attitude and hence what aspects of attitude are worth measuring" (p. 509; see also Uttal 2008). Similarly, Underwood (1957) warns that psychological concepts can easily be misused when one assumes that they "imply an existence of a real state or process in the organism" (p. 212).

The aim of this study is to interpret the debate on human constructs in the light of earlier controversy in psychological measurement. Our focus is on the scientific struggles encountered by Stanley Smith Stevens, an experimental psychologist who in the mid-twentieth century was concerned with measuring concepts such as brightness and loudness. We suggest a resolution to the controversy in human factors constructs, and based on a lexical study of workload (and the NASA Task Load Index specifically), we provide recommendations for improving the quality of human factors research.

2 Stevens' psychophysics research

The history of measurement brings us back to Stanley Smith Stevens (1906–1973). Awarded his PhD in 1933, Stevens became a faculty member in Harvard's new Department of Psychology in 1936. He rose to the rank of professor of Psychology in 1946, and finally, to professor of Psychophysics in 1962 when the university accepted his request to be named as such (Marks 2006; Miller 1974; Teghtsoonian 2001). Similar to human factors researchers operationalizing their constructs (e.g. workload and situation awareness), Stevens operationalised constructs using an approach that seemed at odds with traditional measurement in the natural sciences.

One of Stevens' typical early experimental studies on hearing performed with his supervisor Edwin Boring (Boring and Stevens 1936) investigated what characterizes tones as bright. To define such an abstract notion as brightness, they used clever and carefully defined procedures. Specifically, they used a siren of 40-cm-diameter cut from 1-mm-thick cardboard and perforated with sector-shaped holes of 2 cm in radial dimension. An air blast was delivered from a tank at a pressure of 34 kPa, directed through a 5-m tube and a nozzle with a 0.4-cm hole. The nozzle orifice was positioned at about 0.6 cm from the face of the siren disc. The disc was rotated at a constant speed of 27.5 revolutions per second. An overview of the independent variables (positions of holes and angular sizes of holes) is provided in Table 1.

Table 2 shows comparative judgments carried out by a number of subjects. The results indicated that brightness and loudness varied unambiguously with linear velocity (i.e. all observations showed that 1A was judged brighter and louder than 1B and that 1B was judged brighter and louder than 1C). When the loudness of the comparisons was equated subjectively by adjusting the nozzle, two of the five subjects reversed their judgments. Based on a further analysis of frequency spectra made with a wave analyser, Boring and Stevens found evidence that tones that were regarded bright have relatively strong upper partials. The results further indicated that 2A was judged brighter and louder as compared to 2B and 2C. Regarding the ratio of holes and non-holes, it became clear that brightness does not structurally depend on this ratio. Based on their observations, Boring and Stevens concluded that brightness is a function of intensity and frequency of the stimulus.

By today's standards, the above study has evident weaknesses: it appears to be exploratory rather than confirmatory research and included no frequentist statistical analysis. However, in later works, Stevens developed refined statistical models that allowed for highly effective prediction. For example, in the method of magnitude estimation, stimuli were presented in random order and the participant was asked to assign numbers to the stimuli, which corresponded to his/her subjective impression. When the obtained numbers were plotted against stimulus intensity, a neat power function arose, later named the Stevens' power law. Stevens discovered that for each type of modality, there was a distinctive value of the exponent relating judgment to stimulus intensity (Stevens 1961; Teghtsoonian 2001).

3 The Ferguson committee

Stevens' approach fuelled controversy as to whether it is actually possible to measure psychological attributes in a scientific manner. Between 1932 and 1940, the British Association for the Advancement of Science addressed this question. The Association had appointed a committee of physicists and psychologists to "report upon the possibility of quantitative estimates of sensory events" (Ferguson et al. 1938, p. 277). One of Stevens' sensory scales, the scale of loudness, became the committee's direct target. The British physicist Norman Campbell was an influential

Table 1 Selection of stimuli used in the experiment by Boring and Stevens (1936)

Stimulus	Number of holes N[a]	Mean radius of hole (cm) R	Size of hole (°) H	Size of interval (°) I	Linear velocity (cm/s) V[b]	Frequency (holes/s) F[c]	Ratio H/I
1A	12	15.2	15	15	264	330	1.0
1B[d]	12	11.4	15	15	198	330	1.0
1C	12	7.6	15	15	132	330	1.0
2A	18	11.4	10	10	198	495	1.0
2B[d]	12	11.4	15	15	198	330	1.0
2C	9	11.4	20	20	198	247	1.0
3A	12	11.4	5	25	198	330	0.2
3B[d]	12	11.4	15	15	198	330	1.0
3C	12	11.4	25	5	198	330	5.0

1A, 1B, and 1C give variation of the mean linear velocity of the holes (a higher velocity is achieved by placing the holes more to the outside of the disc)

2A, 2B, and 2C give variation of frequency with H/I and V constant (a higher frequency is achieved by increasing the number of holes)

3A, 3B, and 3C give variation of H/I while keeping V and F constant (a higher H/I ratio is achieved by elongating the holes)

[a] Number of holes = 360/(H + I)

[b] Linear velocity = 2*pi*R*27.5 (apparently Boring and Stevens erred by a factor 10)

[c] Frequency = 360*27.5/(H + I) = N*27.5

[d] These stimuli are identical to each other

Table 2 Comparison judgements among stimuli, selected from Boring and Stevens' (1936) article

Stimulus	1A versus 1B	1B versus 1C	2A versus 2B	2B versus 2C	3A versus 3B	3B versus 3C
Brighter	7/7	7/7	7/7	6.5/7	4.5/7	2/7
Louder	7/7	7/7	5.5/7	3.5/7	6.5/7	0/7
Brighter[a]	5/6	6/6	4/6	5/6	3.5/6	2/6
Denser	4/4	4/4	2/4	4/4	–	–

For example, the "7/7" in the upper left corner of the table means that seven out of seven subjects judged sound 1A as brighter than 1B. 0.5 point is assigned for equality judgements

[a] In these measurements, loudness was equated subjectively for the louder sound, by adjusting the tube to the nozzle

author of the report. Campbell had previously described "fundamental" measurement as the assignment of numbers to represent properties of objects, where an order relationship and a process of addition are satisfied (Campbell 1920; Hand 1996). He pointed out that "any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is in fact meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Ferguson et al. 1940, p. 245, see also Stevens 1946). In their final report (Ferguson et al. 1940), Campbell and the committee concluded that because psychological attributes do not allow additive (concatenation) operations, psychological attributes are not quantitative and therefore not scientifically credible.

4 Stevens' answer to the committee

Stevens met the critique of the Ferguson committee by developing a theory broad enough to cover all forms of measurement. He based his definition of measurement on Campbell: "Measurement … is the assignment of numerals to represent properties" (Campbell and Jeffreys 1938, p. 126). But, inspired by Bridgman's operationalism (Bridgman 1927) and the principles of logical empiricism, Stevens paraphrased this definition: "We may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules" (Stevens 1946, p. 677). In other words, "provided a consistent rule is followed, some form of measurement is achieved" (Stevens 1959, p. 19). Thus, contrary to Campbell's definition, which left no room for measurement outside natural sciences, Stevens' broader definition allowed for measurement in the psychological sciences as well.

Stevens argued that psychologists have other operations than Campbell's concatenation structures, and he extended the concatenation scales to, for example, systems that satisfied ordinality but not concatenation (Hand 1996). In a seminal paper, Stevens (1946) created a taxonomy that identified four classes of scales: nominal, ordinal, interval,

and ratio. He determined which statistical procedures are permissible for each scale in order to leave the scale form invariant: nominal scales allow for summary statistics, such as the mode and chi-square test; ordinal scales allow for percentiles, rank order correlations, and sign tests; interval scales allow calculation of means, standard deviations, *t* tests, and analyses of variance; finally, every type of statistical operation is appropriate for ratio scales. Additionally, he proscribed that higher ranked scales are more restricted concerning invariant transformations. A nominal scale admits any one-to-one substitution of labels assigned to the categories; an ordinal scale allows transformations by any monotonic increasing function; an interval scale is restricted to linear transformations; and a ratio scale admits only multiplication by a constant (Stevens 1946). Stevens' ideas about measurement have since evolved into what is called representational measurement theory (Hand 1996).

Stevens' taxonomy and associated statistical rules were readily adopted by textbook writers (e.g. Siegel 1956 cited 47,603 times in Google Scholar as of 14 March 2014). However, Stevens' measurement scales provided no satisfactory resolution to the controversy as to how to handle psychological constructs. Khurshid and Sahai (1993) offered a bibliography with as many as 300 selected references about Stevens' levels of measurement and their relationship to statistics. One of the most prominent sources of controversy is the distinction between ordinal and interval scales. Considering that psychological constructs are often seen as reflecting an ordinal scale, is it then meaningful to report, for example, a mean brightness score, or is it only meaningful to report a median brightness score? Discussion took the form of a pro-Stevens camp warning about "meaningless statistics" (Marcus-Roberts and Roberts 1987, p. 383), and an anti-Stevens camp arguing that Stevens' statistical rules are too restrictive.

A proponent of Stevens' scales may argue that "a key feature of measurement is that it serves to represent relationships between objects by relationships between numbers" (Hand 2005, p. 81) and that "only certain statistical operations are meaningful" (Hand and Keynes 1993, p. 315). A more liberal statistician may argue: "Approaches to statistics that start from an a priori scale type and then proscribe the kinds of hypotheses that may be considered or the statistical methods and tests that may be computed based on that scale type are simply bad science and bad data analysis" (Velleman and Wilkinson 1993, p. 70). Frederic M. Lord (1912–2000), by some regarded as the Father of Modern Testing, provided a satirical piece against the adoption of Stevens' scales. In this work, Lord (1953) describes a fictitious psychometrics Professor X who sold football numbers using a vending machine. Strictly, the football numbers are only nominal numbers to designate players of a team. However, with the help of a statistician, who argued that "numbers don't remember where they came from" (p. 21), they used "illegal" statistical procedures (e.g. calculating averages, standard deviations, and *p* values) to prove that the vending machine had been tampered with. Subsequently, Professor X recovers from his nervous breakdown and no longer locks the door when he computes the means and standard deviations of his students' test scores.

5 Representational and operational measurement

Based on Stevens' definition of measurement, theorists generated a framework for representational or axiomatic measurement theory. Representational measurement theory offered axiomatic proof of the uniqueness of Stevens' four scale types (Luce and Suppes 2002). For many, Stevens' work received more serious attention only after representational measurement theory was established (Townsend and Ashby 1984). As Hand (2004) explained, in representational measurement, one starts with sets of objects (e.g. rocks) with attributes that can be intrinsic (e.g. rock weight) or extrinsic (e.g. rock speed). The relationships connecting objects to their attributes constitute an empirical relational system. Representational measurement maps this reality to an idealized numerical relational system; that is, it represents the relationships between objects by relationships between numbers.

The operational aspect of measurement seeks to predict, with no reference to an underlying reality, and therefore, any numerical operation may be carried out on the numbers. From an operational (pragmatic) perspective, the numbers are chosen on external grounds, such as practical convenience or presumed theoretical relationships (e.g. predictive and construct validity). According to this perspective, the attributes are defined by the measurement method and the instruments and tools offer the objects their properties and thus their definition as well.

6 Resolution of the conflict: human factors constructs as operational measurement

A resolution to the conflicts in measurement has recently been offered by Hand (1996, 2004), who proposes that representational measurement theory does not describe all measurement activities. Instead, measurements lie on a continuum between representational and operational (pragmatic) measurement. The representational–operational continuum, which underlies all measurement, resolves the apparent disagreement and confusion.

In psychology, sociology, economics, or medicine, the concepts that are measured are often ill-defined and not properly tied to an empirical relational system. Whereas measurement of length or mass sit towards the representational end of the continuum, measurement using a quality of life scale, for example, finds itself close to the operational end (Hand 2005). As Hand (2005) indicates, "merely because a procedure may have weak representational aspects does not mean it is inadequate" (p. 83). Operational measurements procedures, such as a quality of life scale, might be useful to predict the risk of suicide.

The operational measurement seeks to predict, and this can be achieved even without understanding of an underlying mechanism (Hand 1996). Therefore, an operationalist procedure for measuring quality of life will yield a useful measurement, even though it does not reflect an empirical reality. The key in operationalism is that the measurement procedure simultaneously defines concepts and measures them. Hand (1996) states, "an attribute is defined by its measuring procedure, no more and no less, and has no 'real' existence beyond that. In operationalism the attribute and the variable are one and the same" (p. 453). Therefore, special care should be taken towards the construction of the measuring instrument and the precision of the definition.

7 Lexical study: questionnaires for measuring workload

We apply the above ideas to the Dekker–Parasuraman debate about human factors constructs. We argue that human factors constructs, such as workload and situation awareness, are strongly situated to the operational end of the representational–operational continuum. If we follow this assumption, human factors constructs do not reflect an empirical reality, and any discussion about empirical reality (and ordinal or interval scale types for that matter, see Adams 1998; Reid et al. 1981) is irrelevant.

As suggested by Parasuraman et al. (2008) above, human factors constructs should be seen as the results of operations that enable useful prediction. If we accept this operationalist perspective, it becomes crucial to have suitable measurement procedures available. That is, to be useful in prediction, the numerical assignment procedures have to be well defined. Hand (1996) states: "Arbitrariness in the procedure will reflect itself in ambiguity in the results. This is one reason why problems arise in the social and behavioural sciences …. A complete specification of the procedure is often difficult or impossible and different researchers may use the same name for variables that actually have subtly different definitions, leading to different conclusions" (p. 453).

To explore how human factors constructs are measured, we restrict our attention to (mental) workload, arguably the most widely used human factors construct, and its measurement by questionnaires. Questionnaires are powerful tools as they can detect changes in the operator (e.g. resource allocation) that may be impossible to detect by direct observation. Hart and Staveland (1988) point out that "subjective ratings may come closest to tapping the essence of mental workload" (p. 141).

To explore how questionnaires have been used for measuring workload, we applied a so-called culturomic analysis by investigating secular trends in word use (Michel et al. 2010). First, we performed an exploratory literature search of the human factors literature that revealed about 30 different questionnaires used for measuring workload or cost incurred on the operator. Next, for the 22 questionnaires that yielded at least five search results, we registered the number of documents mentioning the specific questionnaire per quadrennium. All searches were conducted in Google Scholar.

The results in Table 3 reveal an explosive use of the NASA-TLX. Other questionnaires, such as the Cooper Harper rating scale and the Subjective Workload Assessment Technique (SWAT) show a modest increase, whereas many other available questionnaires appear short lived. Clearly, in relative terms among questionnaires, the TLX has become a dominant scale in workload measurement. When adopting the operationalist perspective, thereby regarding workload as the result of a measurement procedure that simultaneously defines and quantifies, it can be stated that workload has become synonymous with the TLX.

A positive explanation of the popularity of the TLX could be that it is based on solid evidence regarding predictive validity, is iteratively improved by self-correction, and accordingly, has become the established method for operationalizing workload. A more negative interpretation is that the increase of TLX is simply a Matthew effect ("the rich get richer"; Merton 1968). That is, the TLX is not the most sensitive or predictive-valid questionnaire available, but has become popular because it has reached sufficient escape velocity and is now the obvious choice available to researchers and practitioners.

We argue that the latter explanation is most probable, as there are specific operational (pragmatic) concerns with the NASA-TLX which have persisted since the year the TLX was first published, and which have not resulted in corrective response (cf. Table 4). Other workload questionnaires, which perform well or arguably sustain certain advantages compared to the TLX (Hill et al. 1992; Rubio et al. 2004), do not appear to have been embraced by the human factors community.

Table 3 Number of search results per quadrennium for the most popular workload questionnaires

	1961–1964	1965–1968	1969–1972	1973–1976	1977–1980	1981–1984	1985–1988	1989–1992	1993–1996	1997–2000	2001–2004	2005–2008	2009–2012
"NASA-TLX" OR "NASA Task Load Index"	0	0	0	0	0	0	49	147	283	524	1,097	1,876	2,966
"Cooper–Harper"	0	0	52	76	99	174	198	220	291	287	354	376	423
"Subjective Workload Assessment Technique"	0	0	0	0	0	29	100	108	99	105	111	154	219
"Quantitative Workload Inventory"	0	0	0	0	0	0	0	1	2	8	61	159	287
"Rating Scale Mental Effort"	0	0	0	0	0	0	0	0	8	24	38	67	99
"Subjective Workload Dominance"	0	0	0	0	0	0	0	20	48	25	44	28	32
"Workload" "Activation scale"	0	0	0	0	0	0	5	2	3	7	4	11	16
"Raw Task Load Index" OR "NASA RTLX"	0	0	0	0	0	0	0	11	18	40	44	54	78
"Instantaneous Self-Assessment"	0	0	0	0	0	0	0	2	5	18	42	35	48
"NASA Bipolar"	0	0	0	0	0	0	23	23	16	9	14	20	14
"Overall Workload Scale"	0	0	0	0	0	0	7	11	5	5	15	11	20
"Multiple Resources Questionnaire"	0	0	0	0	0	0	0	0	0	0	7	18	37
"Pro-SWAT"	0	0	0	0	0	2	9	5	2	1	5	2	0
"Workload Profile WP"	0	0	0	0	0	0	0	0	0	0	2	5	18
"Crew Status Survey"	0	0	0	0	0	1	0	5	3	2	3	4	7
"Pilot Objective/Subjective Workload"	0	0	0	0	0	3	4	2	2	1	1	2	1
"Continuous Subjective Assessment"	0	0	0	0	0	0	0	2	3	4	1	2	4
"Workload/Compensation/Interference/Technical Effectiveness"	0	0	0	0	0	3	1	1	1	0	4	2	2
"Sequential Judgment Scale"	0	0	0	0	0	0	0	0	0	0	0	0	0
"Driver activity load index"	0	0	0	0	0	0	0	0	0	0	1	1	14
"Dynamic Workload Scale"	0	0	0	0	0	0	0	0	0	0	1	1	14
"Flight Workload Questionnaire"	0	0	0	0	0	1	0	0	2	1	0	1	0
Total number of search results	0	0	52	76	99	213	396	560	791	1,061	1,849	2,829	4,299
Total number of unique search results	0	0	50	73	99	199	328	423	618	893	1,487	2,415	3,681

Left-hand column shows the syntax used in Google Scholar searches conducted on March 14, 2014. A total of 18 selected false positives have been removed

Table 4 Identified persistent questions regarding the NASA Task Load Index (TLX)

1. The investigator can use a weighting procedure with pairs of cards prior to calculating the total TLX scores, or he can decide to use the unweighted scores (also called the NASA Raw Task Load Index, or NASA RTLX). Various researchers have pointed out that the weighted and unweighted sum scores are highly correlated (Byers et al. 1989; Moroney et al. 1992; Nygren 1991), implying that it is superfluous to invest two minutes in the card sorting required for weighting. Others state that weights provide valuable diagnostic information (Dickinson et al. 1993; Liu and Wickens 1994). What policies are recommended with regard to the predictive validity of the TLX?
2. The official TLX contains items on a 21-tick scale (a 12-cm line divided into 20 equal intervals). The participant should mark one of the 21 vertical ticks. Experience shows that subjects using the paper-and-pencil version are naturally inclined to mark between ticks. The guidelines state that "if a subject marks between two ticks, the value of the right tick is used (i.e., round up)" (NASA 1986, p. 4).
3. The anchors of the performance item are susceptible to misinterpretation. TLX developers observed this in 1986: "Note that 'own performance' goes from 'good' on the left to 'bad' on the right. This order has been confusing for some people" (NASA 1986, p. 11). Accidental scale reversal, of course, has important implications for the obtained results.
4. Two TLX versions seem to exist and both are available from official NASA sources. The difference is in anchors and overall layout. On their official website, NASA offer a TLX paper/pencil version with items that run from "very low" to "very high" (NASA 2014), whereas in the corresponding instruction manual, the items run from "low" to "high" (NASA 2014). Choice of TLX version has important implications, as "we cannot ignore the possibility that the measurement scales themselves may be limited by artifacts such as floor and ceiling effects" (Hancock 1996, p. 1157).

Individual items seem trivial, but could still have important practical consequences for the obtained rating scores

8 Conclusion

Human factors terms, such as situation awareness and workload, are controversial and widely used terms in human factors science. It is a moot point whether these constructs should deserve scientific status, or whether they should merely be seen as folk models. Reflecting on the works by Stanley Smith Stevens and relying on work by theorist David Hand, we suggest a resolution to the debate.

Our suggestion is that human factors constructs are situated to the operational end of a representational–operational continuum. In other words, we agree with Parasuraman et al. (2008) that human factors constructs are not part of an empirical reality. From an operational (pragmatic) perspective, human factors constructs are useful in prediction. As Hart and Staveland (1988) point out regarding workload, "there is no objective standard" (p. 143) and "no physical units of measurement" (p. 143),

Table 5 Matrix of performance and workload associations and dissociations (after Hancock 1996)

		Performance		
		Better	Stable	Worse
Workload	Higher	Operator invests considerable effort which turns out to successful	Operator invests effort or uses adaptive strategies to maintain performance	Association
	Same	Operator is insensitive to own output	Association	Operator is insensitive to own output
	Lower	Association	Operator's skill has developed	Operator gives up

Hancock stated that "if workload response always followed performance variation, then there would be little reason to collect such additional measures" (p. 1156). In other words, workload is not the same as how well a person performs a certain task in terms of speed or accuracy. It is the disassociations between workload and performance that provide predictive information about an operator's skill and strategies

but workload "remains an important, practically relevant, and measurable entity" (p. 139).

Accepting that human factors constructs are the result of pragmatism releases the field from some of its "physics envy" (Hancock and Szalma 2004, p. 500). However, for operationalization to be successful, it is of utmost importance to have suitable measurement procedures available (Table 5).

To explore how questionnaires are used for measuring workload, we carried out a lexical analysis. The explosive use of the NASA-TLX suggests that this scale has become dominant through a Matthew effect, while we found no indication that the TLX is iteratively adjusted towards optimal validity. In fact, some issues, such as inconsistent use of scale anchors, have persisted since the inception of the TLX in 1986. In this respect, we agree with Dekker and Hollnagel (2004) that workload is "a measure defined by consensus, rather than by reference to a model" (p. 83). Apparently, operationalist science is not self-correcting (cf. Ioannidis 2012), and certain procedures become established because of their sheer quantity and availability, a situation which Dekker (2013) characterizes as "a kind of consensus authority: everybody uses it, so everybody uses it" (p. 96).

Our statements are not a critique of the TLX per se. After all, TLX was established after an extensive three-year research effort at a reputable institute, and it sits properly in a web of correlations with external variables (Hart and

Staveland 1988). However, somewhat disconcertingly, workload has now become almost synonymous with the TLX, while any attempts to launch alternative scales appear to be short lived.

After years of research, Stevens discovered his power law and applied it to the dozens of perceptual continua. No such powerful numeric predictive technique seems available in the domain of human factors constructs. One suggested way to improve the situation is to strive towards more powerful prediction of human factors constructs by collecting more quantitative research evidence. The hope is then that eventually, the field will become self-correcting and that highly effective predictive measurement tools will become available.

Another possible way to improve the situation is to move human factors constructs more to the representational side of the continuum and tie them better to empirical relationships (Kantowitz 1992), thereby avoid being "forever bound to measurement which is largely pragmatic" (Hand 2004, p. 82). Parasuraman et al. (2008) observed that human factors constructs are increasingly associated with various measures of brain and autonomic system activity, hence expanding the nomological net of these constructs into the biological domain. Neuroergonomics offers a biological explanation and may therefore contribute to a shift towards the representational dimension, and alleviate some of the problems seen in operationalism. Indeed, brain imaging techniques provide increasing evidence of the correlation between brain activity and cognitive state (Mather et al. 2013).

Acknowledgments I thank Dr. Dimitra Dodou for useful comments and for helping to create Table 3.

References

Adams S (1998) Practical considerations for measuring Situational Awareness. In: Proceedings for the third annual symposium and exhibition on situational awareness in the tactical air environment, Piney Point, MD, pp 157–164

Boring EG, Stevens SS (1936) The nature of tonal brightness. Proc Natl Acad Sci 22:514–521

Bridgman PW (1927) The logic of modern physics. MacMillan, New York

Byers JC, Bittner AC, Hill SG (1989) Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In: Mital A (ed) Advances in industrial ergonomics and safety 1. Taylor and Francis, London, pp 481–485

Campbell NR (1920) Physics: the elements. Cambridge University Press, Cambridge

Campbell NR, Jeffreys H (1938) Symposium: measurement and its importance for philosophy. Proc Aristot Soc Suppl Vol 17:121–151

Dawes RM, Smith TL (1985) Attitude and opinion measurement. In: Lindzey G, Aronson E (eds) The handbook of social psychology. Random House, New York, pp 509–566

Dekker SWA (2013) On the epistemology and ethics of communicating a Cartesian consciousness. Saf Sci 56:96–99

Dekker S, Hollnagel E (2004) Human factors and folk models. Cogn Technol Work 6:79–86

Dekker SW, Woods DD (2002) Maba-maba or abracadabra? progress on human-automation co-ordination. Cogn Technol Work 4:240–244

Dekker S, Nyce JM, Van Winsen R, Henriqson E (2010) Epistemological self-confidence in human factors research. J Cogn Eng Decis Mak 4:27–38

Dickinson J, Byblow WD, Ryan LA (1993) Order effects and the weighting process in workload assessment. Appl Ergon 24:357–361

Ferguson A, Myers CS, Bartlett RJ, Banister H, Bartlett FC, Brown W et al (1938) Quantitative estimates of sensory events: interim report. Br Assoc Adv Sci 108:277–334

Ferguson A, Myers CS, Bartlett RJ, Banister H, Bartlett FC, Brown W et al (1940) Quantitative estimates of sensory events. Adv Sci 1:331–349

Flach JM (1995) Situation awareness: proceed with caution. Hum Factors 37:149–157

Hancock PA (1996) Effects of control order, augmented feedback, input device and practice on tracking performance and perceived workload. Ergonomics 39:1146–1162

Hancock PA, Szalma JL (2004) On the relevance of qualitative methods for ergonomics. Theor Issues Ergon Sci 5:499–506

Hand DJ (1996) Statistics and the theory of measurement. J R Stat Soc A 159:445–492

Hand DJ (2004) Measurement: theory and practice. Arnold, London

Hand DJ (2005) Size matters—how measurement defines our world. Significance 2:81–83

Hand DJ, Keynes M (1993) Letter to the Editor: Velleman PF, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47, 65–72: Comments by Huberty and Hand; rejoinder by Velleman and Wilkinson. Am Stat 47:314–315

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human mental workload, North Holland Press, Amsterdam, pp 139–183. http://humanfactors.arc.nasa.gov/groups/TLX/downloads/NASA-TLXChapter.pdf

Hendy KC, Hamilton KM, Landry LM (1993) Measuring subjective workload: when is one scale better than many? Hum Factors 35:579–601

Hill SG, Iavecchia HP, Byers JC, Bittner AC Jr, Zaklade AL, Christ RE (1992) Comparison of four subjective workload rating scales. Hum Factors 34:429–439

Ioannidis JP (2012) Why science is not necessarily self-correcting. Perspect Psychol Sci 7:645–654

Kantowitz BH (1992) Selecting measures for human factors research. Hum Factors 34:387–398

Khurshid A, Sahai H (1993) Scales of measurements: an introduction and a selected bibliography. Qual Quant 27:303–324

Liu YL, Wickens CD (1994) Mental workload and cognitive task automaticity: an evaluation of subjective and time-estimation metrics. Ergonomics 37:1843–1854

Lord FM (1953) On the statistical treatment of football numbers. Am Psychol 8:750–751

Luce RD, Suppes P (2002) Representational measurement theory. In: Pashler H, Wixted J (eds) Stevens' handbook of experimental psychology, vol 4. Methodology in experimental psychology, 3rd edn. Wiley, New York, pp 1–41

Marcus-Roberts HM, Roberts FS (1987) Meaningless statistics. J Educ Behav Stat 12:383–394

Marks LE (2006) S.S. Stevens: a brief scientific biography. In: Fechner Day 2006: Proceedings of the twenty-second annual

meeting of the international society for psychophysics, St. Albans, England
Mather M, Cacioppo JT, Kanwisher N (2013) Introduction to the Special Section: 20 Years of fMRI—what has it done for understanding cognition? Persp Psychol Sci 8:41–43
Merton RK (1968) The Matthew effect in science. Science 159:56–63
Michel J-B, Shen YK, Aiden AP et al (2010) Quantitative analysis of culture using millions of digitized books. Science 331:176–182
Miller GA (1974) Stanley Smith Stevens: 1906–1973. Am J Psychol 87:279–288
Moroney WF, Biers DW, Eggemeier FT, Mitchell JA (1992) A comparison of two scoring procedures with the NASA Task Load Index in a simulated flight task. In: Proceedings of the IEEE national aerospace and electronics conference 2, Dayton, OH, pp 734–740
NASA (1986) Task Load Index (NASA-TLX). v. 1.0. Paper and pencil package (instruction manual). NASA Ames Research Center, Moffett Field, CA. http://humanfactors.arc.nasa.gov/groups/TLX/downloads/TLX.pdf
NASA (2014) Task Load Index paper and pencil version. NASA Ames Research Center, Moffett Field, CA. http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf
Nygren TE (1991) Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. Hum Factors 33:17–33
Parasuraman R, Sheridan TB, Wickens CD (2008) Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J Cogn Eng Decis Mak 2:140–160
Reid GB, Shingledecker CA, Eggemeier FT (1981) Application of conjoint measurement to workload scale development. Proc Hum Factors Ergon Soc Annu Meet 25:522–526
Rubio S, Díaz E, Martín J, Puente JM (2004) Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and Workload Profile methods. Appl Psychol 53:61–86
Siegel S (1956) Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York
Stevens SS (1946) On the theory of scales of measurement. Science 103:677–688
Stevens SS (1959) Measurement, psychophysics and utility. In: Churchman CW, Ratoosh P (eds) Measurement: definitions and theories. Wiley, New York, pp 18–63
Stevens SS (1961) To honor Fechner and repeal his law. Science 133:80–86
Teghtsoonian R (2001) S. S. Stevens. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social and behavioral sciences 22:15104–15108. Pergamon, Oxford
Townsend JT, Ashby FG (1984) Measurement scales and statistics: the misconception misconceived. Psychol Bull 96:394–401
Underwood BJ (1957) Psychological research. Prentice-Hall, Englewood Cliffs, NJ
Uttal WR (2008) Time, space, and number in physics and psychology. Sloan Publishing. http://j.b5z.net/i/u/2084689/f/Online_Time_Space_Number.pdf
Velleman PF, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47:65–72