Theoretical Issues in Ergonomics Science

ISSN: 1463-922X (Print) 1464-536X (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures

Gerald Matthews, Joost De Winter & Peter A. Hancock

To cite this article: Gerald Matthews, Joost De Winter & Peter A. Hancock (2019): What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures, Theoretical Issues in Ergonomics Science, DOI: 10.1080/1463922X.2018.1547459

To link to this article: https://doi.org/10.1080/1463922X.2018.1547459

Published online: 24 Jan 2019.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
https://doi.org/10.1080/1463922X.2018.1547459

Taylor & Francis
Taylor & Francis Group

What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures

Gerald Matthews[a], Joost De Winter[b] and Peter A. Hancock[a,c]

[a]Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA; [b]Department of BioMechanical Engineering, Delft University of Technology, Delft, The Netherlands; [c]Department of Psychology, University of Central Florida, Orlando, FL, USA

ABSTRACT

We examine the continuing use of subjective workload responses to index an operator's state, either by themselves or as part of a collective suite of measurements. Lack of convergence of subjective scales with physiological and performance-based measures calls into question whether there is any unitary workload construct that underpins conscious experience, physiological state and the individual's profile of task-related performance. We examine philosophical and measurement perspectives on the divergence problem, and we consider three possible solutions. First, difficulties in reliable and valid measurement of workload may contribute to divergence but do not fully explain it. Second, workload may be treated operationally: use of specific measures is justified by demonstrating their pragmatic utility in predicting important outcomes. Third, further efforts may be made to develop representational workload measurements that correspond to real empirical phenomena. Application of formal standards for test validity can identify multiple latent constructs supporting subjective workload, including those defining self-regulation in performance contexts. Physiological and performance-based assessments may define additional, distinct constructs. A resolution of the diversity issue is crucial for ergonomics since the invalid application of workload measurement will threaten exposed operators as well as many others who are served by the complex technological systems they control.

ARTICLE HISTORY
Received 04 June 2018
Accepted 08 November 2018

KEYWORDS
Assessment measures; cognitive workload; convergence; psychometrics; validity

Relevance to human factors/Relevance to ergonomics theory

Workload assessment is important for multiple applications of ergonomics, but subjective and objective workload measures often diverge. Current ergonomics theory fails to explain this divergence. This article provides two possible solutions to this challenge. First, specific workload measures may be treated as operational variables that predict important real-world outcomes. This approach eschews theory in favor of pragmatic utility. It is compatible with emerging "big data" approaches to performance prediction. Second, subjective measures

CONTACT Gerald Matthews gmatthews@ist.ucf.edu Institute for Simulation and Training, University of Central Florida, Orlando, FL, USA

may be reconceptualized as attributes of the operator's self-perceptions, which are only loosely related to neural functioning and task processing. Application of modern standards for test validation, including use of latent factor modeling, may advance the theory of subjective workload, potentially supporting more representational measurement. In any case, ergonomics practitioners should be aware of the limitations of current workload assessments. Practitioners should understand workload theory sufficiently to choose the conceptual framework most suitable for their applied assessment needs.

1. Introduction

In contemporary assessments of how humans work within technical systems, the use of subjective workload measures appears now to have become more popular than ever (see; De Winter 2014; Matthews and Reinerman-Jones 2017; Young, Brookhuis, Wickens, and Hancock 2015). Yet, the scientific credibility and utility of such measures continues to be questioned by some researchers (e.g. Dekker and Nyce 2015). The case against subjective workload scales derives primarily from two fundamental concerns. First, the attachment of numerical values to linguistic terms in an attempt to render private cognition available for public scrutiny has a long and ongoing history in psychology but remains challenging (see Proctor and Xiong 2017). Philosophically-inspired critiques have addressed the conceptual difficulties of quantifying psychological characteristics, especially in the absence of a theory that specifies a normative meaning for the characteristics concerned (Barrett 2005; Hancock, Sanders, and Volante 2015; Michell 1999). The second source of objections derives more directly from empirical concerns. Most especially here, the assessment of subjective workload can fail the key test of reliable convergence with respect to objective physiological and performance-based workload measures (and see also; Hancock and Matthews 2018).

What would most certainly be an exhaustive, and exhausting, review of all the relevant experimental data is beyond the scope of the present disquisition. However, in a broad sense, the current panoply of studies appears to differ quite significantly in the extent to which they show convergence between alternate workload measures derived from self-report, from physiological response and from task performance (see Hancock 2017). Despite these 'internal' assessment difficulties, workload scales still remain undeniably useful in many applied circumstances. They have proven their utility in detecting potentially dangerous levels of task demand, in evaluating the impact on the operator of differing interfaces and varying systems operations, and in predicting an operator's ability to take on extra tasks (see: Matthews and Reinerman-Jones 2017; Parasuraman, Sheridan, and Wickens 2008; Young et al. 2015). Excessive workload may also contribute to occupational health problems such as chronic fatigue (Liu, Fan, Fu, and Liu 2018). In this present work, we explore approaches to resolve the tension between the questionable scientific status of mental workload assessment (henceforth, 'workload') and its pragmatic utility. From the practical, ergonomic perspective, we focus especially on the issue of the lack of convergence of these various standard measures.

The article is structured as followed. First, we define the nature of the convergence problem. We set out the psychometric assumptions of a unitary workload and illustrate instances of divergence between subjective and objective measures that challenge those

assumptions. One possible explanation for the lack of convergence is that there are psychometric deficiencies in one or other type of measure that mask the true workload response. We describe measurement issues that may weaken convergence, but it appears that divergence cannot be attributed to these issues alone. The two sections that follow offer contrasting solutions to the problem, based on operational and representational perspectives on measurement (Hand 2004). According to Hand (1996, p. 448), representational measurement "… seeks to represent or model empirical relationships – and so is about understanding the substantive domain of investigation – whereas … [operational measurement] … seeks to predict. Accurate prediction can be achieved without any understanding of the underlying mechanism." The capacity of workload measures to predict meaningful outcomes such as increased error rates demonstrates their practical utility in support of the operational approach (De Winter 2014). We will also review prospects for developing a more representational perspective on workload, starting from identification of relevant latent constructs using modern psychometric methods and validation standards. Subjective workload may index constructs associated with self-regulation and metacognition in performance settings. We conclude with a brief account of the practical implications that follow from discarding the simple, unitary conception of workload.

1.1. Subjective workload assessment: the convergence problem

Three general types of workload measures have been commonly used for practical assessment. These are (1) subjective assessment scales, (2) psychophysiological responses, and (3) performance measures; the latter being from either primary or secondary tasks, both singly and in combination. Each overall category includes multiple measures that tend to proliferate as research advances. Subjective workload research has accommodated multiple resource theory (Wickens 2008), with the introduction of the Multiple Resources Questionnaire (MRQ: Boles, Bursk, Phillips, and Perdelwitz 2007). Beyond traditional measures based on autonomic and central nervous system response, psychophysiological innovations include measurement of muscular co-contraction (Meulenbroek et al. 2005), cerebral blood flow velocity (Warm, Tripp, Matthews, and Helton 2012) and nose temperature (Orr and Duffy 2007).

Ideally, alternate measures of a common workload construct should converge; that is, task demand manipulations should have similar effects across multiple workload indicators. Measures should also correlate with one another on a between- and within-person basis. Instances of non-convergence are a longstanding challenge for workload assessment (Yeh and Wickens 1988). Evidence on this issue comes from experimental studies concerning the 'AIDs' of workload, i.e. its *associations*, *insensitivities* and *dissociations* (Hancock 2017; Hancock and Matthews 2018). The term 'dissociation' can refer to both divergence of alternate workload measures (Hancock 2017) and to lack of association between workload and primary task performance (Yeh and Wickens 1988). In psychometrics, a fundamental element of validation (Campbell and Fiske 1959) is testing for convergence of measures of the same construct, and divergence of measures of different constructs. One of the themes of this article is that using modern psychometric techniques including analysis of variance-covariance structures to investigate latent factors may contribute to determine convergences and divergences between different workload measures.

1.1.1. Workload as a unitary construct: Psychometric assumptions

Figure 1 illustrates the typical basic assumptions concerning workload measures and their convergence as they might be expressed in a simple latent factor model. External task demands influence an underlying workload construct defined via the three principal types of measures. The latent factor, in turn, influences primary task performance provided that it is resource-limited, and the operator cannot preserve performance through changes in task strategy. The treatment of performance measures in Figure 1 is based on resource theory which provides the foundation for the workload concept (Young et al. 2015). Secondary task performance measures 'spare capacity' that is inversely related to the proportion of resources allocated to the primary task. Thus, measures of RT to secondary task probes may be sensitive to resource variation even when the person is able to maintain stable primary task performance (De Waard and Brookhuis 1997; Gawron 2008). Use of time to index workload also potentially provides the advantage of ratio scaling, i.e. that the scale has a true zero point and differences between values are meaningful (Jensen 2006). Thus, secondary task performance should converge with subjective and physiological measures in defining the latent factor.

Primary task measures are sometimes treated as workload measures, but within the resource/demand model it is preferable to treat primary task performance as an outcome that correlates with workload only under certain conditions (Hart and Wickens 2010). Typically, performance is directly limited by resource availability when the task is moderately difficult (Hart and Wickens 2010). In this case, primary task performance should converge with workload, the case illustrated in Figure 1. However, for easy and very difficult tasks, data-limitations on performance as well as floor and ceiling effects allow workload to vary independently of performance (Hart and Wickens 2010; Vidulich and Tsang 2012). The contingent nature of the workload – primary task relationship differentiates the two constructs, and ensure the utility of workload; otherwise, it becomes essentially redundant in a practical sense and perhaps in a theoretic way also.

1.1.2. Divergence of subjective and objective workload assessments

Studies of the 'AIDs' of workload (Hancock 2017) show a variety of convergences and divergences between subjective and objective measures. A review of this literature is

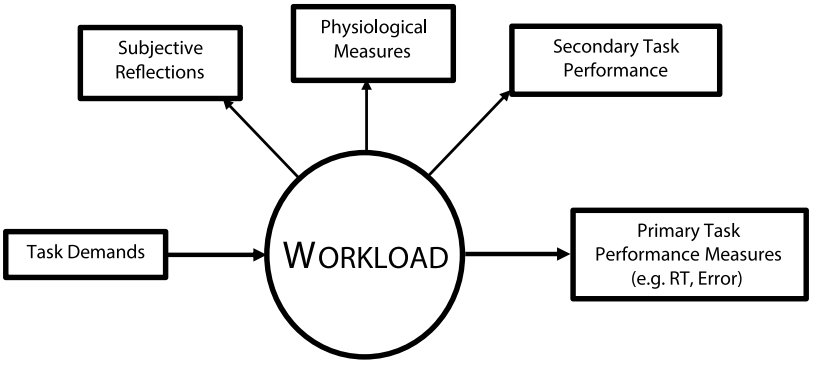

Figure 1. Generalized diagram of the standard understanding of workload as a latent construct: Its antecedent, its reflections in measurable variables, and its output for resource-limited primary task performance.

beyond our present scope (see Hancock 2017; Hancock and Matthews 2018; Wickens et al. 2015). The issue here is the challenge it poses for the unitary latent factor model. First, measures may be empirically uncorrelated with one another, so that no latent factor can be defined psychometrically (e.g. Funke et al. 2013; Myrtek et al. 1994). Matthews et al. (2015a) found that subjective workload and metrics from five different physiological sensors were largely independent of one another. Independence of alternate physiological measures demonstrates difficulty in defining a unitary 'physiological workload' factor. Second, in experimental studies, multivariate patterns of workload response may vary substantially across different task domains. Matthews et al. (2015b) summarized evidence from four simulation studies utilizing multiple sensors (see also Hancock and Matthews 2018, Table 2). Workload manipulations were appropriate to differing applied contexts such as nuclear power plant control and operations of unmanned ground and aerial vehicles. Each manipulation increased subjective workload and one or more of the physiological indices measured. However, the pattern of physiological response proved to be substantially different for each context, i.e. the subjective workload increase was not diagnostic of changes in brain functioning. Third, task demands may produce opposite effects on different workload measures (dissociation), an outcome that is strongly incompatible with the unitary factor model. For example, low heart rate variability (HRV) and high-frequency EEG (beta) are commonly taken as indicators of high workload. However, the studies reviewed by Matthews et al. (2015b) included instances of task demands increasing HRV and decreasing beta power, as well as studies showing the opposite effects. Some studies, of course, do show convergence between subjective, physiological and secondary task measures (e.g. Hwang et al. 2008; Lee and Liu 2003), and there may be stronger evidence for convergence from within-person analyses of physiological and subjective response (e.g. Rendón-Vélez et al. 2016). Nevertheless, a viable unitary factor model should be sufficiently robust to emerge at least somewhat consistently in data from different task domains.

1.1.3. Divergence of subjective workload from primary task performance

In many experimental studies, attention has also fallen on the incidence of dissociation between subjective workload and objective performance measures (Hancock 1996; see also Yeh and Wickens 1988). As previously discussed, the unitary factor model permits a weaker relationship between workload and primary task performance than between the types of measured workload indicator. For example, under data-limited conditions, workload and performance can vary independently (Vidulich and Tsang 2012). Performance and workload may even increase together when task demands elicit higher effort, challenge and enjoyment, for example, highly engaging videogame-like tasks (Abich, Reinerman-Jones, and Matthews 2017). It was suggested by Hancock (2017) that such 'affective' dimensions of task demand had received insufficient attention and the cited studies provide confirmation of this assertion. Such effects might be accommodated within modified resource theories (e.g. Young and Stanton 2002) that permit the quantity of resources available to vary with motivational factors.

Dissociation with tasks for which there is evidence for resource-limitation is especially problematic. Table 1 shows NASA-TLX data from one of the present authors' (GM) studies, based on a sample of various attentional-demanding tasks where evidence from sources

Table 1. Examples of studies providing data on performance correlates of the NASA-TLX.

Study	*N*	Task	Performance Measure	Overall Workload	Own Performance	Effort
Matthews et al. (2010)	187	36 min sensory vigilance task	Perceptual sensitivity (*A'*): final period	−.12	−.37**	.24**
Matthews et al. (2010)	101	36 min cognitive vigilance task	Perceptual sensitivity (*A'*): final period	−.31**	−.45**	.32**
Matthews et al. (2014)	342	12 min sensory vigilance task	Perceptual sensitivity (*d'*): whole task	−.01	−.42**	.23**
Matthews and Campbell (2010)	112	OSPAN task (with time pressure)	Word recall (total correct)	−.16	−.42**	.06
			Arithmetic (% correct)	−.35**	−.42**	.06
Fellner (2008)	121	Discrimination learning (collaborating 2-person teams)	Discriminations (total correct): final period	−.21*	−.50**	.07
Matthews et al. (2017)	150	Simulated UGV operation: surveillance mission	Change detection (% correct)	−.23**	−.23**	.01
			Threat (% correct)	−.02	−.12	−.01
Neubauer et al. (2012)	173	Simulated driving: fatiguing conditions	Variability in lateral position	.00	−.02	.11
			Speed of response to hazard (RT)	.02	−.13	.06

*$p < .05$, **$p < .01$.

Table 2. Five types of evidence supporting test validity.

Type of Validity Evidence	Critical Questions
Evidence based on test content	Is test content sampled systematically and comprehensively?
Evidence based on response processes	Is response free of major bias from extraneous factors?
Evidence based on internal structure	Do psychometric analyses support the proposed dimensional structure of the test?
Evidence based on relations to other variables	Do experimental and correlational studies show meaningful relationships with other related constructs?
Evidence based on the consequences of testing	Does test usage lead to measurable benefits?

such as temporal performance decrement, sensitivity to fatigue and dual-task interference suggests resource-limitation. *N*s were sufficient to estimate correlations with a relatively high degree of confidence. A detailed review is beyond the present scope, but further details of tasks and procedures may be found in each of the source publications cited. In fact, only four out of 10 tasks showed the expected negative correlation between overall workload and poorer performance. Effort ratings correlated with performance only for vigilance. Ratings of poor performance were more reliably associated with performance. Thus, people may commonly have an awareness of their quality of performance (though not always), but this insight is not necessarily associated with a significant workload – performance association.

Again, there may be better evidence for workload – performance convergence from analyses of within-subjects correlations, e.g. between subjective workload and rate of steering movements in simulated driving (Verwey and Veltman 1996). However, such findings do not readily support a general factor model applicable across both persons and task situations. In general, convergence is disappointingly weak at best, even for manifestly attentionally demanding tasks.

1.2. Divergence: deficiencies in measurement of subjective workload

Observations of divergence imply that differing workload measures may not actually index the same construct. A counter-argument, however, is that the deficiency lies in the measures themselves rather than the underlying construct, i.e. that failings of one or more of the various types of measures limit their validity. The general difficulties of subjective assessments are well-known. These include their dependence on unverifiable introspection, their vulnerability to various perceptual and response biases, and the basic difficulty of attaching unequivocal numerical values to conscious experiences (Annett 2002; Muckler and Seven 1992). 'Workload', like most psychological constructs, proves to be broad and often somewhat fuzzily-defined (Van Acker, Parmentier, Vlerick, and Saldien 2018). Thus, this label can and has meant different things in different contexts. Subjective responses are also known to vary across individuals in both their basic conception of the term and in their own personal reports of their experience. For example, vehicle drivers often compensate for increased demands by reducing their speed to regulate such demands. However, their perception of that demand may lead to higher subjective workload ratings; i.e. people tend to assess what happens to them, rather than what they actually experience. For example, in a driving simulator study, Melman et al. (2018) found that ratings of subjective effort were lower when lanes became wider, even though participants drove considerably faster (presumably increasing effort). Factors associated with task demands have been identified in experimental studies reporting the phenomenon of dissociation. Thus, subjective measures are more sensitive to increase in the absolute number of tasks being performed, rather than to response execution demands and/or to resource competition (Vidulich and Tsang 2012; Yeh and Wickens 1988). At present, we possess only a partial list of the various individual difference factors that may pertain to awareness of workload (Damos 1988). We also lack a formalized taxonomy of contexts which might guide our understanding of these, specific performance-environment effects and identify biases in self-perceptions of workload.

There are also issues specific to the design of many extant subjective workload scales. For example, administration of the NASA-TLX scale (Hart and Staveland 1988) requires the provision to respondents of rating scale descriptions. This procedure recognizes, but does not necessarily solve, the problem of variation in interpretations of 'workload.' De Winter (2014) has pointed out persistent issues with the NASA-TLX whose impact on validity has not been either fully aired or fully resolved as of the present time. These include, but are not limited to, the nature of the response scales, whether or not the raw ratings are weighted, and how people interpret the specific 'own performance' rating element which uses a different and inverted response scale to all of the others. Another difficulty is that the task demand commonly varies on a moment-to-moment basis requiring the person estimates workload for what is often conceived as a unitary 'task.' Thus, they need to integrate multiple memories of their recent experience; sometimes this can be done fairly accurately, at other times not. Of course, this integration is also a constructive process that can be disproportionately influenced by workload peaks or deviations from expected workload (and see Hancock 2017; Jansen et al. 2016). It further argues that validity of workload measures may be superior in individuals with better episodic memory, and for tasks which contain relatively few stimuli, relative to those for which many events happen during any particular epoch of interest.

Finally, an issue which is especially salient in physiological assessment is the correction of such indices for *a priori* individual differences. This is usually achieved by comparing

recordings in the experimental state against those taken previously in some resting, or baseline state. For example, resting heart rate is used to reflect purely physiological functions such as those related to persisting factors of health and physical fitness. These are concatenated together with acute factors such as momentary respiration rate and blood oxygenation level. Mental workload may, therefore, be much better represented by the change in heart rate from the baseline state as compared with the raw beats-per-minute value (Roscoe 1978), although there are various statistical issues involved with calculating such reactivity measures (Burt and Obradović 2013). Possible baseline correction issues have often been neglected in subjective workload research. Just as individuals exhibit differential physiological baselines, people may well differ in their perceptions of the workload involved, even in performing cognitively trivial tasks. This form of correction for individual psychological variation is rarely performed but correcting for such baselines might very well improve validity of workload assessments.

1.3. Divergence: deficiencies in measurement of objective workload

Divergence of measures may also reflect measurement shortcomings of objective physiological and performance-based workload measures. Various specific factors may reduce measurement accuracy for certain particular physiological systems. These include artifacts, such as those caused by motion, confounding variables such as ambient light intensity in the case of pupillometry, and the influences of purely physiological factors such as muscular activity in the case of ECG. Experiments, including brain imaging studies (Button et al. 2013), often prove to have insufficient statistical power for analysis.

Beyond these identified measurement issues, a further explanation for poor convergence between measures is that brain response to cognitive demands reflects the action of multiple brain sub-systems whose functioning cannot be reduced to the single attribute of workload. For example, vigilance tasks provide one of the simpler paradigms for investigating workload. Supporting the utility of the construct, higher workload tasks are more prone to perceptual sensitivity decrement (Warm, Dember, and Hancock 1996). However, brain-imaging studies dissociate multiple structures supporting vigilance that are likely sensitive to different demand factors. These include functions such as maintaining task set, monitoring and signaling a need-for-effort, signaling attentional priority and need-for-reorientation, regulating input-output rules, as well as sensory and motor processing (Langner and Eickhoff 2013). The relationships of workload metrics derived from sensors such as EEG and ECG with activity in specific brain areas have often proved enigmatic, although research that now includes studies of joint fMRI and EEG response to cognitive load may serve to clarify this link (e.g. Zhao, Li, and Yao 2017). Yet, another set of issues arises from lack of selectivity; e.g. ECG measures may reflect physical activity and emotional stress as well as workload itself (e.g. Hockey et al. 2009). A final limitation of psychophysiology is that workload response is often dominantly dependent on the individual (Teo et al. 2018). Similar to the response specificity principle for arousal (Stephens, Christie, and Friedman 2010), individuals differ quite widely in which metrics prove most sensitive to cognitive demand manipulations.

Turning to performance-based measures, probe RT methodology has a strong basis in resource theory but also raises assessment issues. One difficulty is that it is hard to control for variation in the voluntary choice of resource allocations across different tasks, especially

in complex, real-world tasks (Young and Stanton 2004). Multiple resource theory poses additional challenges (Wickens 2008); the validity of the probe RT method will depend on the match between the types of resource required for the primary task and for processing the probe stimulus.

1.4. What if workload is non-unitary?

To summarize the foregoing observations, there are some reasons to attribute poor convergence between workload measures to various limitations of all three of the general categories which are the primary means of workload assessment. The larger question is whether lack of convergence is solely attributable to these methodological limitations alone. Imagine that at some future date, we can train participants as expert introspectionists, further that we can also log real-time activity of multiple brain areas with high temporal and spatial resolution, and finally, that we can precisely index resource allocation with probes such as RT measures. The empirical question is then as follows in such a case: would we see high or even determinative convergence between different workload measures, so defining an underlying unitary construct? No doubt methodological advancements can serve to improve convergence, perhaps even substantially so. However, we are skeptical that convergence would then be sufficiently high to declare alternate assessments interchangeable. From a theoretical standpoint, appraisals of subjective state, the exhibition of various physiological responses and resource allocation strategies are distinct constructs that are not required to relate strongly. That is, if conscious awareness of workload reflects a complex, constructive process (Annett 2002), rather than some direct 'read-out' of neural activity, it remains open as to how strong relationships between subjective workload and objective metrics will prove to be, even under ideal measurement conditions.

The current state of the science does not allow a definitive answer to the convergence problem. We cannot preclude the possibility that advances in methodology and measurement technology together with advancements in resource theory will eventually support a latent factor model resembling that of Figure 1. However, given the current lack of evidence for a unitary latent factor (Matthews et al. 2015b), the present research question is how to interpret subjective workload measures in the absence of strong convergence with objective indices. If no single latent construct underpins the various subjective and objective workload measures that are available, they cannot all be considered interchangeable measures of the resource/demand balance.

To answer the interpretation question, we will contrast the representational and operational perspectives on measurement (Hand 2004). Representational measurement requires assignation of numbers to empirically-verifiable properties of objects, such as their mass or their velocity. In the workload context, the most obvious candidates for representational measures would be those directly tied to brain physiology, such as metabolic rate measured in joules or oxygen consumption rate per unit time. However, attempts to measure psychological constructs such as intelligence and personality traits face the objection that they cannot be verified against any 'natural,' quantitative unit of measurement (Michell 1999). By contrast, operational measurements are based on demonstrating that measurement procedures yield meaningful numerical values that can, for example, predict significant real-world outcomes. In part, this justification holds irrespective of a verifiable relationship between the values and actual reality. For example, a subjective quality of life scale might

comprise questions on perceptions of conceptually distinct qualities such as physical health, availability of social support and work satisfaction, that do not define a common latent factor. Nevertheless, such a scale might predict significant outcomes such as the risk of suicide. De Winter (2014) has proposed that subjective workload scales are located towards the operational end of a representational–operational continuum. They are practically useful in ergonomics, even if the numbers cannot be linked to unequivocal objective, observable characteristics of brain functioning or of behavior or even both.

Typically, measures have both an operational and a representational component (Kane 2016). In principle, both perspectives may be useful for understanding workload scores. In the remainder of the present work, we set out two contrasting strategies for interpreting subjective workload measures. First, we elaborate on De Winter's (2014) proposal that workload measures should be treated operationally. Especially in practical ergonomics contexts, we may not be immediately concerned about underlying latent constructs if workload scores can be reliably linked empirically to important outcomes, such as performance failures. Alternatively, we can pursue representational solutions that abandon assumptions which appear to be untenable, i.e. that subjective workload corresponds directly and closely to either brain activity or to processing resource allocation. We will explore the possibility that subjective workload measurement can be validated as a psychological construct linked to self-regulation using standard psychometric criteria. Neither perspective can represent the last word on the issue; thus, we present them as viable alternatives, each of which can generate further ergonomics research and each of which might be applicable in different research or practical contexts.

2. The poor theoretical status of workload and the appropriateness of operationalism

Above, we described that different measures of workload (self-report, physiological and secondary task performance) fail to converge. Here, we clarify the implications of the misplaced expectation that different workload measures ought to converge (i.e. show correlations approaching unity), and accordingly represent a construct 'workload' (Figure 1). One interpretation is that there is no reason to suppose that a true workload exists and can be quantified (see also Michell 1999), and scientists who claim they are measuring workload commit reification, also known as the fallacy of misplaced concreteness (Whitehead 1925/2011).

2.1. The case against the existence of workload

An inspection of the literature shows that the term 'workload' entered the vocabulary of Human Factors scientists in the 1970s and 1980s, without apparent cause or discovery, more or less simultaneously with the introduction of the NASA-Task Load Index (Hart and Staveland 1988). The name of the reputable institute NASA certainly has a scientific connotation, resonates in researchers' memory, and may help attract citations (presumably through self-reinforcement, or the 'Matthew effect'; see De Winter 2014). However, authority and popularity per se do not lend credibility to the theoretical status of workload.

The problematic nature of the theoretical status of workload is apparent by looking up its definitions: In the most highly cited work on workload, Hart and Staveland (1988; 7,941 citations) defined workload as 'the cost incurred by a human operator to achieve a particular level of performance'. In the second most highly cited work, however, Wickens et al. (2015; 7,859 citations) defined workload in terms of a relationship between supply and demand of resources, where the operator is said to be overloaded when the required resources exceed the maximum resources that the human operator can supply. Young et al.'s (2015) 'state of the science' review similarly refers to the resource/demand balance.

The fact that leading Human Factors scientists adopt fundamentally different definitions (i.e. a cost-outcome which is not necessarily bounded vs. a proportional relationship between resources supplied and available) is itself enough reason to cast doubt on whether workload is quantifiable, not to mention that the words 'cost' and 'resources' themselves lack a definition (and see Young and Stanton 2002, adding another layer of verbal sophistication, by arguing that resources are malleable rather than fixed variables). In a recent review, Van Acker et al. (2018) concluded that the workload literature 'suffers from arbitrary selections of various defining variables and a description of these variables at different levels of abstraction'. Also, problematic is the apparent carelessness with which terms are used and interchanged. For example, although it has been argued that workload is conceptually distinct from 'task demands', 'performance', as well as 'effort' (De Waard 1996; Parasuraman et al. 2008), the NASA TLX explicitly asks participants to report 'mental *demands*' (emphasis added), 'performance' and 'effort'. From the literature, it seems that such misuse of workload concepts and definitions are widespread (Van Acker et al. 2018).

It should also be noted that if one searches for the term 'workload' outside the realm of Human Factors, completely different definitions are found. For example, workload is often interpreted as the amount of (physical) work to be completed, and in the area of cognitive load theory (Sweller 1994; Gerjets et al. 2009), workload is divided into extraneous and intrinsic load. A related problem is that the definition of workload fails to demarcate itself from folk psychology (Dekker and Hollnagel 2004). Fuller (2005) argued that the term workload is actually the same as the more commonsense word 'difficulty' (see also De Waard and Lewis-Evans 2014). The aforementioned observations make us believe that workload is a term that has arisen within the human factors community and become an entity without being grounded in empirical reality (see the following quote for a striking example of how a psychological construct can be invented and then may fade away).

> "Let us go back to the late 1930s and early 1940s … In those days, we were talking about level of aspiration. You could not pick up a psychological journal … without finding at least one and sometimes several articles on level of aspiration … It was supposed to be a great powerful theoretical construct that would explain all kinds of things about the human mind from psychopathology to politics. What happened to it? Well, I have looked into some of the recent textbooks of general psychology and have found that either they do not mention it at all—the very phrase is missing from the index—or if they do, it gets cursory treatment in a couple of sentences. We all agree (from common sense) that people differ in what they demand or expect of themselves, and that this probably has something to do, sometimes, with their performance. But, it did not get integrated into the total nomological network, nor did it get clearly liquidated as a nothing concept. It did not get killed or resurrected or transformed or solidified; it just kind of dried up and blew away, and we no longer wanted to talk about it or do experimental research on it." (Meehl 1978)

The lack of convergence between self-reports, physiological measures and secondary task performance becomes apparent when applying a physicalist perspective to these measurements techniques. If we take simulator-based training as an example case, workload could be measured as follows (based on, e.g. de Groot, Centeno Ricote, and de Winter 2012; Melman et al. 2018):

1. **Self-report.** A participant completes a training task in a driving simulator. Once the session is over, the researcher asks the participant to step out of the simulator. The participant moves to an adjacent desk, and the researcher hands the participant a paper form containing the NASA TLX. The participant reads the text and provides marks on scales from 'Very low' to 'Very high' for the TLX items, except for the performance item which ranges from 'Perfect' to 'Failure'. A few weeks later, the researcher calculates the mean score across the six TLX items and notices that 'Perfect' performance means 'Very high' performance, which may explain why some participants report dissociated outcomes for the performance item. The researcher also notices that about half of the participants marked their answers on the ticks, while the other half provided their answers in between the ticks of the 21-tick scales.
2. **Physiological.** A participant completes a training task in a driving simulator. Electrodes are attached to the fingers of the participant, and a weak current is passed between the electrodes. The voltage is continuously recorded by a computer apparatus. A few weeks later, the researcher filters these signals and extracts measures that describe the overall activity of these signals. Incidentally, the researcher observes large individual differences in the voltage signal, and also spots signal artifacts, presumably because the participant was physically moving or turning the steering wheel.
3. **Secondary task.** A participant completes a training task in a driving simulator. Every 4–8 seconds, an auditory beep is produced. The participant has been instructed by the researcher to react as quickly as possible to the beeps by pressing a handheld button. A few weeks later, the researcher computes reaction times by computing the temporal differences between the beep presentation times and the button press times. The researcher then filters out extreme responses (based on e.g. Ratcliff 1993) and computes a mean reaction time per participant. The researcher discovers that when the participant was driving through curves, the reaction times were elevated, relative to response on straight road sections. The researcher considers further studies to investigate whether it is the cognitive demands of tracking a curve or the physical demands of steering that interfere with rapid button pressing.

From the above descriptions of the three types of workload measures, it should be apparent that these measurement procedures and analysis methods carry various pragmatic idiosyncrasies and share no apparent causal relationships. Suppose that an experimenter instructs the participant to drive faster; this may be indeed expected to increase scores on the self-report questionnaire, increase skin conductivity, and increase reaction times to the secondary task; however, this need not be so; it is perfectly imaginable that e.g. the participant reports increased workload while skin conductivity decreases. Hence, there is no particular reason to expect strong convergence: any correlation between these three measures can follow from indirect causal pathways without requiring the explanation of a

unitary workload construct. It is important to note that the aforementioned idiosyncrasies are not removable. For example, it may be argued that the problem of the number of tick options on the TLX could be prevented by using a computerized version instead of a paper-and-pencil version of the TLX. However, doing so necessarily introduces new idiosyncrasies, such as a dependency of the outcome on the size, resolution and brightness of the computer screen, and the quality of the input device. The bottom line is that TLX responses are inevitably pragmatic.

2.2. An operational model for workload

We have herein postulated that the notion of 'workload' bears no theoretical support, and that current ways of measuring workload are highly pragmatic and context-dependent. We believe that there no resolution to this conundrum other than to resort to the somewhat nihilistic and self-contained, yet fruitful, method of operationalism (Bridgman 1927). In operationalism, no underlying reality has to be assumed, and as explained by Hand (1996): '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). According to De Winter (2014), workload as measured by the NASA TLX is just defined as how participants filled out the TLX form and how the researcher processes the data, no more and no less. De Winter (2014) points to Stevens' (1935, 1946) adoption of the operational principle of explicitly rule-based assignation of numbers to objects or events as an inspiration. Stevens' work led him to the well-known power laws in psychophysics. Relationships of this clarity are rarely found in workload research (though see Estes 2015) but may be revealed in future research.

Although, workload does not exist according to the operational viewpoint—in fact, the word workload may just as well be removed from the HF/E literature—the measurements are still potentially informative and useful. For example, a researcher's interpretation of the TLX responses, especially if assessing associations with a performance index such as root mean squared error (RMSE) of lane-keeping performance, may allow for modifications to the driver training program so that the RMSE among future trainees is reduced. Furthermore, a comparison of the average TLX scores obtained from this driving simulator study with other published driving simulator studies may allow for a meta-analytic assessment of why the simulator does or does not produce safe driving outcomes. It remains in principle possible to compute correlation coefficients between the TLX scores and other variables (e.g. physiological and secondary task), and to use factor analysis and extract a latent factor (which may then be called 'workload'). However, it is important to realize that this type of workload is essentially a weighted average of the scores on the individual measures, and so still operational rather than representational. Factor analysis alone is insufficient to establish representational measurement.

Although, operationalist thinking may be regarded as unscientific and causing 'deliberate confusion of what is being measured with how it is being measured' (Michell 2004; see also Eichinger and Bengler 2014), as well as 'undesirable' (De Waard and Lewis-Evans 2014) and an 'ethical cop-out' (Dekker 2015), the perspective here is that operationalism is the right model because it describes the current status of the usage of the term workload within HF/E science.

3. Representational aspects of subjective workload assessments

Having presented the case for an operational definition of subjective workload, we consider next the contrasting position that the prospects for representational measurement merit further exploration. Given a continuum from representational to operational measurement, subjective workload assessments may have representational aspects, even if they are less strongly representational than physics-based variables that we often reify in science.

Representational measurement is conventionally accomplished by mapping measurements to quantitative attributes of real objects (Hand 1996). A potential solution to the difficulties of psychological measurement (Michell 1999) is to base workload assessment on physiological attributes of brain-functioning, such as activity of brain regions activated by task demands. This approach is worth pursuing, but difficulties include the constructive nature of subjective experience (Annett 2002), the complex interconnectivity of different brain areas and the weakness of associations between subjective workload and psychophysiological measures in empirical data (Matthews et al. 2015b). The foundational resource theory for workload (Young et al. 2015) suggests that measurement units might be virtual rather than physical, i.e. the output characteristics of units (e.g. processing rate) within a computational processing architecture (e.g. Anderson 2007). We will identify difficulties for a resource-based approach, and introduce a further possibility, that subjective experience of workload reflects personal interpretations of task demands. Cognitive appraisal theory of emotion provides a useful parallel (Scherer 2009). Emotional response to a stimulus reflects a sequence of evaluations, for novelty, intrinsic pleasantness, goal significance and others. The process can be represented with a computational architecture (Scherer 2009). Similarly, workload might be constructed from evaluations of task demands and their relevance to personal attitudes and goals. Because this is a novel perspective, we must first identify the underlying psychological constructs that may shape subjective workload. In this section, we outline contemporary psychometric test standards that guide identification and validation of constructs and outline their application to identifying a self-regulative basis for subjective workload.

3.1. Contemporary standards for test validity

Sometimes lost in the debate over subjective workload measures is the evolution of perspectives on psychometric test validity. Traditionally, validity was defined as a property of the test itself, inferred primarily from correlations with criterion variables (see Geisinger 1992). The more modern view incorporated into the current *Standards for Educational and Psychological Testing* (APA, AERA, NCME 2014) defines validity as the extent to which evidence and theory support the interpretations of test results entailed by proposed uses of the test. The focus on uses of the test is amenable to applications of psychology. Indeed, the definition appears to represent a shift from a purely representational perspective to one that echoes Hand's (2004) operational measurement, albeit with greater reference to theory as an influential factor in test score interpretation. We thus explicitly acknowledge that controversy continues over the meaning of 'validity' (e.g. Newton and Baird 2016). However, the *Standards* provide a useful stance for re-evaluating the validity of subjective workload scales.

The traditional interpretation of workload would be that scale scores reflect, at some level of validity, a universally-applicable parameter of cognitive, or neurocognitive functioning. This might be thought akin to a tachometer displaying a vehicle engine's RPM. The modern

perspective is that the validity of workload scales should be evaluated in relation to scale usages, such as determining a redline threshold for performance deterioration within a specific task domain (Grier et al. 2008). The case for such usage requires reference to both empirical studies linking workload scores to performance, and a theoretical argument, e.g. that scores reflect some form of resource shortfall. The advantage of a theory-based case for usage is that it can support generalization of findings beyond the immediate operator-task-environment configuration used to establish a redline. For example, if the issue is how many separate, similar gauges can be included in an interface, general resource theory (Kahneman 1973) could support guidelines for an upper limit to that number of displays. In contrast, multiple resource theory (Wickens 2008) predicts that the upper limit could be increased by presenting some information via differing sensory modes. That is, the specific theoretical interpretation of the evidence supporting a test usage makes a difference to application.

3.2. Types of evidence for construct validation

The next logical question to emerge is what theoretical interpretations of workload scores can be justified in relation to typical scale usages. Test standards (APA, AERA, NCME 2014) help to answer this question by discriminating five types of evidence relevant to justification of validity. Each type of evidence suggests critical questions that the researcher can answer in relation to data from studies of the test, as summarized in Table 2. That is, to justify the use of a test, the researcher must provide evidence to answer each question affirmatively.

In the case of workload measures, the resource/demand definition of workload (Young et al. 2015) implies that a primary usage of the test is to determine vulnerability to performance decrement as task demands increase, i.e. the point at which resources become insufficient. For illustrative purposes, we will consider the validity for this purpose of subjective measures, exemplified by the NASA TLX (Hart and Staveland 1988) in relation to the test standards. We will briefly sketch out how well these measures perform against each standard, finding that some types of evidence may not in fact support the usage of subjective workload measures to determine resource insufficiency. We then propose an alternative theoretical perspective on subjective workload, that it reflects self-appraisals of being mentally taxed, and evaluate this perspective in relation to the five types of evidence. For the purpose of the present work, the aim is to elucidate future directions in workload assessment and the types of evidence that may be needful for more representational measurement. We do not aim to articulate a new and comprehensive workload theory which would take considerably more space than available.

1. **Evidence based on test content.** Evidence of this kind is based on rational analysis of test items, including expert evaluations, and featured quite strongly in Hart and Staveland's (1988) initial report on the NASA-TLX. The challenge for a resource interpretation of scores is inferring resource usage from self-reports. The general difficulties of introspection into cognitive processes are reinforced by the issues relating to measure 'dissociations' as we discussed above (and see Hancock 2017; Vidulich and Tsang 2012).
2. **Evidence based on response processes.** The issue here is whether response is biased by factors extraneous to the construct of interest (i.e. resource allocation). For self-report measures, biases due to response styles such as acquiescence and to social

desirability loom large, especially in practical settings where the respondent has a personal stake in the outcome. Where validation evidence is provided by correlations between self-report workload scales and other self-report measures, there may be a concern that correlations are inflated by common method variance (Podsakoff, MacKenzie, Lee, and Podsakoff 2003).

3. **Evidence based on internal structure.** This criterion requires that relationships among test items are consistent with the theorized internal structure of the construct that guides test score interpretations. Calculation of the internal consistency of scale provides a rough check that item scores are compatible with a single underlying construct. Internal consistency has been established for the NASA-TLX, based on the tendency for the ratings to (mostly) intercorrelate positively (Hart and Staveland 1988). However, research has generally neglected more powerful analytic techniques derived from item response theory and Rasch scaling. Latent factor modeling via confirmatory factor analysis (CFA) can separate the latent construct from the measured indicator variables such as test items that are used to infer the latent factor. For practical reasons, subjective workload scales are typically short. This feature hinders comprehensive, systematic sampling of the full range of self-perceptions that may be indicative of subjective workload. One line of work that has used CFA concerns the Paas Cognitive Load Scale (Leppink et al. 2013), an assessment developed primarily for education settings based on a theory of the different types of load experienced by learners. Consistent with theory, Leppink et al. (2013) reported a CFA that identified three correlated factors for intrinsic load from the task, extraneous load (demands of maladaptive instructional features) and germane load (demands of adaptive instructional features). Thus, a higher-level unitary factor emerged, but there was a multifactorial lower-level structure. Furthermore, the higher-level factor did not converge well with workload measured by the NASA-TLX (Naismith, Cheung, Ringsted, and Cavalcanti 2015). More research is necessary to explore the dimensional structure of workload self-assessments.
4. **Evidence based on relations to other variables.** This category of evidence subsumes much of what is traditionally thought of as validity evidence, including correlations of the scale with other variables ('nomological net'), influence of experimental manipulations and natural analogues, and criterion group effects such as expert-novice differences. Most of the evidence of this type for subjective workload scales comes from experimental studies, including evidence for both sensitivity and for dissociations as previously discussed. For other major psychological constructs, there is usually considerably more correlational evidence, including convergence with alternate measures and divergence from distinct constructs. Supporting validity, some subjective scales tend to intercorrelate quite highly (Rubio, Díaz, Martín, and Puente 2004), although some studies show weaker convergence (Funke et al. 2013). Evidence on workload correlations with stress measures has also been reported (Matthews et al. 2002). However, to substantiate a resource/demand interpretation, we would need to show large correlations between subjective workload and performance measures under resource-limited conditions, especially those directly linked to resource utilization such as secondary probe RT measures. While correlations are sometimes found, they do not suggest strong convergence, as highlighted in Table 1. The modest level of association between subjective workload and performance is consistent with broader

concerns about the suitability of subjective measures for predicting objective behaviors (e.g. Af Wåhlberg and Dorn 2015; De Winter, Dodou, and Hancock 2015).

5. **Evidence based on the consequences of testing.** This criterion refers to evidence that benefits expected from test usage are actually realized. It may also cover unintended negative consequences. Given their applied roots, workload measures hold up well against this standard; for example, Parasuraman et al. (2008) provide examples of how workload assessment has improved system design in aircraft. In this case, evidence supports a resource interpretation for predicting outcomes in multi-tasking environments, though favoring a multiple rather than unitary resource theory (Wickens, 2008).

Taken together, the *Standards* provide a roadmap for usage-focused validation of subjective workload scales. For example, redline determination requires evidence that (1) scale content is appropriate for the task domain, (2) scores are not contaminated by social desirability or other artifacts, (3) scores are psychometrically consistent with an underlying latent factor, (4) scores predict performance and objective indices of resource utilization consistent with theory, and (5) incorporating workload assessment into decisions on system design leads to observable improvements in operator performance, safety or other consequential outcomes. Standards 2–3 are generally challenging for subjective workload measure interpretation, in part because of lack of evidence and attention to the relevant issues. Standards 1 and 4 raise additional issues for a resource/demand theory interpretation. The relative success of workload measurement versus standard 5 (Parasuraman et al. 2008) may provide motivation to find stronger evidence in relation to the preceding standards.

Application of the *Standards* may support both operational and representational approaches to measurement. From the operational perspective, all standards may contribute to improve the pragmatic utility of the measure, but standards 4 and 5 are critical, because they require relationships between the workload measure and outcomes that justify real-world, consequential use of the measure. From the representational perspective, validation according to the standards does not in itself provide representational measurement, but it is a necessary first step. In particular, representational measurement requires selection of content according to a theory of what the measurement units are (standard 1), and evidence on internal structure (standard 3) and associations with other constructs (standard 4) to test the theory.

We do not under-estimate the difficulties of such a project. Fayers and Hand (2002) point out that identification of constructs through latent factor modeling (standard 3) is problematic if the measures analyzed include both direct indicators of the construct and causal influences on the construct. For example, in the case of the NASA TLX, should the mental demand rating be considered an index of the external task load that is an external cause of subjective workload, or an index of the subjective experience that is integral to the construct? Another issue discussed by Fayers and Hand (2002) is the weighting of component scores in calculating a measure: the procedure for weighting NASA TLX ratings to calculate overall workload (Hart and Staveland 1988) reflects such concerns. However, there could be many possible quantitative algorithms for representing the construction of subjectively experienced workload from component processes. Developing and testing real-time computational architectures for the constructive process, as for emotional experience (Scherer 2009), provides one way forward.

3.3. Subjective workload as a self-regulative construct

We propose that self-report workload assessments may primarily index aspects of self-regulation, i.e. pursuit of goal-directed behavior that is guided by appraisals of the current status of the self (Carver and Scheier 2001). Theories of self-regulation (e.g. Carver and Scheier 2001; Seufert 2018; Wells and Matthews 2015; Zimmerman 2005) differentiate monitoring and control aspects. People evaluate their mental states and functioning and may try to influence mental state through regulative strategies, including cognitive reappraisal, suppression of unpleasant thoughts or indirect control via changing external contingencies (Ochsner and Gross 2008). Monitoring qualities such as level of demands, effort and performance effectiveness is a form of metacognition i.e. awareness and appraisal of one's mental functioning (Fleming and Lau 2014). Metacognitions can be irrational and subject to various biases (Wells and Matthews 2015), so that it is naïve to suppose that people can readily 'read off' their level of resource investment in processing. In the performance context, people evaluate their level of performance, causal factors contributing to performance, and how it relates to explicit standards or goals (Zimmerman 2005). If performance falls short of target standards, the person chooses between strategies for reducing the discrepancy, including increasing effort, trying a different strategy, lowering standards, seeking assistance, or finding a different path towards the end-goal that performance supports (e.g. requesting a different work assignment).

Workload is a focus for self-monitoring; both low and high levels of workload tend to be uncomfortable (Hancock and Warm 1989). Workload also signifies other important information about the self in the context in which the task is being performed. In a work context, moderate workload ('I am fulfilling my duties') may be preferred to low workload ('I am goofing off my job') and high workload ('I am being exploited'). These evaluations are context-dependent; in some occupations very high workload is expected or valued. Similarly, workload is also a potential driver of control efforts. There is likely a general homeostatic tendency to avoid extremes, as well as more contextually-shaped efforts to regulate workload. In the occupational context, for example, high workload might variously motivate attempts to reappraise the load as a sign of one's value to the company, to simplify the task by cutting corners, to offload work onto others or to request a pay increase. Thus, a person making a workload estimate is not making a dispassionate measurement detached from personal concerns, like reading the temperature from a thermometer, but a self-judgment that may be freighted with emotion and motivation. Monotonous tasks such as vigilance may be stressful because the cognitive work required to sustain attention is typically perceived as disproportionate to the low extrinsic and intrinsic motivations for performance (cf., Hancock 2013).

Existing work on self-regulation has already identified and validated a range of constructs relevant to performance environments, including stress, appraisal, coping and emotion-regulation (Wells and Matthews 2015), so that there is already a psychometric landscape within which workload may be located. Subjective workload scales may be conceptualized in relation to three components of self-regulation that follow in logical sequence (Hofmann, Schmeichel, and Baddeley 2012). People monitor their thoughts, feelings and behaviors in relation to normative standards, they are motivated to reduce discrepancy between actual and target states, and they invest processing capacity into discrepancy-reducing activities. The content of subjective workload scales corresponds to Hofmann et al.'s (2012) monitoring component; items ask people to rate their mental state. Current subjective measures may be poorer at capturing strategies for reducing discrepancy and their performance

consequences. As indicated, we do not aim to present a self-regulative theory of subjective workload, but we will outline how each type of evidence could be pursued.

1. **Evidence based on test content.** Experts would probably agree that scale items represent self-appraisal, but this standard also suggests a wider sampling domain for workload-relevant constructs. Specifically, we could assess further the extent to which workload is perceived as externally-imposed or voluntarily-chosen, how the person would rate the ideal workload for a task, and the extent to which the person is using workload as a proxy for a performance standard.
2. **Evidence based on response processes.** Studies of this kind are generally lacking in the workload literature; self-regulative theory provides a framework for work of this kind. One insight from existing research is that some response distortions have a functional purpose, such as self-enhancement or conformity with others (Paulhus and John 1998). These may be important questions to ask about workload. Does a high rating for effort reflect a self-motivating strategy, as in the self-talk athletes employ to enhance performance? Or is it a delusion of the lazy individual? The answer might be different according to the intended usage of the scale.
3. **Evidence based on internal structure.** Again, this standard primarily provides a call for further research, with an expanded set of workload items sampled according to theory. For example, Hofmann et al.'s (2012) analysis suggests the need to elaborate self-perceptions of standards, motivations and capacity. Confirmatory factor analysis is already widely used in studies of scales for self-regulative constructs. An advantage of conceptualizing workload as a construct of this kind is that tests for internal structure can be designed to confirm that workload is a homogeneous element of self-perception distinct from already-validated constructs (addressing standard 4).
4. **Evidence based on relations to other variables.** In terms of correlational studies, evidence already exists that relates workload to self-regulative constructs. Table 3 shows correlations between NASA TLX score (unweighted mean) and independently validated scales for appraisal and coping, constructs central to cognitive stress theory (Lazarus 1999), for three of the studies previously cited in Table 1. In two of these studies (Matthews et al. 2014; Neubauer et al. 2012), overall workload was unrelated to performance. By contrast, the appraisal and coping data showed fairly consistent patterns of association across studies. Subjective workload was associated with both adaptive (challenge appraisal, task-focused coping) and maladaptive (threat appraisal, low controllability and

Table 3. Correlations between NASA-TLX workload and basic dimensions of situational appraisal and coping; example studies.

			Appraisal			Coping		
Study	*N*	Task	Threat	Challenge	Control.	Task-Focus	Emotion-Focus	Avoidance
Matthews et al. (2014)	342	12 min sensory vigilance task	.31**	.29**	−.37**	.11*	.20**	−.07
Fellner (2008)	121	Discrimination learning (collaborating 2-person teams)	.30**	.31**	−.31**	.30**	.46**	.10
Neubauer et al. (2012)	173	Simulated driving: fatiguing conditions	.45**	.16*	−.29**	.19*	.24**	.05

*$p < .05$, **$p < .01$.
Control. = Perceived controllability.

emotion-focused coping) cognitive processing of task stressors. Again, details of these findings are beyond the present scope, but the data show how workload is systematically related to other self-relevant constructs that define how the person evaluates and deals with stressful encounters. The self-regulative perspective is consistent with the view that subjective experience is actively constructed via cognitive appraisal (not necessarily conscious) of multiple external and internal cues (Annett 2002).

5. **Evidence based on the consequences of testing.** Ergonomics typically and appropriately focused on gains in performance and safety realized via workload assessment (Parasuraman et al. 2008). The self-regulative perspective suggests a wider range of consequences, especially for occupational health and well-being. Research might examine further the utility of workload assessment for predicting outcomes such as job satisfaction, counterproductive behaviors, and burnout.

In framing workload as a self-regulative construct, we advocate the importance of latent factor modeling for future subjective assessments. Latent factor modeling may not satisfy psychometric purists (Barrett 2005), but its separation of the latent construct from observed measurements increases the likelihood that we can identify measurement models that generalize across different constructs. Such constructs are representational to the extent that we can define a consensual methodology for assessment of theoretically-derived components or indicators of workload, guided by the validation *Standards*. Annett (2002) pointed out that the critical issue for measurements in ergonomics, whether objective or self-report is inter-subjectivity, i.e. the degree of shared meaning between independent observers. Fusing modern psychometrics with self-regulative theory may accomplish inter-subjectivity.

4. Conclusions

The central problem addressed in this article is how to interpret subjective workload measures that fail to converge strongly with objective indices. Researchers may need to abandon the convenient but questionable assumption that alternate workload measures are essentially interchangeable, give or take some measurement issues specific to each type of measure. We conclude that the various available workload measures assess not one but several distinct constructs. The multiplicity of constructs implies that some re-thinking of the place of workload measures in ergonomics is necessary. We have proposed alternate solutions to the divergence problem based on operational and representational measurement perspectives. Treating workload operationally will often be sufficient for the practitioner, whereas building a theory of psychological workload potentially accommodating multiple latent dimensions of the construct requires confronting the challenges of a more representational perspective. We finish by considering future research directions and the implications of a non-unitary view of workload for applications in ergonomics.

4.1. Future research directions

The operational perspective frees researchers and practitioners to identify the best algorithms for predicting practically significant outcomes such as safety and productivity irrespective of conventional construct validity. Thus, divergence of measures is not a critical shortcoming, provided that evidence supports the practical usage of a given assessment in

a given domain or context. There is considerable scope for 'big data' approaches to workload assessment. Advances in unobtrusive and wearable sensors imply that rich data sets can be acquired from populations such as vehicle drivers and office workers, subject to appropriate legal and ethical constraints. Analysis of relationships between such data and performance and health outcomes can identify predictors of adverse outcomes such as errors, violations and absenteeism. For example, the neural net of a future automated driving system may be able to predict likelihood of driver error from inputs such as physiological sensors, vehicle control responses and analysis of ambient distractors without ever having to compute workload explicitly. Existing studies of algorithms that aggregate data from multiple sources to identify workload have utilized a range of machine learning classifiers including artificial neural networks, linear regression, linear discriminant analysis and support vector machines (Heard, Harriott, and Adams 2018). Algorithms can also be personalized to reflect individual variation in the responses most sensitive to workload, in effect assessing workload on a within-rather than a between-subjects basis (Teo et al. 2018). In any case, the focus is on validating the algorithm as means for predicting a significant real-world outcome, rather than identifying any latent workload construct.

Current algorithms have various shortcomings described by Heard et al. (2018) including limited generalizability, limited sampling of workload components and lack of verification in practical settings. They may also be difficult to interpret in relation to extant psychological and neuroscience theory. Nevertheless, if algorithms fulfill the operational requirement of predicting real-world outcomes, it may not matter, at least to the ergonomics practitioner, if human factors construct including workload are replaced by computational, data-driven models.

Turning to prospects for representational measurement, the APA, AERA, NCME (2014) *Standards* provide an outline road-map for validating subjective workload as a latent construct or constructs associated with self-regulation. Such constructs may shift measurement towards the representational end of Hand's (2004) continuum, although strongly representational measurement of psychological constructs is hard to attain. Conceptualizing subjective workload as an element of metacognition (Fleming and Lau 2014) allows the construct to be defined independently from its neurological and information-processing concomitants. From this perspective, evidence on convergence and divergence from existing constructs is an important step in validation. Linking existing scales such as the NASA-TLX to latent constructs requires a more systematic approach to construct validation than has previously been conducted.

There remains scope for exploring objective representational assessments for workload constructs. A physiological approach has the advantage of working with physical quantities that are ratio-scaled such as voltages, response latencies and energy consumption. Advancements in functional neuroimaging might eventually support measurement of a unitary 'brain-workload' construct, although the divergence of alternative physiological measures (Matthews et al. 2015a) suggests challenges ahead. Regardless of whether workload in this sense proves to be unitary or multidimensional, the deeper issue is how to relate physical brain-based metrics to virtual information-processing constructs such as resources, and ultimately to performance, in a theoretically coherent fashion.

Improved cognitive neuroscience models may bridge the gap between brain physiology and behavior. For example, Estes (2015) suggested using Anderson's (2007) ACT-R model to predict the activation functions associated with increasing task demands, Estes' (2015)

empirical findings show a close correspondence between the curvilinear activation function and subjective workload to increasing working memory load. The cognitive architecture of ACT-R comprises multiple modules, associated with different brain areas, which might provide a basis for workload conceptualizations associated with multiple resource theory (Wickens 2008). That is, computational modeling may provide a theoretical basis for understanding workload factors that supports representational measurement, i.e. the workload score is interpreted as the activation level of the relevant ACT-R module.

4.2. Practical implications

The popular resource/demand conception of mental workload (Young et al. 2015) provides a convenient cognitive heuristic for researchers. It is simple to understand and apply and coheres with easily-accessible instances of performance failures associated with overload (Matthews, Lin, and Wohleber, in press). Unfortunately, the metaphor is over-simple, as shown by the AIDs of workload (Hancock 2017). How does the practitioner adapt to a more complex scientific reality? There is no simple answer, but we conclude with four general suggestions for practitioners. In each case, there are roles for both operational and representational perspectives.

1. **Define the purpose of the assessment.** The modern conception of test validity, relevant to operational as well as to representational measurement, is that an evidence-based argument must be made for specific purposes for test use (Kane 2016). Thus, clarity of purpose is essential for choosing between the different and potentially divergent workload assessments available. Common purposes for researchers include investigating allocation of attention and validating cognitive and neuroergonomic models of performance (Matthews and Reinerman-Jones 2017). Practitioners address issues such as determining workload redlines, evaluating user comfort and technology acceptance, and predicting loss of productivity. The operational approach to measurement may suffice in each case, but the need to cite data to support the proposed workload usage is paramount. Purposes may include generalization of findings to novel systems, for example, when the technology supporting performance is developing fast. In this case, a theory-based argument assumes greater importance.
2. **Consider a multivariate assessment.** Poor convergence implies that, by and large, adding additional, reliable assessments provides novel information about operator state. From a big data perspective, the more independent sources of information there are the better, assuming the algorithm for workload determination remains computable. From a theory-driven perspective, a multivariate assessment strategy allows for a more complete picture of changes in operator state that may be driven by task demands (Matthews and Reinerman-Jones 2017). Workload response may be more easily interpreted if related but distinct factors such as stress, task engagement and trust are also assessed. For example, increased NASA-TLX scores may be accompanied by higher task engagement if the task is appraised as an enjoyable challenge, but decreased engagement if the task appears pointless or impossible to perform adequately (Abich et al. 2017). Multivariate assessment may also contribute to disambiguating other popular ergonomics constructs for which there are sometimes

dissociations between multiple subjective and objective measures including stress (Matthews et al. 2017), situation awareness (Salmon, Stanton, Walker, and Green 2006) and trust (Chancey, Bliss, Proaps, and Madhavan 2015).

3. **Craft communication of findings.** In practical settings, workload assessment takes place within a social context that may influence study design. The practitioner may need to use workload findings to support a narrative for taking further steps to enhance safety or productivity within the organization concerned. The choice of measures can then take into account the intended audience for the narrative. For example, data may show that both subjective and physiological measures are sensitive to the demand factor of interest, but the subjective measure has higher sensitivity. However, the audience may be skeptical of self-reports. In this case, both types of measures could be included in the workload evaluation, so that changes in the objective measure lend credibility to the subjective measure. From a broader perspective, the practitioner must also communicate the case for consequential validity (APA, AERA, NCME 2014), i.e. how taking decisions based on the workload assessment will lead to desirable outcomes.
4. **Adapt to technological change.** Greater sophistication in workload assessment is important for dealing with the challenges of future technologies. A good example is our coming interaction with autonomous systems. At first, autonomous systems, almost by definition, seem to require no human input, and hence impose minimal workload. But this is not so. If currently conceived architectures are brought to fruition, then humans will team with autonomy, albeit on differing levels of interaction frequency. Much depends on understanding the rate-limiting, human bandwidth in such a pairing. It is here that more precise and targeted workload assessment steps to the fore. Future assessments may measure not just quantitative levels of subjective and objective response, but also address their covariation with interest, creativity, and enjoyment. These hedonomic aspects are liable to feature more and more as the very nature of collaborative activity evolves with conjoined technical capacities. In the end this may not be conventional workload assessment at all, but rather indexing the quality and value of interaction as the nature of work itself radically evolves.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abich, J., L. E. Reinerman-Jones, and G. Matthews. 2017. "Impact of Three Workload Factors on Simulated Unmanned System Intelligence, Surveillance, and Reconnaissance Operations." *Ergonomics* 60 (6): 791–809.

Af Wåhlberg, A. E., and L. Dorn. 2015. "How Reliable Are Self-Report Measures of Mileage, Violations and Crashes?" *Safety Science* 76: 67–73.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (APA, AERA, NCME). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, J. R. 2007. *How Can the Human Mind Occur in the Physical Universe*? New York: Oxford University Press.
Annett, J. 2002. "Subjective Rating Scales: Science or Art?" *Ergonomics* 45 (14): 966–987.
Barrett, P. 2005. "What If There Were No Psychometrics? Constructs, Complexity, and Measurement." *Journal of Personality Assessment* 85 (2): 134–140.
Boles, D. B., J. H. Bursk, J. B. Phillips, and J. R. Perdelwitz. 2007. "Predicting Dual-Task Performance with the Multiple Resources Questionnaire (MRQ)." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49 (1): 32–45.
Bridgman, P. W. 1927. *The Logic of Modern Physics*. New York: Macmillan.
Burt, K. B., and J. Obradović. 2013. "The Construct of Psychophysiological Reactivity: Statistical and Psychometric Issues." *Developmental Review* 33 (1): 29–57.
Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews. Neuroscience* 14 (5): 365–376.
Campbell, D. T., and D. W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (2): 81–105.
Carver, C. S., and M. F. Scheier. 2001. *On the Self-Regulation of Behavior*. New York: Cambridge University Press.
Chancey, E. T., J. P. Bliss, A. B. Proaps, and P. Madhavan. 2015. "The Role of Trust as a Mediator between System Characteristics and Response Behaviors." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (6): 947–958.
Damos, D. L. 1988. "Individual Differences in Subjective Estimates of Workload." In *Human Mental Workload*, edited by P. A. Hancock and N. Meshkati, 231–237. Amsterdam: North-Holland.
Dekker, S. W. 2015. "The Danger of Losing Situation Awareness." *Cognition, Technology & Work* 17 (2): 159–161.
Dekker, S., and E. Hollnagel. 2004. "Human Factors and Folk Models." *Cognition, Technology and Work* 6 (2): 79–86.
Dekker, S. W., and J. M. Nyce. 2015. "From Figments to Figures: Ontological Alchemy in Human Factors Research." *Cognition, Technology & Work* 17 (2): 185–187.
de Groot, S., F. Centeno Ricote, and J. C. F. de Winter. 2012. "The Effect of Tire Grip on Learning Driving Skill and Driving Style: A Driving Simulator Study." *Transportation Research Part F: Traffic Psychology and Behaviour* 15 (4): 413–426.
De Waard, D. 1996. "*The Measurement of Drivers' Mental Workload*." Doctoral diss., University of Groningen, the Netherlands.
De Waard, D., and K. A. Brookhuis. 1997. "On the Measurement of Driver Mental Workload." In *Traffic and Transport Psychology*, edited by J. A. Rothengatter and E. Carbonell, 161–173. Amsterdam: Elsevier.
De Waard, D., and B. Lewis-Evans. 2014. "Self-Report Scales Alone Cannot Capture Mental Workload." *Cognition, Technology & Work* 16 (3): 303–305.
De Winter, J. C. F. 2014. "Controversy in Human Factors Constructs and the Explosive Use of the NASA-TLX: A Measurement Perspective." *Cognition, Technology & Work* 16 (3): 289–297.
De Winter, J. C. F., D. Dodou, and P. A. Hancock. 2015. "On the Paradoxical Decrease of Self-Reported Cognitive Failures with Age." *Ergonomics* 58 (9): 1471–1486.
Eichinger, A., and K. Bengler. 2014. "Representations and Operations: Parts of the Problem and the Solution." *Cognition, Technology & Work* 16 (3): 307–310.
Estes, S. 2015. "The Workload Curve: Subjective Mental Workload." *Human Factors* 57 (7): 1174–1187.
Fayers, P. M., and D. J. Hand. 2002. "Causal Variables, Indicator Variables and Measurement Scales: An Example from Quality of Life." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165 (2): 233–253.
Fellner, A. N. 2008. "*The Effects of Emotional Intelligence on Performance of a Cognitive Task in the Context of Collaboration vs. Competition*." Doctoral diss., University of Cincinnati.
Fleming, S. M., and H. C. Lau. 2014. "How to Measure Metacognition." *Frontiers in Human Neuroscience* 8: 443.

Fuller, R. 2005. "Towards a General Theory of Driver Behaviour." *Accident, Analysis and Prevention* 37 (3): 461–472.
Funke, G., B. Knott, V. F. Mancuso, A. Strang, J. Estepp, L. Menke, B. Miller. 2013. "Evaluation of Subjective and EEG-Based Measures of Mental Workload." In *International Conference on Human-Computer Interaction*, 412–416. Berlin: Springer.
Gawron, V. J. 2008. *Human Performance, Workload and Situation Awareness Measures Handbook*. 2nd ed. Boca Raton, FL: CRC Press.
Geisinger, K. F. 1992. "The Metamorphosis of Test Validation." *Educational Psychologist* 27 (2): 197–222.
Gerjets, P., K. Scheiter, and G. Cierniak. 2009. "The Scientific Value of Cognitive Load Theory: A Research Agenda Based on the Structuralist View of Theories." *Educational Psychology Review* 21 (1): 43–54.
Grier, R., C. Wickens, D. Kaber, D. Strayer, D. Boehm-Davis, J. G. Trafton, and M. St. John. 2008. "The Red-Line of Workload: Theory, Research, and Design." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52 (18): 1204–1208.
Hancock, P. A. 1996. "Effects of Control Order, Augmented Feedback, Input Device and Practice on Tracking Performance and Perceived Workload." *Ergonomics* 39 (9): 1146–1162.
Hancock, P. A. 2013. "In Search of Vigilance: The Problem of Iatrogenically Created Psychological Phenomena." *American Psychologist* 68 (2): 97–109.
Hancock, P. A. 2017. "Whither Workload? Mapping a Path for its Future Development." In *Human Mental Workload: Models and Applications*, edited by L. Longo and M. Chiara Leva. Cham, Switzerland: Springer International.
Hancock, P. A., and G. Matthews. 2018. "Workload and Performance: Associations, Insensitivities and Dissociations." *Human Factors*, in press. doi:10.1177/0018720818809590
Hancock, P. A., T. Sanders, and W. Volante. 2015. "Quantitative Assessment of Qualitative Statements: What People Mean by Once in a Blue Moon." In *Proceedings of the 19th Triennial Congress of the International Ergonomics Association*. Melbourne, Australia.
Hancock, P. A., and J. S. Warm. 1989. "A Dynamic Model of Stress and Sustained Attention." *Human Factors* 31 (5): 519–537.
Hand, D. J. 1996. "Statistics and the Theory of Measurement." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159 (3): 445–492.
Hand, D. J. 2004. *Measurement: Theory and Practice*. London: Arnold.
Hart, S. G., and L. E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In *Human Mental Workload*, edited by P. A. Hancock and N. Meshkati, 139–184. Amsterdam: North-Holland.
Hart, S. G., and C. D. Wickens. 2010. "Cognitive Workload." In *NASA Human Integration Design Handbook (HIDH)*, 190–222. Washington, DC: NASA.
Heard, J., C. E. Harriott, and J. A. Adams. 2018. "A Survey of Workload Assessment Algorithms." *IEEE Transactions on Human-Machine Systems* 99: 1–18.
Hockey, G. R. J., P. Nickel, A. C. Roberts, and M. H. Roberts. 2009. "Sensitivity of Candidate Markers of Psychophysiological Strain to Cyclical Changes in Manual Control Load during Simulated Process Control." *Applied Ergonomics* 40 (6): 1011–1018.
Hofmann, W., B. J. Schmeichel, and A. D. Baddeley. 2012. "Executive Functions and Self-Regulation." *Trends in Cognitive Sciences* 16 (3): 174–180.
Hwang, S. L., Y. J. Yau, Y. T. Lin, J. H. Chen, T. H. Huang, T. C. Yenn, and C. C. Hsu. 2008. "Predicting Work Performance in Nuclear Power Plants." *Safety Science* 46 (7): 1115–1124.
Jansen, R. J., B. D. Sawyer, T. van Egmond, H. de Ridder, and P. A. Hancock. 2016. "Hysteresis in Mental Workload and Task Performance." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 58 (8): 1143–1157.
Jensen, A. R. 2006. *Clocking the Mind: Mental Chronometry and Individual Differences*. Amsterdam: Elsevier.
Kahneman, D. 1973. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.
Kane, M. T. 2016. "Explicating Validity." *Assessment in Education: Principles, Policy and Practice* 23 (2): 198–211.

Langner, R., and S. B. Eickhoff. 2013. "Sustaining Attention to Simple Tasks: A Meta-Analytic Review of the Neural Mechanisms of Vigilant Attention." *Psychological Bulletin* 139 (4): 870–900.

Lazarus, R. S. 1999. *Stress and Emotion: A New Synthesis*. New York: Springer.

Lee, Y. H., and B. S. Liu. 2003. "Inflight Workload Assessment: Comparison of Subjective and Physiological Measurements." *Aviation, Space, and Environmental Medicine* 74: 1078–1084.

Leppink, J., F. Paas, C. P. Van der Vleuten, T. Van Gog, and J. J. Van Merriënboer. 2013. "Development of an Instrument for Measuring Different Types of Cognitive Load." *Behavior Research Methods* 45 (4): 1058–1072.

Liu, H., J. Fan, Y. Fu, and F. Liu. 2018. "Intrinsic Motivation as a Mediator of the Relationship between Organizational Support and Quantitative Workload and Work-Related Fatigue." *Human Factors and Ergonomics in Manufacturing & Service Industries* 28 (3): 154–162.

Matthews, G., and S. E. Campbell. 2010. "Dynamic Relationships between Stress States and Working Memory." *Cognition and Emotion* 24 (2): 357–373.

Matthews, G., S. E. Campbell, S. Falconer, L. Joyner, J. Huggins, K. Gilliland, R. Grier, and J. S. Warm. 2002. "Fundamental Dimensions of Subjective State in Performance Settings: Task Engagement, Distress and Worry." *Emotion* 2 (4): 315–340.

Matthews, G., J. Lin, and R. Wohleber. in press. "Stress, Skilled Performance, and Expertise: Overload and Beyond." In *Oxford Handbook of Expertise*, edited by P. Ward, J. M. Schraagen, J. Gore, and E. Roth. New York: Oxford.

Matthews, G., and L. Reinerman-Jones. 2017. *Workload Assessment: How to Diagnose Workload Issues and Enhance Performance*. Santa Monica, CA: Human Factors and Ergonomics Society.

Matthews, G., L. Reinerman-Jones, J. Abich, and A. Kustubayeva. 2017. "Metrics for Individual Differences in EEG Response to Cognitive Workload: Optimizing Performance Prediction." *Personality and Individual Differences* 118: 22–28.

Matthews, G., L. E. Reinerman-Jones, D. J. Barber, and J. Abich. 2015. "The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (1): 125–143.

Matthews, G., L. E. Reinerman-Jones, R. Wohleber, J. Lin, J. Mercado, and J. Abich. 2015b. "Workload is Multidimensional, Not Unitary: What Now?" In *Foundations of Augmented Cognition*, edited by D. D. Schmorrow and C. M. Fidopiastis, 44–55. New York: Springer International.

Matthews, G., J. S. Warm, L. E. Reinerman-Jones, L. K. Langheim, D. A. Washburn, and L. Tripp. 2010. "Task Engagement, Cerebral Blood Flow Velocity, and Diagnostic Monitoring for Sustained Attention." *Journal of Experimental Psychology: Applied* 16 (2): 187–203.

Matthews, G., J. S. Warm, T. H. Shaw, and V. S. Finomore. 2014. "Predicting Battlefield Vigilance: A Multivariate Approach to Assessment of Attentional Resources." *Ergonomics* 57 (6): 856–875.

Meehl, P. E. 1978. "The Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46 (4): 806–834.

Melman, T., D. A. Abbink, M. M. Van Paassen, E. R. Boer, and J. C. F. De Winter. 2018. "What Determines Drivers' Speed? a Replication of Three Behavioural Adaptation Experiments in a Single Driving Simulator Study." *Ergonomics* 61 (7): 966–987.

Meulenbroek, R. G., G. P. Van Galen, M. Hulstijn, W. Hulstijn, and G. Bloemsaat. 2005. "Muscular Co-Contraction Covaries with Task Load to Control the Flow of Motion in Fine Motor Tasks." *Biological Psychology* 68 (3): 331–352.

Michell, J. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. New York: Cambridge University Press.

Michell, J. 2004. *An Introduction to the Logic of Psychological Measurement*. New York: Psychology Press.

Muckler, F. A., and S. A. Seven. 1992. "Selecting Performance Measures: 'Objective' versus 'Subjective' Measurement." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 34 (4): 441–455.

Myrtek, M., E. Deutschmann-Janicke, H. Strohmaier, W. Zimmermann, S. Lawerenz, G. Brügner, and W. Müller. 1994. "Physical, Mental, Emotional, and Subjective Workload Components in Train Drivers." *Ergonomics* 37 (7): 1195–1203.

Naismith, L. M., J. J. Cheung, C. Ringsted, and R. B. Cavalcanti. 2015. "Limitations of Subjective Cognitive Load Measures in Simulation-Based Procedural Training." *Medical Education* 49 (8): 805–814.
Neubauer, C., L. Langheim, G. Matthews, and D. Saxby. 2012. "Fatigue and Voluntary Utilization of Automation in Simulated Driving." *Human Factors* 54 (5): 734–746.
Newton, P. E., and J. Baird. 2016. "The Great Validity Debate." *Assessment in Education: Principles, Policy and Practice* 23 (2): 173–177.
Ochsner, K. N., and J. J. Gross. 2008. "Cognitive Emotion Regulation: Insights from Social Cognitive and Affective Neuroscience." *Current Directions in Psychological Science* 17 (2): 153–158.
Orr, C. K., and V. G. Duffy. 2007. "Development of a Facial Skin Temperature-Based Methodology for Non-Intrusive Mental Workload Measurement." *Occupational Ergonomics* 7 (2): 83–94.
Parasuraman, R., T. B. Sheridan, and C. D. Wickens. 2008. "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs." *Journal of Cognitive Engineering and Decision Making* 2 (2): 140–160.
Paulhus, D. L., and O. P. John. 1998. "Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles with Basic Traits and Motives." *Journal of Personality* 66 (6): 1025–1060.
Podsakoff, P. M., S. B. MacKenzie, J. Y. Lee, and N. P. Podsakoff. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies." *Journal of Applied Psychology* 88 (5): 879–903.
Proctor, R. W., and A. Xiong. 2017. "The Method of Negative Instruction: Herbert S. Langfeld's and Ludwig R. Geissler's 1910–1913 Insightful Studies." *American Journal of Psychology* 130 (1): 11–21.
Ratcliff, R. 1993. "Methods for Dealing with Reaction Time Outliers." *Psychological Bulletin* 114 (3): 510–532.
Rendón-Vélez, E., P. M. Van Leeuwen, R. Happee, I. Horváth, W. F. Van der Vegte, and J. C. F. De Winter. 2016. "The Effects of Time Pressure on Driver Performance and Physiological Activity: A Driving Simulator Study." *Transportation Research Part F: Traffic Psychology and Behaviour* 41 (Part A): 150–169.
Roscoe, A. H. 1978. "Stress and Workload in Pilots." *Aviation, Space, and Environmental Medicine* 49 (4): 630–636.
Rubio, S., E. Díaz, J. Martín, and J. M. Puente. 2004. "Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods." *Applied Psychology* 53 (1): 61–86.
Salmon, P., N. Stanton, G. Walker, and D. Green. 2006. "Situation Awareness Measurement: A Review of Applicability for C4i Environments." *Applied Ergonomics* 37 (2): 225–238.
Scherer, K. R. 2009. "Emotions Are Emergent Processes: They Require a Dynamic Computational Architecture." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364 (1535): 3459–3474.
Seufert, T. 2018. "The Interplay between Self-Regulation in Learning and Cognitive Load." *Educational Research Review* 24: 116–129.
Stephens, C. L., I. C. Christie, and B. H. Friedman. 2010. "Autonomic Specificity of Basic Emotions: Evidence from Pattern Classification and Cluster Analysis." *Biological Psychology* 84 (3): 463–473.
Stevens, S. S. 1935. "The Operational Basis of Psychology." *The American Journal of Psychology* 47 (2): 323–330.
Stevens, S. S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684): 677–688.
Sweller, J. 1994. "Cognitive Load Theory, Learning Difficulty, and Instructional Design." *Learning and Instruction* 4 (4): 295–312.
Teo, G., L. Reinerman-Jones, G. Matthews, J. Szalma, F. Jentsch, and P. Hancock. 2018. "Enhancing the Effectiveness of Human-Robot Teaming with a Closed-Loop System." *Applied Ergonomics* 67: 91–103.
Van Acker, B. B., D. D. Parmentier, P. Vlerick, and J. Saldien. 2018. "Understanding Mental Workload: From a Clarifying Concept Analysis toward an Implementable Framework." *Cognition, Technology and Work*. 20: 351–365.

Verwey, W. B., and H. A. Veltman. 1996. "Detecting Short Periods of Elevated Workload: A Comparison of Nine Workload Assessment Techniques." *Journal of Experimental Psychology: Applied* 2 (3): 270–285.

Vidulich, M. A., and P. S. Tsang. 2012. "Mental Workload and Situation Awareness." In *Handbook of Human Factors and Ergonomics*, 4th ed., edited by G. Salvendy, 243–273. New York, NY: Wiley.

Warm, J. S., W. Dember, and P. A. Hancock. 1996. "Vigilance and Workload in Automated Systems." In *Automation and Human Performance: Theory and Applications*, edited by R. Parasuraman and M. Mouloua, 183–200. Mahwah, NJ: Erlbaum.

Warm, J. S., L. D. Tripp, G. Matthews, and W. S. Helton. 2012. "Cerebral Hemodynamic Indices of Operator Fatigue in Vigilance." In *Handbook of Operator Fatigue*, edited by G. Matthews, P. A. Desmond, C. Neubauer and P.A. Hancock, 197–207. Aldershot, UK: Ashgate.

Wells, A., and G. Matthews. 2015. *Attention and Emotion: A Clinical Perspective (Classic Edition)*. New York: Psychology Press.

Whitehead, A. N. 1925/2011. *Science and the Modern World*. Cambridge: Cambridge University Press. (Original work published 1925).

Wickens, C. D. 2008. "Multiple Resources and Mental Workload." *Human Factors* 50 (3): 449–455.

Wickens, C. D., J. G. Hollands, S. Banbury, and R. Parasuraman. 2015. *Engineering Psychology and Human Performance*. New York: Psychology Press.

Yeh, Y., and C. D. Wickens. 1988. "Dissociation of Performance and Subjective Measures of Workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 30 (1): 111–120.

Young, M. S., K. A. Brookhuis, C. D. Wickens, and P. A. Hancock. 2015. "State of Science: Mental Workload in Ergonomics." *Ergonomics* 58 (1): 1–17.

Young, M. S., and N. A. Stanton. 2002. "Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 44 (3): 365–375.

Young, M. S., and N. A. Stanton. 2004. "Taking the Load off: Investigations of How Adaptive Cruise Control Affects Mental Workload." *Ergonomics* 47 (9): 1014–1035.

Zhao, X., X. Li, and L. Yao. 2017. "Localized Fluctuant Oscillatory Activity by Working Memory Load: A Simultaneous EEG-fMRI Study." *Frontiers in Behavioral Neuroscience* 11, 215.

Zimmerman, B. J. 2005. "Attaining Self-Regulation: A Social Cognitive Perspective." In *Handbook of Self-Regulation*. (2nd ed.), edited by M. Boekaerts, P.R. Pintrich, and M. Zeidner, 13–39. San Diego: Academic Press.