Cogn Tech Work (2014) 16:303–305
DOI 10.1007/s10111-014-0277-z

COMMENTARY

Self-report scales alone cannot capture mental workload

A reply to De Winter, Controversy in human factors constructs and the explosive use of the NASA TLX: a measurement perspective

Dick de Waard · Ben Lewis-Evans

Published online: 20 May 2014
© Springer-Verlag London 2014

1 Introduction

While one may have the idea that physical workload can be quantified, mental workload certainly cannot. There simply is no SI unit of mental workload. Still, a parallel between physical workload and mental workload can be drawn. What is quantifiable in physical workload are the forces that are needed to perform a task, e.g. to lift an object. However, how physically difficult a task is for someone depends on fitness, training, and so forth. What matters is relative. Specifically, two factors always play a role: on the one hand, the properties of the (mental or physical) task at hand, which we will refer to as the task demands, and on the other hand, the capability of the operator to perform the task, referred to as capacity. Workload in our view is the interaction between these two, in other words, the proportion of capacity that is used to perform a task.

However, mental capacity is not a volume that can be assessed representatively, even though the frequently used term "resource" may give this impression. But even without metrics, it is possible to get an impression of how heavily loaded an operator is. As said, both the task properties and properties of the operator are important for this. Mental workload is the interaction between these two, and it is task difficulty. And in this, we agree with De Winter (2014), that mental workload is assessed by operational measures, just as what IQ tests measure is "IQ", not intelligence or mental capacity. As such, tests for mental workload or IQ only reflect cognitive performance *at the moment of testing*. We also agree with De Winter (2014) that this does not, however, make these operational measures useless.

2 The operator and the task

People differ and are not equally good at performing the same mental task. Apart from innate differences, there are differences in experience. Some tasks can be performed on autopilot, while others require a lot of attention and effort. In this respect, Rasmussen's (1983) and Reason's (1990) division between skill-based, rule-based, and knowledge-based performance is useful (although, again, not a reflection of reality that you can hold and measure, but rather a useful, operational, way to discuss cognition). Reason's division states that knowledge-based performance is effortful and reasoned, rule-based behaviour is less effortful and based on internalised rules, and skill-based performance is mostly automated behaviour. So, if an operator can perform a task on the skill-based level, demands on capacity are low. This means that at the skill level, in general, more tasks can be performed at the same time (i.e. multitasking) without visible deteriorating effects on performance of the skill-based task. Conversely, with knowledge-based task performance, many resources or a large proportion of resources need to be assigned for performance of the task. While in the long term, and with experience, performance can become skill instead of rule or knowledge based and therefore requires fewer resources; shorter-term factors can also have an effect on required resources. After a sleepless night performance on a

This commentary refers to the original article available at: doi:10.1007/s10111-014-0275-1.

D. de Waard (✉) · B. Lewis-Evans
Traffic Psychology Group, Neuropsychology, University of Groningen, Groningen, The Netherlands
e-mail: d.de.waard@rug.nl

B. Lewis-Evans
e-mail: b.lewis.evans@rug.nl

particular task may not be as good as if one is well rested. Being tired, ill, sedated by alcohol or other drugs, or sad, all these state-related factors can have this effect (e.g. De Waard 1996). Individuals can invest extra effort to overcome these short-term factors, increasing mental workload, but maintaining performance (Hockey 1997). That people can act in such a protective fashion in the short term to maintain performance is one of the reasons why an operational measure such as mental workload is useful. In this case, mental workload measurement captures the fact of investing more into a task, thus decreasing an individual's capability to respond to new situations, without noticeable impacts that could be seen if performance alone was monitored.

Furthermore, as far as the operator side is concerned, strategy has an effect on mental workload. Quite often, there are many roads leading to Rome, and not all are equally demanding. The decision to do an acceptable job is less loading than the goal to perform at the ultimate level. This is actually what Hollnagel (2009) calls ETTO, Efficiency Thoroughness Trade Off. ETTO is similar to the speed–accuracy trade-off curve; in that, it is impossible to do most tasks at high speed and at high accuracy. In general, one of the two suffers. More errors are made at high speed, and preventing errors can only be at the cost of speed. With ETTO, thoroughness trades off with efficiency, or in this context, lower mental workload.

3 One measure

Sometimes it seems as if there is a quest to be able to assess mental workload on the basis of one measure, preferably on the basis of task properties. Of course, task properties matter, subtracting 2 from 5 is a less demanding task than dividing 2,315 by 423. However, task complexity is not mental workload, which equals task difficulty, and includes a subjective evaluation. As such, we feel concerned by De Winter's finding (2014) that based on the literature search, the NASA-TLX could start to be seen as synonymous with mental workload. This should not occur.

While we agree with De Winter (2014) that mental workload is an operational concept and not a representational concept, the idea that mental workload can be captured by the use of a questionnaire, and in particular by the use of the NASA-TLX alone, is too simplistic. A situation where workload is synonymous to a TLX score, a situation De Winter describes as becoming reality, is undesirable. Mental workload is a more complex dynamic concept that needs to be assessed by more than just ratings on a subjective scale.

As said, mental workload depends on performance. In very high workload conditions, performance will be affected, but as mentioned these periods can be preceded by periods of performance protection (Hockey 1997) where operators have to try hard, invest effort, but this does not show from the outside, from performance. Self-report measures can reflect this, but not all performance protection is conscious. This shows that one measure does not suffice to capture the complete picture. If one uses only a subjective scale and no other measures, then ratings between conditions can only be compared in within-subject designs. This is because these scales have no actual absolute reference nor are there objective critical levels. Critical levels, however, can be determined for performance (e.g. Brookhuis et al. 2003), which makes assessment of performance along with mental workload measures indispensable.

Another issue of self-report scales is, as De Winter (2014) in our view correctly states, that TLX use is based on researchers simply using what others have used before and not deeply examining the reasons why any particular scale, or scales, should be used. As mentioned by De Winter (2014), the TLX has been questioned on methodological grounds nearly since its inception. For example, Veltman and Gaillard (1996) compared ratings on the TLX with ratings on the Rating Scale Mental Effort (RSME) and found the latter to be more sensitive. However, even if the RSME is more sensitive than the TLX, there are more issues with self-report scales in general that need to be considered. These concerns with self-report include, amongst others, the reliability of reports that are in general created in retrospect, the variation of workload during task performance that cannot be reflected in one rating, and uncertainty over which aspect of mental workload, or some other construct, is actually evaluated in the self-report. Also, the use of popular self-report mental workload questionnaires is mainly limited to Western countries. In other cultures, self-reports may be affected by what is culturally acceptable, e.g. in Eastern cultures, it may be not so easy to state that one had to try hard to complete a task (Widyanti et al. 2013).

In sum, the message should be that although there is no such thing as an attitude, or mental workload, that one can touch, it can be assessed indirectly, as concept. In that we agree with De Winter (2014), although we also see the value of discussing the usefulness of the concept and warning that it should not be seen as a "real" thing that can be touched and quantified. As such, it is perhaps useful that we as practitioners and academics are careful when we talk about such constructs that we do not give the impression to lay-people, students, or ourselves that concepts such as mental workload are anything other than useful, if flawed, operational tools. We also believe that what could perhaps come out of this discussion of the existence and usefulness, or not, of mental workload is a better, more multifactorial way of operationalizing this concept. Specifically, multiple

measures should be taken, performance, self-reports, and if possible physiology, and, very importantly, these do not need to correlate, else the assessment of one type would suffice. Dissociation of measures gives a better view on what has happened during performance of a task, what strategies were applied, and whether the operator had to try hard to protect performance.

References

Brookhuis KA, De Waard D, Fairclough SH (2003) Criteria for driver impairment. Ergonomics 46:433–445

De Waard D (1996) The measurement of drivers' mental workload. PhD thesis, University of Groningen. University of Groningen, Traffic Research Centre, Haren

De Winter JCF (2014) Controversy in human factors constructs and the explosive use of the NASA TLX: a measurement perspective. Cogn Technol Work (this issue). doi:10.1007/s10111-014-0275-1

Hockey GRJ (1997) Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. Biol Psychol 45:73–79

Hollnagel E (2009) The ETTO principle: efficiency-thoroughness trade-off. Why things that go right sometimes go wrong. Ashgate, Farnham

Rasmussen J (1983) Skills, rules, knowledge: signals, signs and symbols and other distinctions in human performance models. IEEE Trans Syst Man Cybern SMC-13:257–267

Reason J (1990) Human error. Cambridge University Press, Cambridge

Veltman JA, Gaillard AWK (1996) Measurement of pilot workload with subjective and physiological techniques. In: Brookhuis KA, Weikert CM, De Waard D (eds) Aging and human factors, proceedings of the europe chapter of the human factors and ergonomics society annual meeting in Soesterberg. University of Groningen, Traffic Research Centre, Haren, pp 107–128

Widyanti A, Johnson A, De Waard D (2013) Adaptation of the rating scale mental effort (RSME) for use in Indonesia. Int J Ind Ergon 43:70–76