The Workload Curve: Subjective Mental Workload

Steven Estes, The MITRE Corporation, McLean, Virginia

Objective: In this paper I begin looking for evidence of a subjective workload curve.

Background: Results from subjective mental workload assessments are often interpreted linearly. However, I hypothesized that ratings of subjective mental workload increase nonlinearly with unitary increases in working memory load.

Method: Two studies were conducted. In the first, the participant provided ratings of the mental difficulty of a series of digit span recall tasks. In the second study, participants provided ratings of mental difficulty associated with recall of visual patterns. The results of the second study were then examined using a mathematical model of working memory.

Results: An *S* curve, predicted a priori, was found in the results of both the digit span and visual pattern studies. A mathematical model showed a tight fit between workload ratings and levels of working memory activation.

Conclusion: This effort provides good initial evidence for the existence of a workload curve. The results support further study in applied settings and other facets of workload (e.g., temporal workload).

Application: Measures of subjective workload are used across a wide variety of domains and applications. These results bear on their interpretation, particularly as they relate to workload thresholds.

Keywords: mental workload, working memory, mathematical models

Address correspondence to Steven Estes, The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102; e-mail: sestes@mitre.org.

HUMAN FACTORS
Vol. 57, No. 7, November 2015, pp. 1174–1187
DOI: 10.1177/0018720815592752

INTRODUCTION

I find—and perhaps this is your experience as well—that an easily managed mental task can become, with just the slightest amount more mental demand, decidedly unmanageable. The relationship between unitary increases in cognitive load and the subjective experience of mental demand seem nonlinear; perceived mental workload is hardly affected at all by increases in demand under low cognitive load but rises quickly and disproportionally as the limits of the cognitive system are approached. In the context of subjective rating scales, it would appear that sometimes 5 is closer to 6 than to 4. That is, the cognitive load required to move a subjective workload rating from 5 to 6 is less than that required to move the rating from 4 to 5. This relationship, which is the basis of my central hypothesis, should take the shape of an *s* or sigmoid curve, as notionally depicted in Figure 1.

The hypothesized asymptote at the top of this *S* curve is the predictable result of using a finite scale (subjective mental workload) to evaluate a conceivably infinite quantity (mental load). Once workload is rated a 10 on a 1-to-10 scale, it does not matter whether task load proceeds to double or triple or quadruple. In each instance, subjective workload shares the basic quality of being "too much" and is therefore a 10. At the lower end of the scale, however, there is a finite beginning to the scale and the workload. There, the relationship between perceived and actual load could be linear, a power function, exponential, and so on. I propose that, as seen in Figure 1, where subjective ratings of workload are low or moderate, something resembling a power function will be observed. With very low subjective workload, unitary increases in cognitive load will result in modest increases in subjective ratings. As subjective workload increases, however, the rater will become more sensitive to his or her diminishing resources, and unitary increases in cognitive load will result in increasingly large jumps in subjective ratings. Throughout the

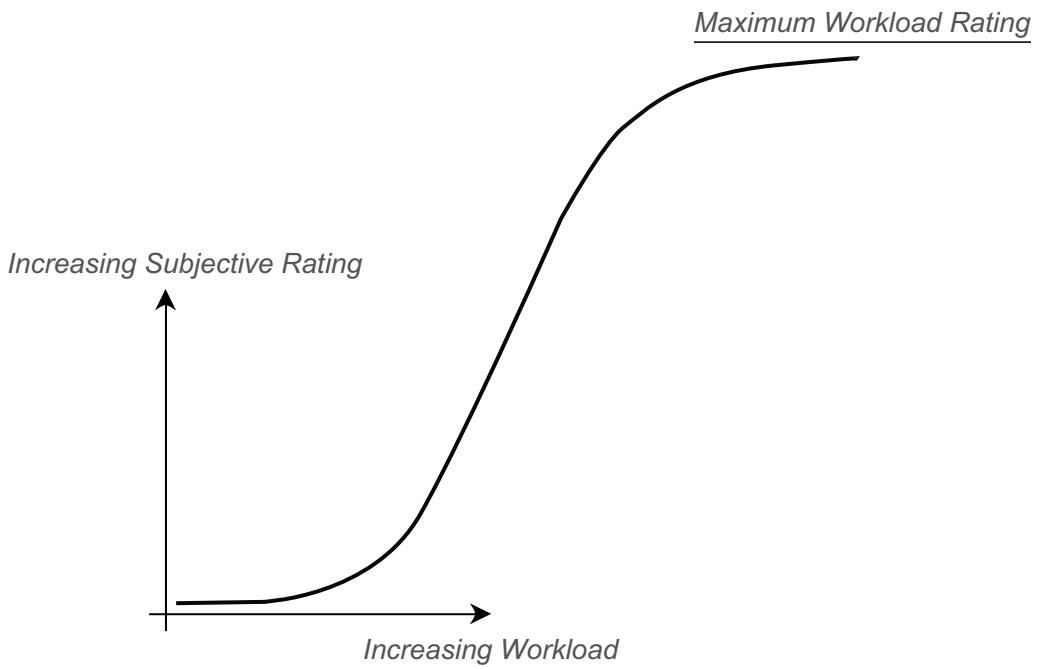

Figure 1. The subjective workload curve.

paper I will refer to my hypothesized relationship between cognitive load and subjective ratings of mental workload as the *workload curve*.

The hypothesized workload curve is unique within the literature. It is commonly assumed that equal intervals in ratings equate to equal intervals of imposed workload (Reid & Nygren, 1988; Young, Brookhuis, Wickens, & Hancock, 2015). If one is of a mind to find them, there are examples of curves in studies of mental workload (Berka et al., 2007; Eggemeier, Crabtree, & Reid, 1982), but they uniformly go uncommented upon and are never the topic of study. There is no mention of a curve in subjective ratings of workload in the seminal references for the most common subjective workload rating tools (Hart & Staveland, 1988; Wierwille & Casali, 1983). Hart and Staveland (1988) acknowledge the possibility of a proportionality between observed ratings and the magnitude of the rated phenomena, though they make no hypothesis as to how that might impact ratings or whether a curve may result. Yet, a curve in subjective ratings of mental workload would have a significant impact on the interpretation and perception of a subjective workload rating.

In this paper I review the evidence for a curve in the most commonly used subjective measures of mental workload and present the results from two studies. Those results are then evaluated in a mathematical model of working memory activation decay (subjective ratings of mental workload are strongly influenced by working memory). I begin with a brief review of mental workload and the impacts of working memory on subjective workload ratings.

Subjective Mental Workload

Workload, to oversimplify, is complex. It is multidimensional and its magnitude is the result of interactions between the human, the task, and the environment (Hart & Staveland, 1988; Simon, 1969; Wickens, 2008). Ultimately, documentation of the workload curve must take into account all of these variables. But I require a starting point, and the evaluation of the mental dimension of workload is a reasonable place to begin if for no other reason than it is difficult to quantify, and there is some appeal in dealing with the most difficult elements of a problem first.

A universally accepted definition of mental workload has been elusive. For this paper, *mental*

workload is defined in the strictest sense: the work done by the mental system. Somewhat less recursively, mental workload is the cognitive and perceptual processing expended in the course of completing a task (Eggemeier & Wilson, 1991), where processing includes the storage, maintenance, manipulation, and retrieval of information within working memory and long-term memory as accomplished through control of the locus of attention.

The measurement of workload has been a topic of interest in the applied community since at least the 1950s. By the late 1960s, workload had become an area of significant research, with a variety of techniques being developed to measure it. Wierwille and Williges (1978) classified these techniques into three categories: performance measures, psychophysiological measures, and subjective assessment. Those categories still accurately classify the vast majority of mental workload assessment techniques used today (Gawron, 2008).

Of the many dozens of methods within those categories proposed for measurement of workload—including dual-task tests, performance measures, heart rate, respiration, pupil dilation, functional magnetic resonance imaging, and infrared spectrometry—subjective workload measures have assuredly been the most widely used, which is likely attributable to their usability and face validity. As Moray et al. (1979) put it, "if the person feels loaded and effortful, he is loaded and effortful whatever the behavioral and performance measures may show" (p. 105).

Representative subjective workload measurement techniques, such as NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988), subjective workload assessment technique (SWAT; Reid & Nygren, 1988), and Cooper-Harper (Cooper & Harper, 1969), all produce a scalar rating of workload. Cooper-Harper's scale is ordinal, and NASA-TLX and SWAT are continuous. Several studies have shown strongly correlated ratings across these and other subjective workload measures (Hess, 1971; Rubio, Diaz, Martin, & Puente, 2004; Vidulich & Tsang, 1985). Many measures, like NASA-TLX and SWAT, are multidimensional and make allowances for distinguishing between different sources of workload, including mental workload.

Although there are numerous subjective techniques, including open-ended scales, there is a very limited set that sees consistent, applied use. The most popular, if the frequency of study is any indicator, is by far NASA-TLX. In her retrospective on its use, Hart (2006) found over 550 studies of NASA-TLX. To be clear, this number reflects not just studies that made use of NASA-TLX but 550 studies *of* NASA-TLX. Because of their overwhelming popularity, closed, bipolar rating scales for mental workload are of particular interest for this paper.

Consciousness, Working Memory, and Subjective Mental Workload

When mental workload is being measured subjectively, one may reasonably ask, "What is it that is being measured?" It does not seem, for example, that individuals sense the workload involved in visual perception; it is not effortful to see, although an incredible amount of neural processing is required. Instead, one's perception of workload is influenced almost solely by processes of which one has some conscious awareness (Vidulich, 1988; Yeh & Wickens, 1988).

In cognitive psychology, consciousness is thought to reside in working memory (Baddeley, 2007; Hassin, Bargh, Engell, & McCulluch, 2009). As the location of consciousness in the cognitive system, working memory has been attributed a central role in subjective ratings of mental workload (Gopher & Braune, 1984; Ericsson & Simon, 1980).

Yeh and Wickens (1988) found that the majority of variables found to affect subjective workload are related to working memory demands. Those variables include capacity (Hauser, Childress, & Hart, 1982), presentation rate (Daryanian, 1980), processing rate (Tulga & Sheridan, 1980), attention allocation, and decision alternatives.

Judgment of Mental Workload

The variables catalogued by Yeh and Wickens (1988) are a product of the capacity and durability limitations of working memory. One prevalent theory as to why those limitations exist is decay theory. According to decay theory, the strength of a memory, determined by its level of activation, fades over time (Baddeley, 1975;

Brown, 1958). Further, the pool of activation is limited and must be spread across all chunks in working memory (Just & Carpenter, 1992). In order to be recalled, a chunk's activation must exceed a threshold (Barrouillet, Bernardin, & Camos, 2004), and therefore the initial strength of the memory trace and resultant decay are critical to determining the probability of recall.

Although decay is a critical element of working memory, it seems unlikely that, in generating an estimate of mental workload, one directly measures decay of memory activation. More probable is the hypothesis that someone asked to rate his or her mental workload, lacking a direct measure of working memory activation, bases his or her rating on the *effects* of working memory activation and decay.

One could theorize many mechanisms by which activation influences judgments of mental demand. For instance, metamemory and learning research has documented the ability to estimate remaining working memory capacity and rates of forgetting as determined by activation (Amichetti, Stanley, White, & Wingfield, 2013; Bunnell, Baken, & Richards-Ward, 1999; Halamish, McGillivray, & Castel, 2011; Kornell, Rhodes, Castel, & Tauber, 2011). It may be that the accuracy of judgments about available capacity increases as available working memory capacity decreases and that this process in turn gives rise to the workload curve. Whatever the precise process, it is my contention that these judgments are based on the effects of working memory activation. Evidence for this hypothesis is discussed later in the modeling section of the paper. It is worth noting both that judgments in ratings of workload more have been discussed before in the literature (Hart & Staveland, 1988) and that they are thought to be relative to prior experience rather than absolute (Sheridan & Simpson, 1979).

In summary, I hypothesize that effects of working memory activation decay are critical to the assessment of mental workload:

- When subjective mental workload rated is plotted as a function of a measure of the imposed workload, the result is curvilinear.
- The relationship between subjective and imposed workload takes the shape of an *S* curve (the workload curve).
- The workload curve results from judgment of the effects of working memory activation decay.

STUDY 1

To test these hypotheses, I performed two web-based recall studies and one modeling exercise. Study 1 required serial recall of a digit span in order of presentation. After each trial, participants were asked to rate the mental demand of the recall task.

Participants

Study 1 included 102 participants. All participants were employees of the MITRE Corporation and participated voluntarily and anonymously. The study was deemed exempt by MITRE's institutional review board (IRB) under the provisions of 45 CFR 46. Participants were recruited via an internal newsletter. MITRE is a technical company, reflected in the demographic information provided by participants, 52% of whom described their job as some form of engineering. Other job descriptions included computer scientists (7%), managers (6%), information technology (IT) professionals (5%), and administrators (4%). No further demographic information was collected as additional details, like age and gender, would in some cases allow identification of the participant.

Procedure

Participants accessed an internal website to complete the study. On the welcome page, they were given a high-level description of the task. Participants then completed two practice problems representing the easiest and most difficult recall spans. After the practice problems, participants completed 24 digit span recall trials.

In each trial, the participant was presented with a digit span of varying length. Presentation time was determined by multiplying the span length by 500 ms. Once the presentation time elapsed, the span was removed and the participant was provided an open text field for entering the digit span as he or she recalled it. There was no time limit on recall.

After entering each span, the participant was asked to rate the mental difficulty of the task on a scale of 1 to 10. Not unlike NASA-TLX,

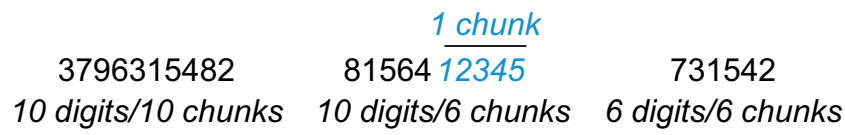

Figure 2. Chunks versus digit spans.

mental difficulty was described as a rating of the mental effort required to complete the task. Decimal ratings were allowed. The scale was bipolar with adjectival labels of *low* and *high* at each end. Once all trials were completed, participants were asked to provide some information on recall strategies used during the study.

Manipulations

A repeated-measures design, the study was block ordered and included a "no-chunks" and a "chunks" condition, each with 12 levels of digit span to be recalled (resulting in a total of 24 trials). In the no-chunks condition, each digit span level corresponded with the number of digits actually presented to the participant (e.g., at Level 10, the participant was presented with 10 digits to recall). A random-number generator was used to create the digit spans. Each span was then inspected manually to ensure that no obvious patterns were included. Commonly chunked digits, such as local area codes or ZIP codes, were replaced. When possible (Levels 1 to 9), an integer was used only once. Integers appeared, at most, twice in the span. Although care was used to ensure no obvious chunks were included, this precaution did not preclude the participants from developing a chunking strategy based on nonobvious, personal information.

The second condition is referred to as the chunks condition. As with the no-chunks condition, participants were presented a digit span, recalled that span, and provided a rating of mental difficulty. However, as part of the chunks condition, one digit in the string was replaced with a five-digit chunk based on an obvious pattern (either 12345 or 54321) such that a condition span of six would contain five digits followed by a five-digit chunk. Participants were told to expect these patterns in some spans. To ensure participants recognized them as chunks, those digits, otherwise black, were colored blue. I have hypothesized that working memory decay, and therefore chunk activation, plays a central role in ratings of mental workload. This condition is therefore included to verify that workload ratings were based on the number of chunks in the span rather than the number of digits or apparent length of the span. For my hypothesis to be supported, a 10-digit number composed of six chunks (as in the center of Figure 2) should be rated more like the six-digit number containing no chunks on the right of Figure 3 than the 10-digit number containing no chunks on the left.

The spans used for all 20 trials are provided in Table 1.

Results

The primary concern was the shape of subjective mental workload: Will an *S* curve be produced when plotting workload ratings as a function of the number of chunks in the span? Qualitatively, as seen in Figure 3, the answer is yes. The shape was confirmed quantitatively when the data was fit to a four-parameter logistic (4PL) model (Baud, 1993), which produced an R^2 value of .99 (root mean square error [RMSE] = 0.16).

On the secondary question of equivalence, I found a largely favorable result. A repeated-measures ANOVA showed, as would be expected, no statistically significant difference between the two chunk conditions, $F(1, 101) = 2.9$, $p = .09$. Likewise, eta squared showed chunk condition accounted for none of the total variance ($\eta^2 = .00$). Testing for equivalence using inferential confidence intervals (Tryon, 2001) with a criterion of a 0.75 scale degree, seven of the 12 span lengths were shown to be statistically equivalent ($p < .05$) when compared across chunk conditions (Table 2).

Discussion

Study 1 showed that, indeed, participants' judgment of mental workload took the shape of

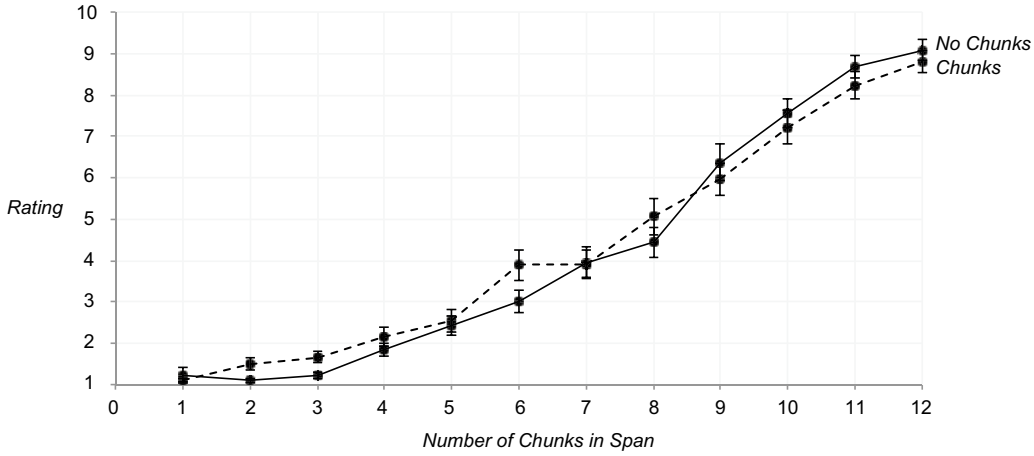

Figure 3. Ratings as a function of chunks in digit span with confidence intervals.

TABLE 1: Study 1 Digit Spans

Length	No Chunks	Chunks
1	5	12345
2	73	6 54321
3	185	94 12345
4	4967	296 54321
5	43795	3172 12345
6	154859	21531 54321
7	5439156	981564 12345
8	15362084	2451386 54321
9	945132708	73016829 12345
10	1594862730	953472680 54321
11	38170649527	5203147968 12345
12	715039428603	16428530920 54321

an *S* curve, as nonlinear increases in ratings of workload occurred with unitary increases in the number of chunks in the span. And with regard to the number of chunks in the span, the data, in both practical interpretation and statistical testing, favor the conclusion that mental difficulty was judged by the number of chunks rather than the number of digits in the span. This latter point will be important in the modeling section of the paper, but for now, I turn to a second test of the hypothesized workload curve. In this second study, a different type of demand will be placed on working memory to see if the sigmoid shape remains.

STUDY 2

The digit span study relied on working memory in the central executive and articulatory rehearsal loop (Kahana, 2012). Capacity and durability in the rehearsal loop, however, differs from that of the visuospatial sketchpad (Card, Moran, & Newell, 1983). To determine if the *S* curve persists when processing load is placed on the visuospatial sketchpad subsystem, a visual pattern was used for the second study. For this study, I hypothesized that the *S*-shaped workload curve seen in the digit span would be retained though steeper due to the higher demands placed on working memory.

TABLE 2: Inferential Confidence Intervals (ICIs)

Span	1	2	3	4	5	6	7	8	9	10	11	12
ICI	0.33	0.55	0.59	0.61	0.48	1.35	0.56	1.19	1.01	0.90	0.89	0.65

Participants

Thirty volunteers participated in Study 2. As with the first study, all participants were employees of the MITRE Corporation and completed the study voluntarily. Limited demographic information was collected to ensure anonymity. The study was deemed exempt by MITRE's IRB under the provisions of 45 CFR 46. Participants were recruited via an internal newsletter. Fifty-four percent of participants described their job as engineering. Other job descriptions included IT professionals (6%), computer scientists (5%), managers (3%), and administrators (3%).

Procedure

Participants accessed an internal website to complete the study. The welcome page provided a high-level description of the task. This description was immediately followed by training for the first of three sets of trials. The training included instructions on how the trials should be completed, example problems with solutions, and three practice problems. Each participant completed all levels of every condition. Conditions were presented in blocks and block order was varied. Prior to each block, participants were given instructions on how to proceed and then completed three practice trials. The three practice trials represented the easiest, moderate, and most difficult levels of the condition.

The visual pattern was shown to the participant using a wheel consisting of six colored buttons (example in Figure 4). In each trial, buttons were highlighted for 1 s in a predetermined pattern. The pattern was repeated back by the participant by pressing the buttons on the color wheel. The entire pattern was shown only once. Response time was unlimited and after each trial, the participant was asked to rate the mental difficulty of the task with the same bipolar scale used in the first study. In addition to collecting the scale rating, correctness of the response and the total response time were calculated. Upon completion of the study, participants completed a brief exit survey. This procedure was modeled after the electronic game *Simon*, variants of which have been widely used to test visual memory span (Cleary, Pisoni, & Geers, 2001; Gendle & Ransom, 2006; Humes & Floyd, 2005).

Figure 4. Visual pattern study button wheel.

Manipulations

Instructions for a set of trials and the length of the visual span were manipulated. Trial instruction conditions included "as seen," "reverse," and "offset." In the as-seen condition, the participant repeated the pattern back in the order it was presented. The reverse condition required the participant to repeat pattern back in the opposite order, beginning by pressing the button highlighted last, first. This condition necessitated manipulation of information in working memory but not the creation of new chunks for storage. The offset condition required both manipulation and the creation of new working memory chunks. In this condition, the participant repeated the pattern back in the order presented but with a one-position-clockwise offset as illustrated in Figure 5. Span length varied from one item in the visual pattern to 10 items.

Results

As expected, the workload curves found in Study 2 were even more pronounced than those

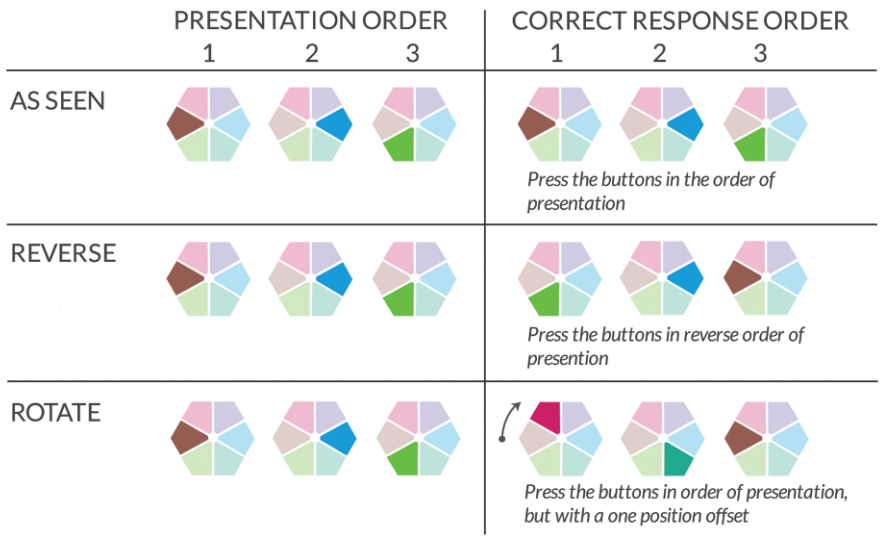

Figure 5. Visual pattern study instructions by condition.

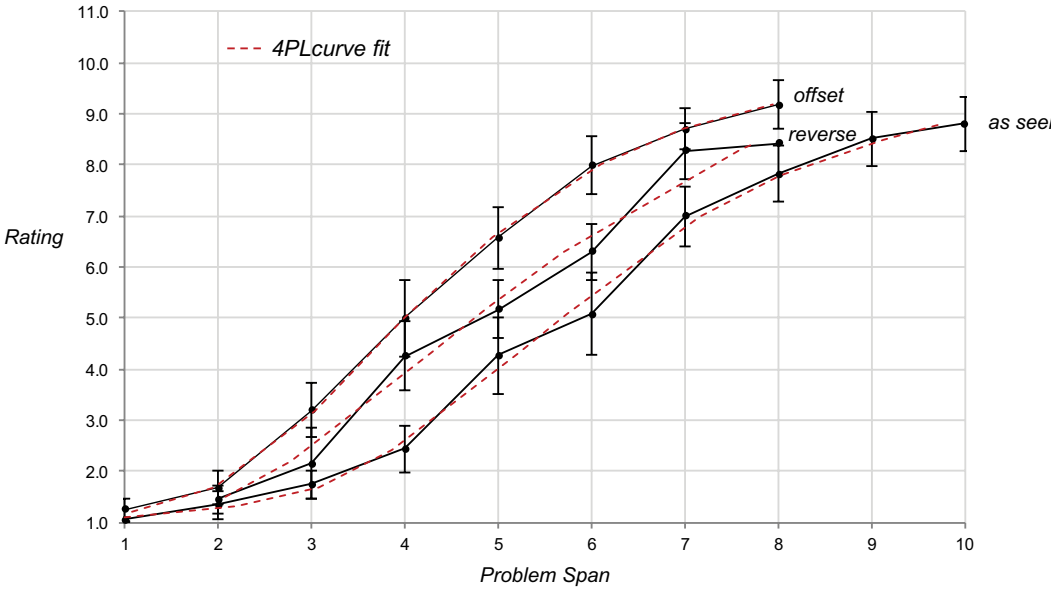

Figure 6. Visual span subjective ratings as function of span length with confidence intervals.

seen in the digit span (Figure 6). Using the as-seen curve as an example, ratings at the bottom of the curve increased less than half a point between spans, jumped to nearly two full points between spans in the middle of the curve, and then subsided back to increases of half a point or less near the top. Although the increases happened more quickly in the other conditions, the basic pattern remained. Using 4PL functions, the as-seen (R^2 = .99, RMSE = 0.24), reverse (R^2 = .98, RMSE = 0.24), and offset (R^2 = .99, RMSE = 0.08) showed strong fits as sigmoid curves. When sigmoid model curves were applied to individual data, the RMSEs increased across the board but were, nonetheless, within one scale degree of the observed value (as seen, RMSE = 0.75; reverse, RMSE = 0.49; offset, RMSE = 0.98).

The ordering of the results was also as expected, with the as-seen condition rated the easiest and offset the most difficult. A repeated-measures ANOVA showed those differences to be significant, $F(2, 28) = 63.7$, $p < .05$. There was a significant main effect for span, $F(6, 24) = 181.6$, $p < .05$, as well. According to Cohen's (1988) guidelines for interpretation of the eta-squared test, the effect sizes for the instruction condition and span were both large ($\eta^2 = .34$ and $\eta^2 = .80$, respectively).

Discussion

Once again, distinct *S* curves are seen. With small visual spans, unitary increases in workload correspond to modest increases in workload ratings. These modest increases are followed by noticeable jumps in ratings as the participant becomes more sensitive to diminishing working memory resources and asymptotes as those resources become overwhelmed. Although these results need to be tested further in ecologically valid environments, they provide a strong initial argument for the workload curve and rethinking how one interprets subjective workload ratings collected via popular methods, like NASA-TLX.

The results also uncover an unfortunate, if logical, interaction between people's limited working memory capacity—estimates range from seven chunks (Miller, 1956) to as low as three or four chunks (Cowan, 2001)—and perceived mental workload. In most situations, larger jumps in subjective ratings with unitary increases in load would be preferred—particularly if boredom is of concern—when working memory demand is low. Smaller jumps would be preferred when working memory is moderate and people begin to approach the limits of their capacity. However, the workload curve shows the opposite to be true. Given that the practitioner is often trying to manage user workload, knowledge of the curve is valuable as it gives guidance on the magnitude of mental demands necessary to increase a user's engagement and emphasizes how carefully workload must be managed passing the midpoint of the subjective rating scale.

A MODEL OF THE SUBJECTIVE MENTAL WORKLOAD CURVE

In the section on judgment, I hypothesized that the subjective workload curve arises from judgments about the availability of information in working memory, which is not to say that other parts of the cognitive system do not impact subjective ratings of mental workload. Rather, I am hypothesizing that availability plays the *central* role in these judgments and that knowledge of availability alone is sufficient to replicate the workload curve. If this hypothesis is true, then a model of working memory availability (i.e., activation decay) should predict ratings of mental difficulty in conditions when pure recall plays a lesser role (the offset condition of the visual span study) just as accurately as it does ratings in near-pure recall conditions (the as-seen condition).

Activation-Based Model

In his famous paper, "You Can't Play 20 Questions With Nature and Win," Allen Newell (1973) argued that experimental psychology, although adept at answering binary questions about psychological phenomena, was not advancing cognitive psychology toward a unified understanding of the mind. Newell believed that a unified theory required the development of cognitive architectures: software that implemented human cognitive capacities and constraints such that they could be used to test a theory's plausibility within the broader cognitive system.

One of the many cognitive architectures that arose as a response to Newell's challenge is adaptive control of thought–rational (ACT-R; Anderson, 2007). Relevant to the work at hand, research on memory using the ACT-R software and formulations has been extensive, with hundreds of published papers on the topics of memory activation, decay, or interference (ACT-R-related research is archived at http://act-r.psy.cmu.edu/). For this particular paper, my objective was to provide plausible support for the hypothesis that the workload curve demonstrated in the first two studies arises from judgments based on the effects of working memory activation and decay. As such, I relied on memory activation and decay functions that have been well established within the ACT-R community (Altmann & Schunn, 2002; Altmann & Trafton, 2002; Anderson, Reder, & Lebiere, 1996; Böhm & Mehlhorn, 2009; Pape & Urbas, 2008; Sohn, Anderson, Reder, & Goode, 2004).

First, the activation of a memory trace was calculated using the following equation taken from the literature (Altmann & Trafton, 2002):

$$a = ln\ (\frac{n}{\sqrt{T}}), \quad (1)$$

where a is activation, n is the number of times that chunk is rehearsed, and T is the total time the trace is held in memory (and the determinate of decay). Next, in order to mimic the division of activation across all working memory chunks, I reduced the activation as a function of the number of chunks in the problem span:

$$d = a + \frac{1}{c} - 1, \quad (2)$$

where d is the divided activation, a is activation, and c is the number of chunks the activation must be divided among. The idea of limited activation source pools and their distribution among all the chunks held in working memory has been previously documented in the literature (Anderson et al., 1996).

These two basic equations allowed me to model a relationship between the number of chunks to be memorized (visual pattern span), decay over time, and a subjective rating of mental demand. The first equation requires a reasonable estimate of the number of times each chunk is rehearsed between storage and final recall. In order for the model to have explanatory power, I should set free parameters, like rehearsal, once and use those settings for modeling the results of all three visual pattern conditions. The model could not be applied to the digit span study as response time—required to set T—was not collected. For this model, rehearsal (n) was set to three retrievals of the chunk (a plausible level of rehearsal that provided the best overall fit).

With the parameters set, an activation value was calculated for each participant response using the condition to determine number of chunks and response time for the decay period. The data were then collapsed across all three visual pattern conditions and an average activation generated for each degree of the subjective rating scale. It was necessary to round each raw subjective rating to the nearest whole number in order to ensure a sufficient number of observations at each point in the rating scale. A logarithmic curve, seen in Figure 7, was then fit to the data from Study 2 (the model curve). This step allowed evaluation of the fit of averaged data for each condition to the model curve. The best trend and magnitude of fit (Schunn & Wallach, 2005) were found for the as-seen condition (R^2 = .99, RMSE = 0.27). This finding makes intuitive sense as this condition was closest to what may be called a pure working memory test (i.e., place information in working memory and repeat it back verbatim).

The reverse condition equaled the trend of the as-seen condition but showed a slightly larger error value (R^2 = .99, RMSE = 0.35). The values for the offset condition, although lower still, were not noticeably worse (R^2 = .97, RMSE = 0.47), and the estimated subjective workload ratings were, on average, within half a point of the observed workload ratings. So, although increasing recruitment of differing cognitive resources does seem to impact the accuracy of the model, in these tests, working memory activation levels alone were able to provide a very good prediction of the observed subjective ratings of mental difficulty and, more to the point at hand, account for the subjective workload curve.

CONCLUSIONS

With hindsight, it makes sense that when it comes to mental workload, 5 is sometimes closer to 6 than to 4. In psychology, many, if not most, well-established effects exhibit curvilinear relationships. Fitts's law (Fitts, 1954), Hick's law (Hick, 1952), the Yerkes-Dodson law (Yerkes & Dodson, 1908), Stevens's power law (Stevens, 1957), subitization (Jevons, 1871), and the power law of practice (Newell & Rosenbloom, 1981) are all curvilinear processes, as is activation decay (Anderson, 2007; Byrne & Bovair, 1997; Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012) and the probability of memory recall over time (Bachelder, 2000; Mueller & Krawitz, 2009; Taatgen, 2000). Those curves exist for a variety of reasons. Cognitive processing, for example, is thought to happen along a curve for expediency; it is unnecessary and inefficient to process sensory input linearly (Burns, 2014). In the case of workload, the modeling exercise conducted in this paper supported the idea that the mental workload curve may bend

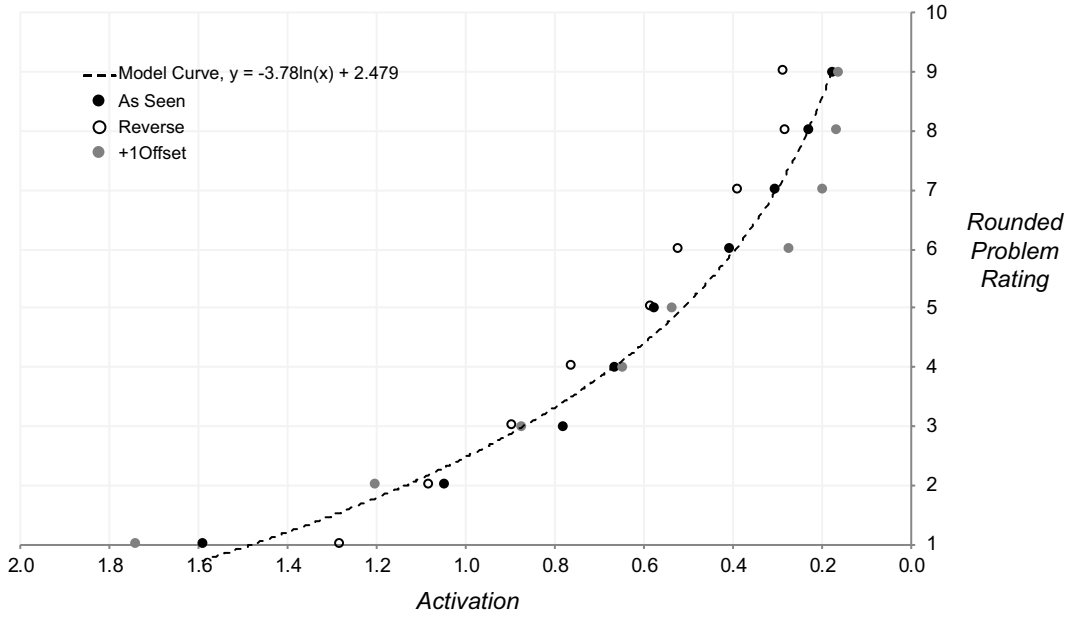

Figure 7. Model fit for problem rating as function of activation by condition.

as a result of the subject's reflecting on the difficulty he or she has maintaining information in working memory (i.e., availability).

Although perhaps intuitive, the existence of a subjective workload curve has, until now, never been formally documented. And although this research needs to be replicated with ecologically valid tasks, it acts as a starting point for refactoring the interpretation of subjective ratings of workload. The activation model gives credence to the theory proposed here that the judgments that give rise to the workload curve are based on the effects of activation and decay. This theory suggests that practitioners who wish to manage the mental workload imposed by a tool must manage not only the total amount of information the user must hold in working memory but how long it must be held there.

When subjective measures are used to compare workload under differing conditions, the workload curve indicates that the true magnitude of the difference is dependent on where on the scale the ratings lie. The position of the rating on the scale may likewise tell something about the stability of the rating, as ratings in both studies became more sensitive under increasing loads.

With regard to the impact of the workload curve, consider how widespread the use of subjective workload ratings are in safety-critical domains, like aviation. The *Journal of Aviation Psychology*, for example, publishes applied research related to aviation safety. Using its online search capabilities, I found that of 566 articles, approximately 25% included measures of subjective workload. One in 10 included NASA-TLX specifically. For these practitioners who so commonly use subjective workload ratings to ensure system safety, a more thorough understanding of subjective workload ratings is always of value.

Moving forward, several questions need to be answered. First, can the workload curve be found in more complex mental work? Ecologically valid tests will be required to answer this question, but they will be difficult as they require formulating and modeling working memory processing in complex environments. To that end, the formulations found in the modeling section here will be integrated into the Cogulator cognitive modeling tool (http://cogulator.io).

Second, does the curve seen in subjective ratings of mental workload exist in other dimensions of workload? Some evidence of a temporal workload curve can be found in existing studies (e.g., Dijksterhuis, de Waard, Brookhuis, Mulder, & de Jong, 2013), and temporal workload will

likely be the next dimension of workload I investigate. Until those studies take place, I hope that others find the workload curve useful in their applied work and add to it with their own research.

ACKNOWLEDGMENTS

This work was produced for the U.S. government under Contract DTFA01-01-C-00001 and is subject to Federal Aviation Administration (FAA) Acquisition Management System Clause 3.5-13, Rights in Data–General, Alt. III and Alt. IV (October 1996). The contents of this document reflect the views of the author and the MITRE Corporation and do not necessarily reflect the views of the FAA or the DOT. Neither the FAA nor the Department of Transportation makes any warranty or guarantee, expressed or implied, concerning the content or accuracy of these views. My thanks to Ronald Chong for his help in reviewing the experimental design, data, and document. Thanks to Christopher DeSenti for brainstorming the concept and reviewing the document. Thanks to John Helleberg for his thoughts on the concept, data, and design. Finally, my thanks to Valerie Gawron and Kevin Burns for their feedback on the results.

KEY POINTS

- Evidence was found for the existence of a workload curve.
- *S* curves characterize the relationship between working memory load and subjective ratings of workload.
- I hypothesize subjective mental workload is driven by the availability of working memory traces (activation), and models support that hypothesis as a plausible theory.

REFERENCES

Altmann, E. M., & Schunn, C. D. (2002). Integrating decay and interference: A new look at an old interaction. In W. Gray & S. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 65–70). Fairfax, VA: Cognitive Science Society.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*, 39–83.

Amichetti, N., Stanley, R., White, A., & Wingfield, A. (2013). Monitoring the capacity of working memory: Executive control and effects of listening effort. *Memory & Cognition*, *41*, 839–849.

Anderson, J. (2007). *How can the mind occur in the physical universe?* New York, NY: Oxford University Press.

Anderson, J., Reder, L., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*, 221–256.

Bachelder, B. (2000). The magical number 4 = 7: Span theory on capacity limitations. *Behavioral and Brain Sciences*, *24*, 87–185.

Baddeley, A. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589.

Baddeley, A. (2007). *Working memory, thought, and action*. New York, NY: Oxford University Press.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*, 83–100.

Baud, M. (1993). Data analysis, mathematical modeling. In R. F. Masseyeff, W. Albert, & N. A. Staines (Eds.), *Methods of immunological analysis: Vol. 1. Fundamentals* (pp. 656–671). New York, NY: VCH.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*, 231–244.

Böhm, U., & Mehlhorn, K. (2009). The influence of spreading activation on memory retrieval in sequential diagnostic reasoning. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 188–193). Manchester, UK.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*, 12–21.

Bunnell, J., Baken, D., & Richards-Ward, L. (1999). The effect of age on metamemory for working memory. *New Zealand Journal of Psychology*, *28*, 23–29.

Burns, K. (2014). *Entropy and optimality in abstract art: An empirical test of visual aesthetics*. Retrieved from http://www.askhow.org/papers/Entropy%20and%20Optimality%20Pre-Print.pdf

Byrne, M., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science*, *21*, 31–61.

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human computer interaction*. Hillsdale, NJ: Lawrence Erlbaum.

Cleary, M., Pisoni, D. B., & Geers, A. E. (2001). Some measures of verbal and spatial working memory in eight- and nine-year-old hearing-impaired children with cochlear implants. *Ear & Hearing*, *22*, 395–411.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cooper, G., & Harper, R. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Tech. Rep. TN D-5153). Washington, DC: NASA.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114.

Daryanian, B. (1980). *Subjective scaling of mental workload multitask environment* (Unpublished master's thesis). Massachusetts Institute of Technology, Cambridge.

Dijksterhuis, C., de Waard, D., Brookhuis, K., Mulder, B., & de Jong, R. (2013). Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. *Frontiers in Neuroscience*, *7*(149). Retrieved from http://journal.frontiersin.org/article/10.3389/fnins.2013.00149/full

Eggemeier, F. T., Crabtree, M. S., & Reid, G. B. (1982). Subjective workload assessment in a memory update task. In *Proceedings of the Human Factors and Ergonomics Society 26th Annual Meeting* (pp. 643–647). Santa Monica, CA: Human Factors and Ergonomics Society.

Eggemeier, F. T., & Wilson, G. F. (1991). Performance-based and subjective assessment of workload in multi-task environments. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 217–278). London, UK: Taylor & Francis.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

Fitts, P. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*, 381–391.

Gawron, V. J. (2008). *Human performance, workload, and situational awareness measures handbook*. Boca Raton, Florida: CRC.

Gendle, M., & Ransom, M. (2006). Use of the electronic game Simon as a measure of working memory span in college age adults. *Journal of Behavioral and Neuroscience Research*, *4*, 1–7.

Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*, *26*, 519–532.

Halamish, V., McGillivray, S., & Castel, A. (2011). Monitoring one's own forgetting in younger and older adults. *Psychology and Aging*, *26*, 631–635.

Hart, S. G. (2006). NASA–Task Load Index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society 26th Annual Meeting* (pp. 904–908). Santa Monica, CA: Human Factors and Ergonomics Society.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, Netherlands: North-Holland.

Hassin, R. R., Bargh, J. A., Engell, A., & McCulluch, K. C. (2009). Implicit working memory. *Consciousness and Cognition*, *189*, 665–678.

Hauser, J. R., Childress, M. E., & Hart, S. G. (1982). Rating consistency and component salience in subjective workload estimation. In F. George (Ed.), *Proceedings of the 18th Annual Conference on Manual Control* (pp. 127–149). Washington, DC: NASA.

Hess, R. (1971). *The use of a nonadjectival, nonordinal, linear rating scale in a single axis compensatory tracking task* (Unpublished master's thesis). Naval Postgraduate School, Monterey, CA

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11–26.

Humes, L., & Floyd, S. (2005). Measures of working memory, sequence learning, and speech recognition in the elderly. *Journal of Speech, Language, and Hearing Research*, *48*, 224–235.

Jevons, W. S. (1871). The power of numerical discrimination. *Nature*, *3*, 281–282.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kahana, M. (2012). *Foundations of human memory*. New York, NY: Oxford University Press.

Kornell, N., Rhodes, M., Castel, A., & Tauber, S. (2011). The ease-of-processing heuristic and stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*, 787–794.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Moray, N., Johansen, J., Pew, R. W., Rasmussen, J., Sanders, A. F., & Wickens, C. D. (1979). *Mental workload: Its theory and measurement*. New York, NY: Plenum Press.

Mueller, S., & Krawitz, A. (2009). Reconsidering the two-second decay hypothesis in verbal working memory. *Journal of Mathematical Psychology*, *53*, 14–25.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York, NY: Academic Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*, 779–819.

Pape, N., & Urbas, L. (2008). A model of time-estimation considering working memory demands. In *Proceedings of the 30th Annual Cognitive Science Society* (pp. 1543–1548). Austin, TX: Cognitive Science Society.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185–218). Amsterdam, Netherlands: Elsevier.

Rubio, S., Diaz, E., Martin, J., & Puente, J. M. (2004) Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and Workload Profile. *Applied Psychology: An International Review*, *53*, 61–86.

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115–154). Saarbrueken, Germany: University of Saarland Press.

Sheridan, T. B., & Simpson, R. W. (1979). *Toward the definition and measurement of the mental workload of transport pilots* (Tech. Rep. No. R 79-4). Cambridge, MA: MIT Flight Transportation Laboratory.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Sohn, M. H., Anderson, J. R., Reder, L. M., & Goode, A. (2004) Differential fan effect and attentional focus. *Psychonomic Bulletin and Review*, *11*, 729–734.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Taatgen, N. (2000). Dispelling the magic: Towards memory without capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Tryon, W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

Tulga, M. K., & Sheridan, T. B. (1980). Dynamics decision and workload in multitask supervisory control. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-10*, 217–231.

Vidulich, M. A. (1988). The cognitive psychology of subjective mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 219–229). New York, NY: Elsevier.

Vidulich, M. A., & Tsang, P. (1985). Assessing subjective workload assessment: A comparison of SWAT and the NASA-bipolar methods. In *Proceedings of the Human Factors and Ergonomics Society 29th Annual Meeting* (pp. 71–75). Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*, 449–455.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors and Ergonomics Society 27th Annual Meeting* (pp. 129–133). Santa Monica, CA: Human Factors and Ergonomics Society.

Wierwille, W. W., & Williges, R. C. (1978). *Survey and analysis of operator workload assessment techniques* (Tech. Rep. No. S-78-101). Systemetrics, Inc.

Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*, 111–120.

Yerkes, R., & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Steven Estes lives in Savannah, Georgia. He holds a BA in history (1996) from the University of Georgia in Athens and an MA in human factors and applied cognition (2002) from George Mason University in Fairfax, Virginia. He is currently a principal human factors engineer at the MITRE Corporation's Center for Advanced Aviation System Design in McLean, Virginia. Prior to working for MITRE, he was employed as a human factors engineer at Gulfstream Aerospace. Publications include the book chapter "Macrocognition in Systems Engineering: Supporting Changes in the Air Traffic Control Tower," published in the book *Naturalistic Decision Making and Macrocognition* (Burlington, VT: Ashgate, 2008). He is the developer of Cogulator, an applied tool for workload assessment and task time estimation (http://cogulator.io). Research interests include cognitive engineering and human computer interface design.

Date received: October 14, 2014
Date accepted: May 27, 2015