

Target Paper

Subjective rating scales: science or art?

JOHN ANNETT*

Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

Keywords: Subjective rating scales; Intersubjectivity; Measurement; Multivariate analysis; Annoyance; Comfort; Effort; Fatigue; Presence; Stress, Urgency; Usability; Workload.

Subjective rating scales are widely used in almost every aspect of ergonomics research and practice for the assessment of workload, fatigue, usability, annoyance and comfort, and lesser known qualities such as urgency and presence, but are they truly scientific? This paper raises some of the key issues as a basis for debate. First, it is argued that all empirical observations, including those conventionally labelled as 'objective', are unavoidably subjective. Shared meaning between observers, or intersubjectivity, is the key criterion of scientific probity. The practical steps that can be taken to increase intersubjective agreement are discussed and the well-known sources of error and bias in human judgement reviewed. The role of conscious experience as a mechanism for appraising the environment and guiding behaviour has important implications for the interpretation of subjective reports. The view that psychometric measures do not conform to the requirements of truly 'scientific' measurement is discussed. Human judgement of subjective attributes is essentially ordinal and, unlike physical measures, can be matched to interval scales only with difficulty, but ordinal measures can be used successfully both to develop and test substantive theories using multivariate statistical techniques. Constructs such as fatigue are best understood as latent or inferred variables defined by a set of manifest or directly observed indicator variables. Both construct validity and predictive validity are viewed from this perspective and this helps to clarify several problems including the dissociation between measures of different aspects of a given construct, the question of whether physical (e.g. physiological) measures should be preferred to subjective measures and whether a single measure of constructs which are essentially multidimensional having both subjective and physical components is desirable. Finally, the fitness of subjective ratings to different purposes within the broad field of ergonomics research is discussed. For testing of competing hypotheses concerning the mechanisms underlying human performance, precise quantitative predictions are rarely needed. The same is frequently true of comparative evaluation of competing designs. In setting design standards, however, something approaching the level of measurement required for precise quantitative prediction is required, but this is difficult to achieve in practice. Although it may be possible to establish standards within restricted contexts, general standards for broadly conceived constructs such as workload are impractical owing to the requirement for representative sampling of tasks, work environments and personnel.

*e-mail: j.annett@warwick.ac.uk

1. Introduction

The founding fathers of the science of ergonomics cautioned against the use of subjective methods. Morgan *et al.* (1963: 48), for example, warned that 'objective measurements in which performances are recorded in quantitative terms are strongly preferred to subjective opinion, comments and ratings'. Yet rating scales are now in common use and feature in the majority of published research reports (six out of seven in the issue of this journal current at the time of writing).

Subjective workload is one of the most debated measures in ergonomics and a number of subjective rating scales have been developed. The Cooper–Harper Scale (Cooper and Harper 1969) was originally developed as a checklist to be used by pilots to assess the handling characteristics of aircraft, whereas the more recent methods such as Subjective Workload Assessment Technique (SWAT; Reid and Nygren 1988) and NASA-TLX (Task Load Index; Hart and Staveland 1988) focus principally on the subjective response of the user to the demands of the task. Although ostensibly referring to the inner experience of the operator, these ratings are generally taken to reflect the nature of the task and its demands on human physical and mental resources. *Effort* and *fatigue* as measured by such as scales as Perceived Exertion (the RPE and CR10 scales) developed by Borg (1998) and the Swedish Occupational Fatigue Inventory (SOFI; Åhsberg 1998) are clearly subjective. *Usability* assessment typically involves an audit of various objective features of the system using checklists and report forms (Hulzebosch and Jameson 1996) but may also include subjective ratings such as the Software Usability Measurement Inventory (SUMI; Porteus *et al.* 1993) or Brooke's (1996) System Usability Scale (SUS). Klein-Teeselink *et al.* (1999) whilst describing methods of collecting objective performance data as a means of assessing interface usability also recommended the use of questionnaires and interviews to assess the enjoyment and satisfaction associated with their use. The quality of *presence* or *being there* is a recent development (Ijsselsteijn *et al.* 1999) relevant to the design of simulations and virtual environments. It is an essentially subjective phenomenon that therefore requires subjective measurement. Scales for the measurement of *comfort*, or *annoyance*, have also been widely used in ergonomics especially in the field of seat design (Shackel *et al.* 1969) and vibration (Griffin *et al.* 1982) and various environmental and climatic factors (Rohles and Konz 1987). The perceived *urgency* of auditory warning signals (Hellier *et al.* 1995) is another case in which subjective response to a physical input might be used to evaluate and optimize equipment design.

Although objective measures of operator performance are generally preferred in evaluation studies these are often supplemented by subjective ratings of experts, such as instructors. This use of subjective judgement is familiar to all teachers and academics and it has particular relevance to ergonomics when complex performance characteristics are being assessed. Recent examples include the assessment of team performance in terms of characteristics such as *situation awareness* (Endsley 1996) or *communication and coordination* (Smith-Jentsch *et al.* 1998).

This paper does not attempt a detailed review of the many subjective rating scales currently in use, nor even to provide a detailed critique of the specific measurement techniques employed, but rather will address some of the fundamental problems concerning the use of subjective evidence in the practice of ergonomics. First is the problem of subjectivity *per se*. Are subjective methods fundamentally flawed and outside the scope of true scientific investigation or are they acceptable, if less

compelling, measurement tools? Related to this first issue are questions about the nature of measurement and whether subjective data comply with the logical requirements of true scientific measurement. Given these problems can be resolved, there are questions concerning the place of subjective scales within the context of ergonomics research and practice. How are theoretical constructs such as *workload* and *usability* validated, how useful are they in design and evaluation compared with objective measures and how does one interpret the dissociations often found between subjective and objective measures of what is ostensibly the same construct?

2. Subjectivity

2.1. *Subjectivity and introspection*

Weber (1830, trans Ross 1972: 91) distinguished between *objective* and *subjective perceptions*:

Objective perceptions arise when we not only sense the change in our own organs caused by the stimulus of the perceived objects but also seem to sense the object itself . . . other parts of the human body admit only subjective perceptions: their nature is such that we perceive only changes in our organs coming into contact with other objects.

The distinction is between judgements of objects and events in the external world, that is potentially public information, and judgements of internal sensations and feelings obtained by introspection which are essentially private.

The subjective scales used in ergonomics reflect both objective and subjective judgements in Weber's sense. For example, *usability* and *urgency*, would probably have been classed by Weber as objective since they refer to the qualities of objects and events external to the observer, whilst others such as *fatigue* and *effort*, referring primarily to the observer's introspective experience, would be classed as subjective. *Workload* may lie somewhere in between the two being in part a measure of the observer's private sensations but in part an assessment of the task. Ratings of felt *pain* on the one hand and *usability* on the other are equally subjective in the current use of that term but would have been classed by Weber as respectively subjective and objective. The difference is principally one of *attribution*. Pain is typically attributed to an internal cause, such as a rotten tooth, whilst usability is more likely to be attributed to the physical characteristics of the equipment. The important difference between subjective qualities such as fatigue, anxiety, and pain or workload, usability and presence is that, in some the external stimulus conditions to which the experience is attributed can be more readily reproduced, and hence there is greater potential for independent observers to agree or disagree.

The classical objective/subjective dichotomy may be too absolute. Rather we should consider human judgements as varying in terms of the degree of *intersubjectivity*, that is the degree of shared meaning between independent observers of the same object or event. Two people looking at the same instrument under the same circumstances will normally agree on the reading, which can then be taken as objective 'fact', but perfect agreement is the exception rather than the rule. In practice, two observers may make their measurements with slightly different instruments, or under slightly different conditions or just at different times. This is normal in science and the process of verification and falsification therefore depends on agreement concerning the *acceptable extent of disagreement* between different observations, known as *experimental error*.

2.2. Sources of unreliability

Sources of disagreement, or interobserver unreliability, arise in a variety of recognizable ways and can often be dealt with if we know how the disagreement arose. Muckler and Seven (1992) have reviewed the well-known sources of interobserver unreliability including simple errors of observation and recording to prejudice, halo effects and even deliberate distortion. The methods of sampling and experimental control developed in the psychological laboratory are aimed principally at reducing these sources of error.

In judgements of internal sensations, the use of descriptors whose meaning is shared is often valuable. In describing pain, for example, analogy to external objects is frequently used as in 'sharp', 'piercing', 'dull' and so on, and anchor statements referring to publicly observable objects or events such as feeling 'worn out' or 'yawning' are common in subjective rating scales (Åhsberg 1998).

Disagreement about subjective observations of public events can arise simply from the failure to agree in advance the criteria by which observations should be classified. That favourite tool of ergonomists, the checklist, is a common example of a procedure for enhancing shared meaning. Each individual checklist item should be unambiguous (is 'X' the case?) and the list of relevant items is agreed between participants, or at least omissions or disputed items may be identified. The widespread use of checklists supports the view that agreement on criteria is important and Stanton and Young (1999) find that interobserver reliability of checklists 'surpasses all methods'.

The checklist principle can be taken even further into the realm of quantification. Take, for example, the 'subjective' assessment of how well a team is performing in terms of characteristics such as 'communication' and 'coordination'. The TARGETS (Targeted Acceptable Responses to Generated Events or Tasks) methodology (Dwyer *et al.* 1997, Annett *et al.* 2000) defines these general concepts of team skill in terms of specific behaviours that occur in the context of identifiable 'trigger' events when the team is performing a task. The behaviours can then be simply counted and a score derived from the probability of their occurrence in the given context and we then have an objective measure of the team skill variable. The scores can be taken as objective since each of the contributory observations can be readily agreed by independent observers.

Classical psychophysics recognized a number of systematic sources of error in making comparative judgements (Woodworth 1938). The second of two weights typically appears heavier than the first. This effect, known as *time error*, can be attributed to the time between the presentation of the standard and the variable stimulus when the standard is held in memory. In psychophysical experiments the time delay is often only a matter of seconds but in the direct rating methods of the kind used by ergonomists the standard is not physically present and can only be recalled from memory. Ratings given after the end of a test period when the observer has been working on a task or has been exposed to some environmental stress must be entirely dependent on memory. The alternative, to call for ratings during testing (Tattersall and Foord 1996) may of course disrupt the performance but there has been little systematic study of possible time error in relation to ergonomics rating scales. Even if the observer is required to produce ratings immediately following the test, the very fact that it is a numerical assessment implies a comparison with a range of related items. The rater then has to call to mind those experiences thought to be relevant and to arrange them on a scale, placing the item to be rated on that scale by

comparison with those drawn from memory. This dependence on memory for previous experiences may well depend on the extent of previous experience and an expert may be expected to be able to draw on a wider, probably more representative range, of comparators.

Context effects have been widely recognized as a common source of bias in subjective judgement. The best-known examples include the geometrical illusions such as the Müller–Lyer and cases such as the size–weight illusion in which the larger object appears lighter than a standard of equal mass but smaller volume. Ergonomics ratings are rarely if ever free of context and they are typically multidimensional, meaning that variations in one sensory attribute may well affect judgements in another. For example, a number of different task features have been shown to affect overall ratings of workload (Wierwille *et al.* 1985) so that individual features may have different weightings when in different combinations. Adaptation level may well play a part in ratings of workload and effort. Colle and Reid (1998), for example, found that much higher workload ratings of a standard task were found when raters had experienced a relatively easy range of tasks and the converse was found when the range was of high difficulty.

2.3. *Limitations of consciousness*

Methods involving introspection present a number of special problems. It is generally recognized that consciousness is limited in the sense that attention is not necessarily paid to every source of stimulation, even when the relevant sense organs may be activated. Unattended sensory activity cannot form the basis of a subjective report (Nisbett and Wilson 1977). Ergonomists will be familiar with the distinction between automatic and controlled processing (Shiffrin and Schneider 1977) and with the limitations of working memory (Baddeley and Hitch 1974) both of which present problems for the completion of post-performance ratings. Even highly experienced operators can sometime find difficulty in identifying the cues they use and in reporting familiar experiences and actions.

Introspective psychologists from Wundt to Titchener believed that introspection was itself a learnable skill and went to some trouble to train their ‘observers’. Expert test pilots and vehicle drivers clearly learn through long experience to identify and describe subtle dynamic features of the systems they are testing even if the terms they use, such as ‘understeer’, remain mysterious to the uninitiated. Little is known about this learning process but it is nonetheless clear that the subjective ratings of these skilled practitioners can be of greater value to the designer than the immediate reactions of a sample of naive individuals because they have learned how to identify subtle characteristics of the system. The choice of a sample of naive participants from which to record subjective reactions may, however, be entirely appropriate to the design of public systems destined for the use of untrained individuals. When to ask for subjective ratings from a small sample of experts and when to collect data from a large random sample depends not on some abstract notion of scientifically acceptable practice but on a rational appraisal of the purpose of the exercise and the likely value of the data for that purpose. For example, Whitaker and Marsh (1997) employed experienced air-traffic controllers as participants in a simulated trials of a new European Air Traffic Management system. Both objective performance measures and subjective ratings were taken during extensive runs comparing several different versions of the system. The results yielded useful insights yet it was clear to the investigators that the raters were far from passive users and they found it difficult

to accept that an experimental trial is different from a pre-operational evaluation. Their involvement as potential future users of the system may have distorted their judgements by either over- or underestimating the workload provided by different configurations. In an earlier study, Kelly *et al.* (1995) noted that their expert participants frequently ignored the specially designed computer-assisted tools and continued to do their usual job. It was not surprising that their judgement of the usefulness of these tools was rather negative.

2.4. *Subjective experience as cognitive appraisal*

The concept of *cognitive appraisal* arising from the James–Lange theory of emotion (James 1890), elaborated by Schachter (1964) as *cognitive labelling theory* and modified by Lazarus (1982) as *cognitive appraisal theory*, raises some fundamental questions about the relationship between our feelings and the physiological mechanisms which supposedly underlie them. The James–Lange theory proposed that ‘I feel frightened because I am running away from the bear,’ rather than ‘I am running away because I feel frightened.’ A well-known experiment by Schachter and Singer (1962) involved administering doses of adrenaline together with sometimes misleading information about the symptoms likely to arise from the drug. It was found that reported subjective experience was clearly affected by the information provided by the experimenter and by the behaviour of a stooge supposedly receiving the same drug but behaving as if either angry or euphoric.

A careful reading of the literature on stress (Cox 1978) and fatigue (Holding 1983) leads to an interpretation of these two kinds of experience as providing the basis for choices of action. Thus, a physical stimulus may trigger a stress response, a combination of autonomic and behavioural activity, only if it is perceived or appraised as dangerous and demanding avoidance behaviour. Holding noted that one of the common effects of prolonged engagement with a demanding ‘task is not just to generate feelings of fatigue but also changes of strategy, especially taking risky shortcuts’.

The primary role of consciousness is not just to experience raw pleasure or pain but to assess the current situation in relation to personal goals. The experience of subjective workload is telling us if more effort or a different technique is required, or the goal is to be downgraded or abandoned. Experience of fatigue is telling us to adopt some less effortful activity, discomfort invites attempts to modify aspects of the environment, usability determines consumer choice, urgency signals the need to change task priorities.

2.5. *Summary*

In summary, subjectivity is inescapable in science as in everyday life. As Muckler and Seven (1992) remarked, ‘The distinction between objectivity and subjectivity is not a useful way of distinguishing among human performance measures.’ Progress in science depends on the degree of shared meaning between individuals concerning their observations and experiences. Scientific method attempts to maximize intersubjectivity, that is to minimize disagreement between independent observers. In the practical world of ergonomics the precise levels of control striven for, and sometimes achieved, in the psychophysics laboratory are rarely possible but the users of rating scales should take every care to avoid the principal sources of error outlined above. Since some sources of disagreement are well known we can take steps to maximize interobserver reliability. In the case of objects and events external to the

observer agreed observational categories and criteria can greatly reduce 'subjectivity' whilst internal events are often communicated best by the use of well-chosen analogy. Although it may be possible to train individuals to make consistent introspective judgements we know little of this process but it is clear that conscious experience can be quite limited, especially when carrying out practised tasks and sensory data which inform conscious experience may be highly selected. In making use of subjective response data, it is important to appreciate how the conditions of data collection and its perceived purpose may affect the observer's interpretation of the task and the related subjective experience. The conscious experience, on which subjective ratings are inevitably based, may be described as a process of cognitive appraisal of the demands of the task and environment on the resources of the individual to be used in conjunction with aims and intentions to control the choice of strategies.

3. Measurement

3.1. *Psychometrics and quantitative structure*

The psychophysical methods devised by Fechner (1860) relate the intensity of sensation to the physical magnitude of the stimulus. There are, however, other psychological variables for which there exists no comparable physical scale, and their measurement is termed *psychometrics*. Fechner himself adapted his psychophysical methods to the purpose of measuring abstract attributes such as 'beauty' (Boring 1957). Ergonomics constructs such as workload, fatigue, comfort, etc. are analogous to familiar psychometric constructs such as intelligence or anxiety and share many of the logical and technical problems associated with their use.

It has been argued (most recently by Michell 1997) that psychometrics is unscientific since attempts to measure psychological attributes violate the basic requirements of scientific measurement, notably these attributes lack *quantitative structure*. An essential feature of measurement is agreement on standard units (metres, grams and so on) which are always equal and can be counted and summed, the principle of *additivity*. Quantitative structure can be empirically determined, but whereas the quantitative nature of the attribute 'length' is almost intuitively obvious, the appropriate tests for quantitative structure in psychological attributes are less clear. Attributes such as intelligence and extraversion, although assigned values on a numerical scale, clearly lack an identifiable unit of measurement. Such scales are at best ordinal since, without agreed units, they cannot possess interval or ratio properties. This, according to critics such as Michell, throws serious doubt on their scientific value. There are a number of ways in which this apparently fundamental criticism has been met.

3.2. *Additive conjoint measurement*

The method of *additive conjoint measurement* developed by Luce and Tukey (1964) is claimed to provide a test of quantitative structure based on the ordinal relationship between measurements. It is applicable where the construct being measured has two or more components each of which can be manipulated independently. If we take the hypothetical example of subjective workload as a dependent variable which has both a physical and a mental component each of which can be varied so as to produce three levels of each, the data would be summarized in a 3×3 cell matrix such as that shown in figure 1.

| | Lo | Med | Hi |
|-------------------------|--------|--------|--------|
| | Mental | Mental | Mental |
| | (M1) | (M2) | (M3) |
| Lo Physical (P1) | 1 | 2 | 3 |
| Med Physical (P2) | 2 | 3 | 4 |
| Hi Physical P3 | 3 | 4 | 5 |

Figure 1. Conjoint measurement of physical and mental workload.

In this artificially simple example, the numbers in the cells represent subjective ratings of workload produced in response to the nine combinations of physical and mental task difficulty. All the numbers in the rows and columns are ordered in the same way. It is also true that the value in P2,M1 is equal to or greater than the value in P1,M2 and that $P3,M2 \geq P2,M3$ and $P3,M1 \geq P1,M3$. The fact of common ordering of row and column entries is known technically as *single cancellation* and the three comparisons comprise the criterion of *double cancellation*. Kline (1998) argues that if the data meet both criteria then the dependent variable, in this instance subjective workload, can be said to have quantitative structure in the sense required by classical measurement theory.

The question of quantitative structure is purely empirical and has rarely been asked of psychometric test data and even more rarely for ergonomics scales. An exception is Reid and Nygren's (1988) claim of additivity for the Subjective Workload Assessment Technique (SWAT) on the basis of conjoint measurement of the three dimensions of *time load*, *mental effort load* and *psychological stress load*. Each was rated on a three-point scale such as 'Often have spare time ...' (rated 1), 'Occasionally have spare time ...' (rated 2) and 'Almost never have spare time ...' (rated 3). In developing this procedure, the participants were required to imagine task situations and events to match the 27 ($3 \times 3 \times 3$) combined descriptors which were then placed in a rank order. From the mean rankings of the different descriptors a scale value was assigned, thus the scale value of the descriptor rated 1 for time load, 3 for effort and 2 for psychological stress was found to have a value of 52.1 on a scale 0 to 100. Given a margin of error variability of 5–10% in

participants' rankings, it is claimed that SWAT meets the additivity criterion and could therefore be said to have quantitative structure.

3.3. Stevens's solution to the problem of quantification

Stevens (1951, 1956) was well aware of the quantitative limitations of the measurement scales used in psychophysics and offered an alternative solution in the method of *direct estimation*. In this method, magnitudes of sensation can be directly mapped onto numbers (both whole numbers and ratios). Stevens's method of *free modulus estimation* simply requires the observer to match stimulus magnitude with a freely chosen number. A measurement scale is derived simply by mapping the numbers provided by the participant or observer over a series of stimuli onto the physical magnitude of the stimuli. (See, for example, the method for measuring the perceived urgency of warning signals used by Hellier *et al.* 1995.) In the case of psychometric scales, that is where there is no identifiable physical counterpart, the numbers can still be assumed to represent the magnitude of the unknown psychological dimension by the following logic. The method of *cross-modal matching* in which observers are asked to select, for example, a light which is as intense as a given sound is loud, provides the paradigm in which the number system is taken to be just another sensory dimension (Stevens 1971). The numerical estimation of psychological magnitudes is then simply a special case of cross-modal matching.

Consider, as Stevens's proposal would imply, the concepts of *number* and *magnitude* as psychological phenomena. Butterworth (1999) has argued for the existence of an innate sense of number (sometimes referred to as *numerosity*) which does behave in the way we expect of other sensory/perceptual dimensions. Neonates have been shown to respond to differences between the numbers of objects in a visual array, at least up to the number four and at least one case of an inability to perceive number as such has been described by Cipolotti *et al.* (1991). A study by Underwood (1966) found that the difference threshold for the numerosity of arrays of dots was roughly consistent with Weber's Law *provided* the standard comprised more than 16 dots. At smaller numbers observers did considerably *better* than predicted on the basis of a simple sensory dimension and might well have been employing a different strategy, perhaps based on counting or *subitization*, that is recognizing a visual pattern such as the dots on a domino piece. The truth is we know rather little about how potentially 'pure' numerosity is influenced by the nature of the objects being judged and the circumstances under which the judgements are made and we cannot assume that the rigorous conditions of the psychophysics laboratory are reproduced.

Homo sapiens is by no means the only species able to judge numerosity. Both predator and prey animals generally know when they are outnumbered, and hence when to stay and fight or make a run for it. However, we among all species have developed the trick of counting by making use of verbal symbols as counters in short term memory. This trick enables us to know that the quantity 1001 is bigger than 1000, and to engage in the manipulation of numbers known as arithmetic, something we would not be able to do by unaided perception. Counting and calculation in general is, unlike an intuitive impression of numerosity, a *cognitive* operation, a conventional procedure we learn to perform once we have acquired some basic linguistic skills. It is not the same as a sensory dimension such as distinguishing the weights of different objects which we do not have to be taught and so Stevens's basic premise must be rejected. However, the method of direct estimation can still be

useful in establishing the ordinal relationships between items varying on a given dimension.

3.4. *Measurement and reality*

Whether or not a thing can be measured in the strict sense of the term, is bound up with what is believed to be its *real* nature and in particular whether or not it has quantitative structure. If there is a two-way relationship between the attribute being measured and the numbers used to represent that attribute, then this kind of measurement is deemed *representational*. The question is whether proposed measures of these attributes are representational. If they are not, then they cannot be properly measured nor can psychometrics be deemed scientific. All the quantitative methods used in psychology contain normally hidden assumptions about the reality to which they refer.

Classical experimental procedures in psychology are generally based on assessing the effects of treatments, independent variables mostly defined in physical terms such as the strength, duration and type of the stimulus, on other physical events such as response types, frequencies and latencies as the dependent variables. The underlying reality comprises the hypothesized mechanisms that mediate between stimulus and response. Reaction time, and variants on it are presumed to reflect real information processing; errors and omissions in recall tests provide measures of memory capacity and so on. Measures such as these are accepted as valid for testing hypotheses about the underlying physical mechanisms and their quantitative structure is not in doubt.

Specifying the quantitative structure of individual characteristics such as abilities and personality traits is more difficult. Operationalism, 'characteristic X is what tests of X measure', has been the fall-back position to justify psychometric procedures, but is rejected by the pure measurement theorists who can find no tangible reality and hence no quantitative structure behind attributes such as intelligence or extraversion. Psychometrics attempts to model individual differences, that is the many ways in which individuals vary, in terms of their ability to cope with different kinds of problems and their typical responses to standard situations — but what is the underlying reality? Kline (1998) has addressed this fundamental issue in his *The New Psychometrics*. Psychometrics is not simply the statistical exploration of arbitrarily collected test data, but a genuine attempt to come to grips with processes underlying the observed patterns of intercorrelations.

Attempts to identify physical features of the nervous system which account for observed variations in performance and personality, such as Eysenck's theory of inhibition/excitation as the mechanism underlying the introversion/extraversion dimension (Eysenck 1967) have met with limited success. However, we can determine the empirical structure of abilities and traits in terms of their interrelationships using multivariate techniques such as factor analysis. The reality underlying test results may be said to lie in the structure of the relationships between data points. However, establishing these relationships is but one step towards a more complete understanding of the physical mechanisms which give rise to the observed patterns of behaviour. Whilst intelligence and personality traits are regarded by some as essentially social constructs which may never be characterized in purely physical terms ergonomics constructs appear to be much closer to features of the physical world, the environment and work done. The patterns we find in the data derived from subjective rating scales of workload, fatigue and so on comprise the first step towards developing hypotheses concerning the underlying physical mechanisms, but

in the meantime they have practical value. The pragmatic response to the objections of the strict measurement theorists is that these methods work in the sense that they can be successfully used in conjunction with appropriate theories to predict further empirical observations. Measurement systems in which the numbers do not strictly represent the physical properties of the things being measured, but are used to indicate the empirical relationships between measures of theoretical constructs are referred to as *index measurement* (Dawes 1972). The distinction between representational and index measurement will be illustrated in the next section where we consider the *validity* of subjective scales.

4. Validity

4.1. Construct validity

Workload, fatigue, comfort, urgency, etc. are constructs based on common experience and shared meaning. The first task for the scale developer is to establish construct validity by providing a coherent theoretical framework that fits the relevant observations. The construct *fatigue* illustrates some of the implications of the special role of subjective measures within the theoretical framework. Fatigue has been in use as a technical term at least since the late 19th century when W. H. Rivers, who had studied under Kraepelin, introduced the topic to the newly established Cambridge Psychological Laboratory (Hearnshaw 1964). During the First World War it became a substantive research area but has resisted precise definition (Holding 1983, Åhsberg 1998). Muscio (1921) questioned whether a test of fatigue was possible and proposed the term be dropped (in 1929 the Industrial *Fatigue* Research Board, established in 1916, became the Industrial *Health* Research Board). The problem was that fatigue was at the same time both a subjective phenomenon, feeling tired and so on, and a pattern of behaviour including taking shortcuts, making errors and omissions and exhibiting irritable behaviour (Bartlett 1943). Some took these signs and symptoms to be due to a complex but unitary physiological state. It became clear that fatigue is a *multidimensional construct* composed of at least the three elements, namely certain kinds of subjective experience and aversive behaviour loosely linked with certain physiological processes.

Many ergonomics constructs comprise both objective and subjective elements but even the subjective experience itself may be multidimensional. For example, SWAT acknowledges three principal subjective dimensions, *time load*, *mental effort load* and *psychological stress load* (Reid and Nygren 1988), whilst the NASA-TLX comprises six distinct subscales including *overall workload*, *physical effort* and *fatigue* (Hart and Staveland 1988).

Probably the principal reason why Muscio and others found fatigue a slippery concept is the lack of shared meaning, or intersubjectivity, discussed in Section 2. This problem can be greatly reduced, if not entirely eliminated, by the use of modern methods of data collection and analysis. Åhsberg (1998), for example, collected 172 verbal expressions used by a sample of people to describe their own experiences of fatigue following work. These were first edited down to 95 items and then incorporated into a questionnaire answered by 705 people from various occupations about how they felt after work. The ratings applied to these items were subjected to factor analysis and five factors were extracted (*lack of energy*, *physical exertion*, *physical discomfort*, *lack of motivation* and *sleepiness*) and these are interpreted as the dimensions of experienced fatigue.

Many characteristics of work such as duration, frequency and energy cost can be specified in physical terms and are quantifiable in the representational sense, but the data derived from subjective ratings may be said to reflect the *perceptions* of the raters, as reflected in their choices and preferences. Multivariate techniques, including Multidimensional Scaling (MDS) which map perceptions in n -dimensional space, avoid two problems. First, the participant/observer is not required to provide numerical ratings but only judgements of similarity (or dissimilarity) between pairs of objects. Preferences can, of course, be derived from ratings and rankings, but the observer is not technically required to provide them directly. MDS solutions for ratio scale data have been found very similar to those obtained from purely ordinal information (Borg and Groenen 1997). The second advantage of MDS is that there is no need to attempt to describe or prescribe the precise attributes to which the observer must pay attention and no assumptions need be made about the contents of the observer's consciousness. The effective dimensions are inferred from the data.

There is, however, one very important condition that it may not always be easy to satisfy in practice, that is the selection of a set of items or descriptors from which to construct a model. For example, in a study of workload Derrick (1988) employed 18 laboratory tasks involving tracking at different levels of difficulty, either singly or with secondary auditory and visual detection tasks and a tone judgement task, all to be rated in terms of their similarity of task difficulty. Analysis of the data identified three dimensions of perceived task difficulty, one associated with *processing resource cost*, the second with the *input modality* and the third with *time stress* and *response complexity*. These are dimensions that may not have occurred spontaneously to the observers, but were determined by mapping the statistical structure of their responses onto a theory of the processes underlying task difficulty. This methodology suited the purpose of the study, namely to investigate the resource model of workload (Derrick and Wickens 1984), but cannot be interpreted as identifying *all* the dimensions of workload which might be involved in a wide range of industrial tasks as, for instance, is attempted in the study of perceived fatigue by Åhsberg *et al.* (1997). A structural equation model (confirmatory factor analysis) based on responses to 25 verbal expressions describing subjective experience such as feeling worn out, numb, indifferent, sleepy and so on, provided the measures, or *indicators*, of five *latent variables*, was found to provide an acceptable fit of the data to the model.

The SOFI studies (Åhsberg 1998) were based on a representative sample of observers of all ages and both sexes who were employed across a range of 16 different occupations from bus driving to nuclear plant operation. Although these studies were broadly based, they do not claim to be a comprehensive measure of fatigue for all aspects of all occupations. The typical workload study employs fewer than 100 raters, often drawn from a very specific population such as air-traffic controllers or students. As Nygren (1991) has pointed out, scales such as the Cooper–Harper, based on checklists, have often lacked validation against external criteria. The same can be said of usability scales such as SUMI (Porteous *et al.* 1993). We have to conclude from the range of published studies that, whilst no doubt tapping a great deal of common ground, we are a long way from prescribing single measures of constructs such as fatigue and workload which are applicable across a fully representative range of industrial tasks.

4.2. *Predictive validity*

The justification often advanced on behalf of psychometric measures is that, whatever the doubts of some pure measurement theorists, they *work*. The use of subjective rating scales, as with physical measures of work and the environment, is justified to the extent that they have *predictive validity*, that is test results predict performance on some specified criterion measure or measures. It is vital that the criteria should be agreed in advance either by scientific consensus or, in the case of practical ergonomics, by the stakeholders of the system under investigation.

Underlying both construct validity and predictive validity there should be a theory about how variations in some measure of that construct influence one or more criterion variables. Generally in ergonomics these can be of three kinds. They may be behavioural such as speed, errors and omissions; they may be physiological such as heart rate, $\dot{V}O_{2max}$, EMG, EEG, etc.; they may be subjective reports such as preferences, ratings and rankings. In any given study, one of these types of criterion variable may be of primary interest. Behavioural variables such as those related to pilot error have been predominant in ergonomics but physiological variables, especially if linked to occupational injury and disease, also merit attention. The third category becomes especially significant in consumer ergonomics and in user acceptance, including usability and presence. All three classes of outcome variables are implicated in fatigue and should play their part in developing an explanatory theory. Furthermore, the three types of variable are often found to be correlated leading to the possibility that each may be useful as a predictor of one or more others. Physiological measures of workload have in particular proved attractive to investigators because they appear immune to the biases and distortions of subjective judgement and are acceptable as ‘hard’ evidence.

Figure 2 represents a generic model of a typical ergonomics construct, such as workload. The construct itself is not predefined but is envisaged as a complex response to a *challenge*. The challenge comprises those variables defining the task, the object or environment to which the observer is responding. The construct itself is a *latent variable* defined in terms of a set of *manifest* (or *indicator*) variables, that is the actually observed data.

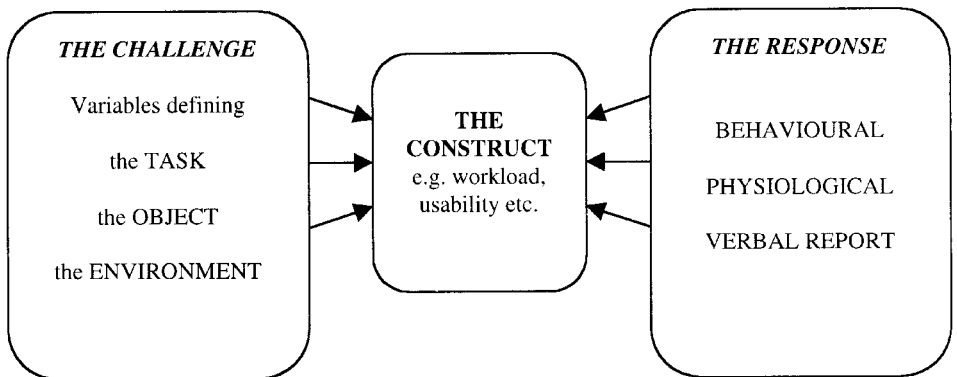


Figure 2. Generic model of ergonomics constructs as latent variables defined in terms of two sets of manifest variables.

On the right hand side are represented various classes of dependent variables which theory supposes are likely to vary with different values of the construct. For example, we might suspect that increases in workload produce deteriorations in performance as well as verbal complaints. If the construct were, say, *urgency* (of a warning signal), then the challenge might be specified in terms of characteristics such as loudness, pitch, interruption rate etc. and the dependent variables we might expect to vary with *urgency* might be giving priority of attention to the signal, possibly a change in skin conductance or heart rate, and probably a verbal comment, 'I must act on this at once.'

The challenge is specified in terms of attributes of the task the investigator presumes to be responsible for generating workload such as information to be processed or energy to be expended over the period in which the task is performed. Appropriate measures would be selected which specify the response to the challenge, for instance physical work accomplished, information transmitted, duration and location of work, etc. If the construct were *comfort* then the challenge might be specified in terms such as the hardness of the seat, the frequency and amplitude of vibration, or combinations of these. If climate is of concern then measures such as temperature, wind chill and so on might be selected. For each study, an appropriate selection of measures is made. Whiteside *et al.* (1988) specified the *usability* of an electronic conferencing system in terms of 10 measures. These included both performance measures such as number of successful interactions in 30 minutes, error rate, and learning rate and subjective responses including preference over alternative systems and attitudes towards various aspects of the system. The results were used to improve the design of the system by an iterative process rather than with the objective of establishing general usability criteria.

4.3. Dissociations between subjective and objective measures

Attempts to relate *perceived effort* with physiological measures and *mental workload* with error rate, for example, have been only partially successful. It seems that subjective experience and physically measurable variables are not always monotonically related. Hankins and Wilson (1998) found measures of heart rate, heart rate variability, ocular activity and EEG were related to different levels of task activity during an actual flight and some of these physiological measures were intercorrelated, but none of them correlated significantly with the TLX subjective workload scores nor were they found to be related to observed task performance criteria. Results such as these reinforce the appeal of 'objective' measures of mental workload based on physiological data. Obviously more empirical work might clarify these uncertainties and new techniques which are more sensitive and convenient are always possible (Mital and Govindaraju 1999) but we may suspect a more fundamental problem and this is to do with the functional role of conscious experience which is to carry out a *cognitive appraisal* of available sensory information leading to an appropriate course of action, as discussed in Section 2.

Yeh and Wickens (1988) offered an explanation for dissociation between subjective and performance measures of workload. Using several laboratory tasks, including tracking, a memory task and a simulated air-traffic control task, they investigated the hypothesis that, since working memory is closely associated with conscious awareness, demands on this particular resource will be more strongly represented in subjective reports. Performance may not always deteriorate with an increase in subjective workload simply because the operator, as a result of appraising

the disparity between demand and performance, invests more resources in order to meet task demands. They concluded that in evaluating a system, subjective measures may be useful in indicating potential performance problems if task demands are further increased. This may well be true but only serves to illustrate the general proposition that the choice of performance, physiological and subjective criteria must always be made with the purpose of the investigation in mind. Often performance measures provide the essential criteria by which choices are made between different designs but we must reckon with the costs of performance in other terms. For example, strongly measurable physiological responses may indicate not just current workload but may, if continued, be the warning signs of future organic disease. In the field of consumer ergonomics, subjective criteria are likely to be preferred since consumer choice is all-important. Even when the operator has no choice in the equipment or the tasks to be performed with it, user dissatisfaction may have consequences in the longer term even if there are no detectable effects on current performance.

4.4. *Conclusions on predictive validity*

Having inspected a wide range of reports involving subjective ratings, it is hard to avoid the conclusion that studies of the predictive validity of scales based on them are few and far between. There could be a number of reasons for this. First, publications on scale development are rightly concerned primarily with establishing construct validity, the internal coherence and theoretical integrity of the proposed measure. Second, subjective ratings are often employed in the developmental stages of a project and long-term follow-up of relatively successful and unsuccessful projects, in the same field, are extremely rare. Terman's validation of the Stanford Binet intelligence scale (Terman and Merrill 1937) involved follow-up studies correlating measures of life success, such as socio-economic status, with childhood IQ. Although ergonomists may be convinced of the value of their predictions, substantive comparative evidence is rarely offered, probably because poorer designs tend not to survive.

5. Utility

There is only one golden rule for the selection of measurement instruments, whether subjective or objective, and that is that they must be appropriate to the purpose of the study. The catch is, of course, that appropriateness is itself a matter of judgement. Ergonomics may be scientific, but planning research is an art! There are several practical guides to the use of subjective methods such as Sinclair (1995) and Charlton (1996) and a host of texts on questionnaire design, rating methods and the relevant statistical techniques. I shall not attempt to summarize them here, but it might be useful to discuss the kinds of factors that can affect the choice of suitable measures.

5.1. *Relevance*

The first question to ask is whether subjective experience of the user/operators is relevant to the purpose of the study. If the chosen criterion variable is acceptable levels of *comfort*, *noise annoyance* or qualities such as *usability* or the *presence* of a VR system, then subjective measures are clearly indicated. In other cases, subjective measures may be seen as contributory or convenient rather than crucial. For example, in designing a warning signal it may appear appropriate to use a subjective

measure. However, *perceived urgency* is unlikely to be the sole criterion attribute unless it is closely related to the appropriate behaviour *in the context in which it is to be used*. We need to know not only if the signal *sounds* urgent, but also if the operator responds to it by making an appropriate reassignment of task priorities, and we may not be entitled to assume we know the two are well correlated. Subjectively measured workload may be of interest in itself, but more often we are also interested in using it as a predictor of operator performance. It has been suggested that subjective measures of mental workload can predict performance better than physiological measures during primarily cognitive tasks (Sheridan and Stassen 1979) and that subjective measures are more sensitive than performance measures at low task loading. As Muckler and Seven (1992) remark, objective measures tell us what is happening, but subjective measures can tell us how we are coping and thus provide a warning of possible future changes in performance.

5.2. *Level of measurement*

The intended use of a measure is also highly relevant to the choice of instrument and hence the type of data to be collected. In the early stages of an enquiry, say the development of a new design, diagnosis may be more important than quantification *per se*. If we need to know in what ways the design may fit, or fail to fit, the requirement, an open-ended questionnaire is the obvious starting point, especially when the system under development is complex and has many attributes. Only after having established that attribute X is subjectively significant, and perhaps having further developed the design, do we become interested in *how much* better (or worse) is the relevant feature.

Although the use of interval or ratio measures is routinely recommended as superior to the qualitative and ordinal measures, experimental comparisons of two or more conditions, treatments or designs often require only an ordinal result. In such cases, it is often enough to show that A is preferred to B without having to show by how much. However, full quantification would appear to be required if the aim were to establish general design standards. In this case, we may need measures which at least approximate interval scales (Charlton 1996) and many *ad hoc* subjective scales do not meet this requirement. Only scales for which there are established norms standardized on the appropriate populations of both tasks and users are likely to have interval properties. Their development is a long, complex and expensive business by comparison with the 'quick and dirty' questionnaire or rating scale which served a useful purpose in the design of a specific system. In fact, there are relatively few published scales suitable for this purpose. Most studies of mental workload using scales such as SWAT and NASA-TLX have been carried out on air pilots or air-traffic controllers. Little information is available about the use of these measures in other tasks, even those which might be closely related such as the broad range of vehicle control (cars, trains, cranes) or command-and-control in military or emergency services, telephone call centre operation, to say nothing of the almost infinite range of information-processing tasks from school teaching to share dealing. Sanders (1979: 48) concluded that 'the prospects of measuring mental load by subjective judgements are not high. Perhaps the method is applicable to limited sets of task conditions'. Whilst subjective responses to information overload, such as a feeling of being under stress may be common to a wide variety of situations, the specific set of circumstances which give rise to this response are almost infinitely variable according to specific features of the task combined with the individual's

training and experience and motivation. Subjective comparisons of similar tasks made by individuals who have similar training and motivation can be useful in problem diagnosis and developmental ergonomics (Kelly and Goillau 1996, Whitaker and Marsh 1997). The use of subjective scales for establishing general standards such as ISO 10075: 1991 and 10075-2: 1996 is highly problematic.

5.3. *Cost and convenience*

The utility of subjective measures can be compared with that of physiological and behavioural measures. By common consent subjective, or pencil and paper, methods often present a cheaper alternative to those involving instrumentation, be it sensors for physiological recording or video for recording behaviour. Subjective ratings are, in one sense, non-intrusive in that they do not require the attachment of electrodes, breathing bags and the like or taking blood and urine samples, but may require the participant to perform an additional task such as answering a question or pressing a button to register some current subjective state such as stress. The usual alternative is to record ratings and opinions some time after performing the task, but this introduces the possibility of time error as discussed in Section 2. It is always possible to carry out independent studies in order to provide estimates of the effect of this 'secondary task' on primary task performance and the possible bias on ratings introduced by time delay, but few investigators can afford this luxury and in practice make their choice intuitively. Typical of the problem of choosing between subjective and physiological measures is the complaint of Lindh and Gårder (1993) attempting to assess differences in stress on drivers induced by different road markings. Road traffic researchers, they suggest, are sceptical of subjective measures yet the favoured alternative physiological measure of Respiratory Sinus Arrhythmia required the participant to follow the instruction 'inhale deeply now' and compliance could not be guaranteed. Under these circumstances, the choice of subjective measures was entirely reasonable.

5.4. *Ease of data collection*

As discussed in Section 3, there are a variety of techniques for obtaining subjective data, including direct numerical estimation and paired comparisons, both with a number of subvarieties (Sinclair 1995, Charlton 1996, Stanton and Young 1999). In purely technical terms, paired comparison using forced choice has been the preferred method since the early days of psychophysics. The paired comparison method reflects the basic function of human judgement, namely that of choice between alternatives based on weakly specified criteria. Rank ordering and direct numerical estimation can be seen as derivative of preferences between pairs contained in the overall set, be it of objects varying in weight, tasks varying in difficulty or displays varying in presence. A special advantage of comparison methods as used in multidimensional scaling is that it is not even necessary to specify the attribute or dimension to be compared, but simply to ask for a judgement of similarity/dissimilarity. MDS data analysis then allows the determination of the stimulus dimensions used by the respondent, whether consciously or not, and may additionally prevent the investigator from imposing his or her own presuppositions on the nature of the judgement by posing particular questions or addressing them to irrelevant attributes.

Forced choice, which is a normal feature of paired comparisons, has long been understood to be the most sensitive psychophysical method (Woodworth 1938). The

'balanced' Likert-type five- or seven-point scale with a central indifference point which is often the modal response, can be wasteful since it can reflect the unwillingness of the participant to make a judgement. Its use is sometimes justified in terms of an appeal to the respondents who resist the requirement to make a guess (Charlton 1996). Another possible justification is that inspection of the data can quickly reveal the shape of the response distribution which may, in turn, affect the choice of statistical analysis.

Paired comparison methods also carry a penalty that can be more important in practical ergonomics than in the psychological laboratory, and that is the provision of an adequate number of instances. In many ergonomics studies only one pair is ever considered, the old design and the new. The use of a number of respondents permits us to estimate the population preference but in itself tells us nothing about the judgmental dimensions and their weightings which they apply to the attributes of interest. An alternative is to invite paired comparisons between *imagined* or remembered samples with the problem of time error or even differential experience previously referred to. Such methods are likely to reflect the opinion, including the biases, of the respondents, but not necessarily their response to the actual situation. Ratings based wholly or mainly on judgements made from memory are intrinsically suspect.

6. Discussion

The historical distinction between subjective and objective knowledge is not quite what it may have seemed. I have argued that all knowledge is based on subjective experience. What really matters in establishing scientific truth is the method by which independent observers agree on the meaning of their individual observations. Complex features of the external world such as the handling characteristics of a vehicle or the degree of cooperation in a team may be judged subjectively by experts, but the reliability of these judgements depends heavily on the use of agreed criteria and so methods based on detailed checklists provide the best assurance of intersubjectivity.

Other strictly subjective scales, such as those concerned with *fatigue*, *pain* and *annoyance*, for example, refer quite specifically to 'private' experience and here we need to take into account a variety of factors known to affect judgement. Practice and familiarity have a considerable effect on conscious experience with both positive and negative consequences. An expert may be better able than a novice to identify some subtle variation in the attributes of a complex task or object, but sheer familiarity with the task may mean that significant cues go unnoticed and unreported. Other factors that affect conscious experience include the individual's overall appraisal of the available information.

Subjective rating scales are based on the assumption that the human participant normally responds quantitatively to variations in the specified sensory attribute of the stimulus object or situation. However, careful examination of the evidence suggests subjective judgements are based on a more complex mechanism referred to here as *cognitive appraisal*. The selectivity of consciousness is greatly influenced by the individual's current agenda, the goals, motives and plans as perceived in the context of a personal model of the world. Subjective experience can depend as much on this interpretative process as on the actual state of the proprioceptors, and this is especially likely to be so in emotionally charged situations where a change of strategy is likely and the behavioural choice is between approach and avoidance or between

fight and flight. Such judgements may be especially vulnerable to this kind of distortion, but there is no reason to believe that, despite this 'subjectivity', that physiological indicators are any more scientifically valid than subjective measures.

Subjective attributes, it is said, lack *quantitative structure* and thus fail to meet the key criterion on which scientific measurement is based. Psychometricians have offered several solutions including direct estimation of magnitudes and ratios and the demonstration of additivity by *additive conjoint measurement*. Direct estimation methods, whilst giving the appearance of interval or ratio properties, are essentially ordinal. Ordinal, sometimes called non-metric, measurement is the basis of subjective scaling and this is facilitated by the use of mathematical techniques, including varieties of multivariate analysis such as factor analysis, multidimensional scaling and structural equation modelling.

Without exception, the attributes that ergonomists aspire to measure by subjective ratings are multidimensional. They are constructs whose everyday meaning can be refined by multivariate statistical methods. Multidimensional scaling can be used to identify the principal sources of variance, or dimensions, of subjective judgement without requiring the participant or observer to indulge in the difficult and often suspect process of introspection. The construct validity of subjective measures such as *fatigue* and *workload* can be greatly enhanced by the use of these methods, but of course the choice of items to be judged, questions to be asked and populations to be sampled are always in the hands of the investigator and have to be defensible in terms of the wider theoretical and practical context.

The classical justification of any scientific method is that 'it works' and the validity of subjective scales depends on how well they can predict the outcomes of interest. Very often, the principal outcome or dependent variable is human/system performance, but we may also be interested in predicting physiological responses, especially if these are thought to be related to health problems. We may even be interested in predicting purely subjective measures such as user acceptance regardless of whether these are correlated with performance or physiological measures. In areas such as *comfort* and *annoyance*, and perhaps *workload*, subjective measures may have their own intrinsic value within the context of the specific investigation.

The overall impression given by published studies using subjective scales is that it is easy to underestimate the amount of effort required to establish both construct and predictive validity. There are few scales so well established that both kinds of validity are unquestionable. It may well be that in the rapidly changing technology of the modern world it is just not feasible to provide psychometric standards which will fit all possible circumstances. Many, perhaps the majority, of ergonomics studies concern the evaluation of specific equipments, tasks and working environments within an iterative design process. To be useful in this context subjective measures are not required to reflect absolute standards. The judgement that this design is preferable to some other may be all that is needed, or indeed all that is possible. It can, however, be useful to discover by the use of multivariate analyses which are the most significant dimensions in judgements of subjective attributes, as has been done to varying degrees for some ergonomics constructs.

To conclude, the choice of measures, whether subjective or objective, always has to be justified in terms of the specific aims of the investigation. Physical measures are not to be preferred purely on the grounds that they are 'objective'. The choice of any measure must be justified in terms of the purpose of the investigation. The supposedly unscientific nature of measures based on human perception and

judgement is no bar to the use of mathematical modelling techniques in many branches of ergonomics enquiry. At the same time, the use of such measures to define design standards is premature. Predictive validity is quite different from construct validity and normative data are rarely adequate for the former purpose. The complexity of the task of the ergonomist in making these choices means that the practice of this particular branch of science is an art!

Acknowledgements

I am grateful to Neville Stanton and several anonymous referees for helpful comments and suggestions, to my colleague Greg Hunt for a discussion of contemporary measurement theory and to all those who answered my appeals for information on various rating scales.

References

- ÅHSBERG, E. 1998, *Perceived Fatigue Related to Work* (Stockholm: Arbetslivsinstitutet).
- ÅHSBERG, E., GAMBERALE, F. and KJELLBERG, A. 1997, Perceived quality of fatigue during different occupational tasks: development of a questionnaire, *International Journal of Industrial Ergonomics*, **20**, 121–135.
- ANNETT, J., CUNNINGHAM, D. and MATHIAS-JONES, P. 2000, A method for measuring team skills, *Ergonomics*, **43**, 1076–1094.
- BADDELEY, A. D. and HITCH, G. 1974, Working memory, in G. Bower (ed.), *The Psychology of Learning and Motivation*, Vol. VIII (New York: Academic Press), 47–90.
- BARTLETT, F. C. 1943, Fatigue following highly skilled work, *Proceedings of the Royal Society, Series B*, **131**, 247–257.
- BORG, G. 1998, *Borg's Perceived Exertion and Pain Scales* (Champaign: Human Kinetics).
- BORG, I. and GROENEN, P. 1997, *Modern Multi-dimensional Scaling: Theory and Applications* (New York: Springer).
- BORING, E. G. 1957, *A History of Experimental Psychology*, 2nd edn (New York: Appleton-Century-Crofts).
- BROOKE, J. 1996, SUS: a 'quick and dirty' usability scale, in P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland (eds), *Usability Evaluation in Industry* (London: Taylor & Francis), 189–194.
- BUTTERWORTH, B. 1999, *The Mathematical Brain* (London: Macmillan).
- CHARLTON, S. G. 1996, Questionnaire techniques for test and evaluation, in T. G. O'Brien and S. G. Charlton (eds), *Handbook of Human Factors Testing and Evaluation* (Mahwah: Erlbaum), 81–99.
- CIPOLOTTI, L., BUTTERWORTH, B. and DENES, G. 1991, A specific deficit for numbers in a case of dense acalculia, *Brain*, **114**, 2619–2637.
- COLLE, H. A. and REID, G. B. 1998, Context effects in subjective mental workload ratings, *Human Factors*, **40**, 591–600.
- COOPER, G. E. and HARPER, R. P. 1969, *The Use of Pilot Ratings in the Evaluation of Aircraft Handling Qualities*. NASA Technical Note TND-5153 (Moffett Field: NASA Ames Research Center).
- COX, T. 1978, *Stress* (London: Macmillan).
- DAWES, R. M. 1972, *Fundamentals of Attitude Measurement* (New York: Wiley).
- DERRICK, W. L. 1988, Dimensions of operator workload, *Human Factors*, **30**, 95–110.
- DERRICK, W. L. and WICKENS, C. D. 1984, *A Multiple Processing Resource Explanation of the Subjective Dimensions of Operator*. Technical Report EPL-84-2/ONR-84-1 (Champaign: Engineering Psychology Laboratory, University of Illinois).
- DWYER, D. J., FOWLKES, J. E., OSER, R. L. and LANE, N. E. 1997, Team performance measurement in distributed environments: the TARGETs methodology, in M. T. Brannick, E. Salas and C. Prince (eds), *Team Performance Assessment and Measurement* (Mahwah: Erlbaum), 137–153.

- ENDSLEY, M. 1996, Situation awareness measurement in test and evaluation, in T. G. O'Brien and S. G. Charlton (eds), *Handbook of Human Factors Testing and Evaluation* (Mahwah: Erlbaum), 159–180.
- EYSENCK, H. J. 1967, *The Biological Basis of Personality* (Springfield: Thomas).
- FECHNER, G. T. 1860, *Elemente der Psychophysik* (Leipzig: Breitkopf & Hartel).
- GRIFFIN, M. J., PARSONS, K. C. and WHITHAM, E. M. 1982, Vibration and comfort: IV. Application of experimental results, *Ergonomics*, **25**, 721–739.
- HANKINS, T. C. and WILSON, G. F. 1998, A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight, *Aviation, Space and Environmental Medicine*, **69**, 360–367.
- HART, S. G. and STAVELAND, L. E. 1988, Development of the NASA-TLX (Task Load Index): results of empirical and theoretical research, in P. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: Elsevier), 139–183.
- HEARNSHAW, L. S. 1964, *A Short History of British Psychology, 1840–1940* (London: Methuen).
- HELLIER, E., EDWORTHY, J and DENNIS, I. 1995, A comparison of different techniques for scaling perceived urgency, *Ergonomics*, **38**, 659–670.
- HOLDING, D. H. 1983, Fatigue, in G. J. R. Hockey (ed.), *Stress and Fatigue in Human Performance* (London: Wiley), 145–167.
- HULZEBOSCH, R. and JAMESON, A. 1996, FACE: a rapid method for evaluation of user interfaces, in P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland (eds), *Usability Evaluation in Industry* (London: Taylor & Francis), 195–204.
- JISSELSTEIJN, W. A., DE RIDDER, H. and FREEMAN, J. 1999, Measuring presence: an overview of assessment methodologies, in G. W. M. Rauterberg (ed.), *IPO Annual Progress Report 34* (Eindhoven: University of Technology), 37–47.
- ISO 10075, 1991, *Ergonomic Principles Related to Mental Workload — General Terms and Definitions* (Geneva: International Standards Organization).
- ISO 10075-2, 1996, *Ergonomic Principles Related to Mental Workload — Part 2: Design Principles* (Geneva: International Standards Organization).
- JAMES, W. 1890, *Principles of Psychology* (New York: Holt).
- KELLY, C. J. and GOILLAU, P. J. 1996, *Cognitive Aspects of ATC: Experience from the CAER and PHARE Simulations*. Report of the ATC Systems Group (Malvern: Defence Research Agency).
- KELLY, C. J., GOILLOU, P. J., FINCH, W. and VARELLAS, M. 1995, *CAER Future Systems 1. (FSI) Final Trial Report. DRA/LS(LSC4)/CTR/RPT/CD246/1.0*.
- KLINE, P. 1998, *The New Psychometrics: Science, Psychology and Measurement* (London: Routledge).
- KLEIN-TEESELINK, G., SIEPE, H. and DE PIJPER, J. R. 1999, Log files for testing usability, in G. W. M. Rauterberg (ed.), *IPO Annual Progress Report 34* (Eindhoven: University of Technology), 63–71.
- LAZARUS, R. S. 1982, Thoughts on the relations between emotion and cognition, *American Psychologist*, **37**, 1019–1024.
- LINDH, C. and GÄRDER, P. 1993, The use of subjective rating in deciding RTI success, in A. M. Parkes and S. Franzén (eds), *Driving Future Vehicles* (London: Taylor & Francis), 391–400.
- LUCE, R. D. and TUKEY, J. W. 1964, Simultaneous conjoint measurement: a new type of fundamental measurement, *Journal of Mathematical Psychology*, **1**, 1–27.
- MICHELL, J. 1997, Quantitative science and the definition of measurement in psychology, *British Journal of Psychology*, **88**, 355–383.
- MITAL, A. and GOVINDARAJU, M. 1999, Is it possible to have a single measure for all work? *International Journal of Industrial Engineering — Theory, Applications and Practice*, **6**, 190–195.
- MORGAN, C. T., COOK, J. S., CHAPANIS, A. and LUND, M. 1963, *Human Engineering Guide to Equipment Design* (New York: McGraw-Hill).
- MUCKLER, F. and SEVEN, S. 1992, Selecting performance measures: 'objective' versus 'subjective' measurement, *Human Factors*, **34**, 441–455.
- MUSCIO, B. 1921, Is a test of fatigue possible? *British Journal of Psychology*, **12**, 31–46.

- NISBETT, R. E. and WILSON, T. 1977, Telling more than we can know: verbal reports as mental processes, *Psychological Review*, **84**, 231–259.
- NYGREN, T. E. 1991, Psychometric properties of subjective workload measurement techniques — implications for their use in the assessment of perceived mental workload, *Human Factors*, **33**, 17–33.
- PORTEOUS, M., KIRAKOWSKI, J. and CORBETT, M. 1993, *SUMI Handbook* (Cork: Human Factors Research Group, University College Cork).
- REED, C. B. and NYGREN, T. E. 1988, The subjective mental workload assessment technique: a scaling procedure for measuring mental workload, in P. A. Hancock and N. Meshkati (eds), *Human Mental Workload* (Amsterdam: North Holland), 185–218.
- ROHLES, F. H. and KONZ, S. A. 1987, Climate, in G. Salvendy (ed.), *Handbook of Human Factors* (New York: Wiley), 696–707.
- SANDERS, A. 1979, Some remarks on mental load, in N. Moray (ed.), *Mental Workload: Theory and Measurement* (New York: Plenum), 41–77.
- SCHACHTER, S. 1964, Interaction of cognitive and physiological determinants of emotional state, in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Vol. 1 (New York: Academic Press), 49–79.
- SCHACHTER, S. and SINGER, J. E. 1962, Cognitive, social and physiological determinants of emotional state, *Psychological Review*, **69**, 379–399.
- SHACKEL, B., CHIDSEY, K. and SHIPLEY, P. 1969, The assessment of chair comfort, *Ergonomics*, **12**, 269–306.
- SHERIDAN, T. B. and STASSEN, H. G. 1979, Definitions, models and measures of human workload, in N. Moray (ed.), *Mental Workload: Its Theory and Measurement* (New York: Plenum).
- SHIFFRIN, R. M. and SCHNEIDER, W. 1977, Controlled and automatic human information processing: II Perceptual learning, automatic attending and a general theory, *Psychological Review*, **84**, 127–190.
- SINCLAIR, M. A. 1995, Subjective assessment, in J. R. Wilson and E. N. Corlett (eds), *Evaluation of Human Work: A Practical Ergonomics Methodology*, 2nd edn (London: Taylor & Francis), 69–100.
- SMITH-JENTSCH, K. A., JOHNSTON, J. H. and PAYNE, S. C. 1998, Measuring team related expertise in complex environments, in J. A. Cannon-Bowers and E. Salas (eds), *Making Decisions Under Stress* (Washington, DC: American Psychological Association), 61–87.
- STANTON, N. A. and YOUNG, M. S. 1999, *A Guide to Methodology in Ergonomics* (London: Taylor & Francis).
- STEVENS, S. S. 1951, Mathematics, measurement and psychophysics, in S. S. Stevens (ed.), *Handbook of Experimental Psychology* (New York: Wiley), 1–49.
- STEVENS, S. S. 1956, The direct estimation of sensory magnitudes — loudness, *American Journal of Psychology*, **69**, 1–25.
- STEVENS, S. S. 1971, Issues in psychophysical measurement, *Psychological Review*, **78**, 426–450.
- TATTERSALL, A. J. and FOORD, P. S. 1996, An experimental evaluation of instantaneous self-assessment as a measure of workload, *Ergonomics*, **39**, 740–748.
- TERMAN, L. M. and MERRILL, M. A. 1937, *Measuring Intelligence* (Boston: Houghton-Mifflin).
- UNDERWOOD, B. J. 1966, *Experimental Psychology*, 2nd edn (New York: Appleton-Century-Crofts).
- WEBER, E. H. 1830, *The Sense of Touch*, trans. H. E. Ross (London: Academic Press, 1972).
- WHITAKER, D. and MARSH, D. 1997, Programme for Harmonised Air Traffic Management. PD/1 Final Report, PHARE/NATS/PDI 10.2/SSR;1.1.
- WHITESIDE, J., BENNETT, J. and HOLTZBLATT, K. 1988, Usability engineering: our experience and evolution, in M. Helander (ed.), *Handbook of Human-Computer Interaction* (Amsterdam: North Holland), 791–817.
- WIERWILLE, W. W., RAHIMI, M. and CASALI, J. G. 1985, Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity, *Human Factors*, **27**, 489–502.
- WOODWORTH, R. S. 1938, *Experimental Psychology* (London: Methuen).
- YEH, Y. Y. and WICKENS, C. D. 1988, The dissociation of subjective measures of mental workload and performance, *Human Factors*, **30**, 111–120.

Copyright of Ergonomics is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.