Characteristics of measures

The measures that can be used for the assessment of mental workload have different properties. The properties range from very general aspects to very specific. A general aspect is, for instance, the amount of equipment that is needed. A more specific, and from a scientific perspective more important property is the validity of a measure. Is the measure reflecting the concept of mental workload as intended, or is it reflecting other concepts, e.g., physical workload? O’Donnell & Eggemeier (1986) categorize the criteria for the selection of a workload-assessment technique on the basis of the following properties of the technique: sensitivity, diagnosticity, primary-task intrusion, implementation requirements and operator acceptance.

Sensitivity

Is the technique able to reflect changes in workload? Sensitivity of a measure should be defined within region of performance. In the previous chapter a model that described the relation between workload, performance and demand was presented. A primary-task performance measure cannot possibly be sensitive to mental workload in region C or A, simply because in the region’s definition included no change in performance. However, in the D and B regions changes in performance do reflect changes in workload. It is also likely that an operator is quite capable of indicating overload when demands are in the C region, and therefore a self-report measure’s sensitivity can easily be different from performance measures per region. Evaluation of measures should therefore always be linked to the region of performance.

Diagnosticity

How capable is the measure in reflecting demands on specific resources? Diagnosticity is the ability to discern the type or cause of workload, or the ability to attribute it to an aspect or aspects of the operator’s task (Wierwille & Eggemeier, 1993). A measure is said to be diagnostic within the context of the multiple-resource theory (Wickens, 1984) if it is sensitive to specific resource demands and not to others. Measures can be highly diagnostic and reflect a variation at a certain stage or on a certain locus of demand or they can be low on diagnosticity and reflect general demands. Pupil diameter is an example of a measure that reflects general demands and is low in diagnosticity. Pupil diameter is equally responsive to manipulations of different stages, such as response load or encoding and central processing load (Beatty, 1982). It is not sensitive to a specific type of resource expenditure. Other measures, e.g. some of the secondary task measures, are highly diagnostic. An example of a highly diagnostic measure is the evoked brain potential. The amplitude of the so-called P_{300}-component
of the evoked brain potential is sensitive to perceptual/central demands of a primary task (Gopher & Donchin, 1986). The choice for a diagnostic measure depends upon the measurement objective. If a general workload level has to be established, diagnosticity is not the most important selection criterium. If, however, the source of workload has to be traced a diagnostic measure can prove to be very useful and may guide to solutions of high workload demand.

Primary-task intrusion

The degree to which a technique degrades ordinary or primary-task performance is called primary-task intrusion. The disruption in ongoing task performance as a result of the application of the measurement technique is an undesirable property and should be minimized. Secondary-task techniques probably have the largest degrading effect on the primary task (Eggemeier et al., 1991). In particular the addition of an artificial secondary task may contaminate performance on the primary task. Self-report measures taken after completion of the task and most physiological measures seem to degrade primary-task performance the least.

Implementation requirements

Implementation requirements refer to practical constraints, such as the requirement of specific equipment or operator training. In field studies in particular, implementation requirements can become important. For example, the amount of equipment that is needed to measure eye movements may limit its use to laboratory settings (e.g. Unema, 1995). The same applies to the conditions in which some of the low-amplitude signal physiological measures can be properly assessed. Too much equipment might even result in primary-task intrusion.

Sometimes, in order to reach a stable or a reasonable performance level, subjects have to be trained extensively. In particular the requirement to obtain a reasonable dual-task performance can necessitate training. This is not necessarily a problem, but it does affect the time required before measures can be taken.

Operator acceptance

The degree of approval of the technique by the operator is referred to as operator acceptance. The operator’s opinion about a measurement technique, especially the use of self-reports, largely affects the correctness and accuracy of the measure. In general acceptance is higher if the technique is less intrusive or artificial, while the face validity of specific measurements may enhance operator acceptance. If the use and usefulness of some of the measures is not clear to the operator, explanation about the measures’ use is worthwhile and can help the operator to accept them (O’Donnell & Eggemeier, 1986). If the secondary-task technique is employed, operator acceptance is of primary importance. Acceptance can be enhanced by trying to let
the secondary task resemble activities that occur in the normal course of the operator’s performance. In pilot performance, for example, activities like radio communication could be used (O’Donnell & Eggemeier, 1986)

Sensitivity, diagnosticity and primary-task intrusion are of major importance, while the latter two criteria, implementation requirements and acceptance, should be considered additional selection criteria. Some authors propose a slightly different categorization of criteria. Wickens (1992) added ‘Selectivity’ and ‘Bandwidth & Reliability’ to the list:

**Selectivity**

Selectivity is the selective sensitivity to mental workload and not to changes in such factors as physical load. Selectivity denotes the validity of the measure for workload assessment. A measure can be sensitive to mental workload only or be sensitive to other factors as well, in particular to physical load. If the measure is also sensitive to other factors, this may or may not be a reason to discard it for mental load measurement purposes, depending upon task and test environment. For instance, a measure that is sensitive to both physical and mental workload can be used as mental workload indicator when no physical effort is required.

**Bandwidth and reliability**

Bandwidth and reliability refer to the workload’s estimate that has to be reliable both within and across tests. Stability of a measure between tests is what Wierwille & Eggemeier (1993) call ‘transferability’. Measures that were developed in a laboratory setting do not have to indicate workload equally well in the field. Between applications much will depend upon the region of task performance. A measure sensitive to low levels of workload only will not be able to discriminate between levels within high demand situations. Comparison of results obtained in the different environment with samples taken from the same population should give a good estimate of reliability.

**Interdependence characteristics**

The above described characteristics are not independent of each other. A highly diagnostic measure is only sensitive to variations in workload in specific computational processes. Therefore diagnosticity restricts sensitivity. Also, diagnosticity presupposes selectivity. An other interrelation exists between bandwidth and sensitivity. Bandwidth is no more than the definition of the restricted area of test environments in which a measure is sensitive. Interaction between characteristics can be expected to be particularly high in the case of secondary tasks. For example, a diagnostic measure that is sensitive to secondary-task
performance can only be reliable if primary-task intrusion of the secondary task is low.

The most desirable characteristics of measures of mental workload are high sensitivity, preferably in a wide bandwidth, high reliability and low primary task intrusion. Diagnosticity can also be of major importance, in particular, if a certain stage of information processing is suspected to be affected.

The different measures and their characteristics will be discussed individually in the next chapter. Three groups of measures can be distinguished: self-reports of mental workload, task performance parameters and physiological indices. Overall experience with the measure’s characteristics in the laboratory and field experiments will first be discussed, while in chapter 5 the range is narrowed to the applied domain of traffic research.