한국감성과학회지, Vol. 5, No. 4, pp. 11-32. December 2002

지상 교통에 있어서 운전자 상태의 주관적 척도: 비판적 고찰 및 응용을 위한 제언

Subjective Measures of Operator Status in Surface Transportation: A Critical Review and Recommendations for Application

Heidi D. Howarth*, Young-Woo Sohn**

Abstract : This article evaluates the existing subjective measures that have been utilized in surface transportation to assess various aspects of operator status such as fatigue, sleepiness, arousal, mood, etc. Specifically, the representative six subjective instruments - Epworth Sleepiness Scale, Karolinska Sleepiness Scale, Pearson and Byars Fatigue Checklist, Stanford Sleepiness Scale, Stress-Arousal Checklist, and NPRU Mood Scale - are compared and contrasted in terms of reliability,. validity, sensitivity, and appropriateness for application. Recommendations for application of the subjective measures in surface transportation are discussed.
Key words : Subjective measures, Fatigue, Sleepiness, Transportation

요 약 : 이 논문은 피로, 졸음, 각성, 기분 등과 같은 운전자 상태의 다양한 측면을 측정하기 위해 지상교통 분야에서 사용되어온 주관적 척도들을 평가한다. 구체적으로, 대표적인 주관적 측정 방법인 Epworth Sleepiness Scale, Karolinska Sleepiness Scale, Pearson and Byars Fatigue Checklist, Stanford Sleepiness Scale, Stress-Arousal Checklist 와 NPRU Mood Scale 등이 신뢰성, 타당성, 민감성과 응용을 위한 적절성의 측면에서 비교 및 대조된다. 결론에서 이러한 주관적 척도들을 지상교통 분야에 응용하기 위한 제언이 논의된다.
주요어 : 주관적 척도, 피로, 졸음, 교통

1. Introduction

1.1 Operator Status Defined

Researchers have spent years studying and debating the relative differences and similarities between human states, such as "fatigue", "tiredness", "sleepiness", and "drowsiness." Though some have used these terms interchangeably, most researchers will probably agree that they do not all mean the same thing, and indeed may even interact with each other. Related to these states are assessments of "mood." Reviewing the notion of what mood means, Nowlis (1965) suggests that it is comprised of factors that occur as states and vary temporally, allowing for the understanding, prediction, and control of behavior. Accordingly, mood may be conceptualized as a multi-faceted construct. Relevant to the current topic, and central to the most well-known mood scales with published psychometric data are the factors of Fatigue-Inertia, Vigor-Activity, Tension-Anxiety, Stress, and Arousal (Horne, 1991; Mackay, Cox, Burrows, & Lazzerini, 1978). The relationship between various measures of mood and measures of states, such as "sleepiness," has been examined in various research efforts (e.g., Angus & Heslegrave, 1985; Carskadon,

* University of Connecticut
** 연세대학교 문과대학 심리학과 조교수, Tel : 02)2123-2444, Fax : 02)365-4354, E-mail : ysohn@yonsei.ac.kr

1979; Herscovitch & Broughton, 1981; How et al., 1994).

This article does not attempt to resolve the definitional issues discussed above. Instead, since there are no clear boundaries between each of the aforementioned states, it utilizes a more general, all-encompassing term: operator status. Operator status is used to refer to the "state" of an operator as it pertains to his/her capacity for work, as influenced by job factors, such as scheduling, the time and duration of work, and the task itself. The idea behind the development of the term "operator status" is to provide a useful means of referring generally to the human states that others combine, confuse, and misuse, without falling into those same traps. In addition, the use of a broader definition allows for the review of a wider scope of the literature than would a more narrow definition.

1.2 Factors that Impair Operator Status

As noted above, there are many factors that may impact operator status, both alone and in conjunction with one another. They include issues such as irregular work schedules, tight work schedules, long work periods, sleep debt, the timing of work, and the task of vehicle/vessel operation itself (Dalziel & Job, 1997; Hartley, Arnold, Smythe, & Hansen, 1994; Horne & Reyner, 1995; McDonald, 1989; Wiley, Shultz, Miller, Mitler, & Mackie, 1996). "On-call" bus drivers provide a good example of surface transportation operators faced with multiple job-related factors that could impact their status, not to mention their safety and the safety of others. These operators work irregular schedules, often during non-routine hours (possibly extended due to overtime) and are, additionally, expected to adhere to strict time schedules. Moreover, they may also be exposed to environmental factors, such as excessive noise, vibration, bad weather, or heavy traffic (Evans, 1994).

1.3 Assessing Operator Status

Clearly, operator status is a complex and multi-faceted construct. It includes aspects of the quality and intensity or degree of an individual's experience and may be experienced or exhibited either psychologically or physiologically. In the classic fatigue literature, Bartley and Chute (1947) argued that only the psychological aspects of this phenomenon should be considered, because they are always directly experienced by the individual, while physiological impairment is not. This is deemed a rather extreme position today, however, and is not accepted by most researchers (Brown, 1994). Nevertheless, it is indeed likely that the majority of operators are faced with physical demands that are well within their capabilities and do not affect their ability to maintain safe work practices. The true challenge lies psychologically, not physically, in the demand for sustained attention during an often monotonous (i.e., overlearned) task . Though there are undeniably both physical and psychological components to operator status, the current literature has tended to focus less upon measuring decrements that are physical (i.e., muscular) in nature, and more upon those that are psychological.

1.4 Surface Transportation Defined

It is important to clarify for the purposes of this article, that the term surface transportation will be considered to encompass both land and water transport modes. This means that the literature that was surveyed included samples of operators of commercial motor vehicles (CMV; heavy trucks and motorcoaches), light rail operators, train engineers, and maritime crews. While school busses are also

technically considered CMVs, these drivers were not considered, as they work regular day schedules in jobs that are markedly different in scope, compared to other CMV operators. As a special case of operators, car drivers were, however, considered where appropriate. Although these "lay" operators do not drive for a living and therefore are not subject to all of the work-related factors that professional operators face on a chronic basis, there is a fair amount of research that focuses on some of the same issues, such as the effects of driving for long distances or during circadian downtimes.

1.5 Methodologies for Measurement

Two general methodologies for subjective measurement have developed over time. These are psychophysical techniques and rating scales. The most common psychophysical techniques for ratio scaling include the estimation and production methods. Estimation methods involve either estimating the relative magnitude of various stimuli by assigning numerical values to a series of comparative stimuli in reference to a standard stimulus of a particular numerical magnitude, or estimating the percentage magnitude of comparative stimuli relative to the standard (Kinsman & Weiser, 1976). Production methods, on the other hand, start with a standard stimulus of given intensity and require the production of a stimulus that is either a multiple or specified magnitude of the original (Kinsman & Weiser, 1976).

Within the field of psychology, estimation methods most often take the form of visual analog scales (VAS). The technique itself is quite simple and involves a horizontal line, anchored at each end using terms that are presumed to represent the extremes of the state of interest (e.g., Lee, Hicks, & Nino-Murcia, 1991; Pivik, 1991). Individuals are instructed to regard the line as a range of feelings along the given dimension, and to place a mark along it that is representative of how they feel at the moment. Responses are quantified in arbitrary units as the distance between the left end of the line and the participant's mark. While various researchers have constructed VASs for use in their studies, it appears that no standardized forms of VASs have been employed within the surface transportation literature to date.

The focus of this article is on the use of rating scales to assess subjective aspects of states (e.g., operator status). There are generally three types of rating scales: nondimensional, single-point measures, unidimensional rating scales, and multidimensional rating scales. According to Kinsman and Weiser (1976), nondimensional, single-point measures are the simplest means whereby an individual's state may be assessed. These measures typically consist of a single verbal report during work regarding a level of "tiredness" or undifferentiated "fatigue." The literature using such measures appears to have both originated and ceased during the 1950's, however, and does not exist within surface transportation. Kinsman and Weiser (1976) provide two reasons for this. First, since it is a nondimensional, single-point measure, the state in question is purportedly being experienced as an "all-or-none" event. Both conceptually, and psychometrically, this idea is difficult to justify. Experience alone suggests that subjective states vary in intensity, and logically, expressions of intensity increase quantitatively in some way the longer one performs a task. Second, from a psychometric standpoint, single-point measures suffer from much lower reliability than multi-point, dimensional measures of subjective status.

Unidimensional rating scales differ from

nondimensional scales, in that there are levels to the assessments of state. Judgements are made where an individual chooses a single point on a scale to represent the level that they are experiencing of the state in question. Because unidimensional scales allow for the quantitative measurement of levels of a particular variable, and thus produce a range of scores, the result is an increase in reliability over that for nondimensional scales (Kinsman & Weiser, 1976). There are a number of unidimensional rating scales that have been used successfully in the surface transportation literature. These include the Epworth Sleepiness Scale, the Karolinska Sleepiness Scale and the Pearson and Byars Fatigue Checklist, each of which will be detailed in the subsequent section.

The above unidimensional rating scales have undoubtedly provided researchers with valuable insights into the subjective measurement of operator status in surface transportation. However, as Bartley and Chute (1947) noted, the subjective assessment of state is likely more complex than a single (subjective) quality is capable of capturing. For this reason, multidimensional scales that possess a number of subjective qualities with differing levels have been developed as an additional and more intricate means of measurement. The multidimensional scales that have appeared within the surface transportation literature include the Stanford Sleepiness Scale, the Stress-Arousal Checklist, and the NPRU Mood Scale.

2. Subjective Measures of Surface Transportation

2.1 Reliability and Validity of Subjective Measures in Surface Transportation

There are two basic properties of empirical measurement, as put forth and detailed in a paper by Carmines and Zeller (1980). These are reliability and validity. In the case of unidimensional and multidimensonal scales, reliability refers generally to the degree to which repeated questionnaire administration yields the same results. Establishing reliability is important for measurement instruments because without it, there is a greater chance that obtained results are due to random error. Random error is problematic because its effects are unsystematic, and therefore unpredictable. As the process of measurement itself produces random error, indicators will always include it to some degree, however ensuring sufficient reliability levels will minimize these effects. Nonrandom error, on the other hand, refers to a systematic bias that may exist in a measurement instrument. It is this type of error that lies at the heart of validity. Being more conceptual in nature than reliability, validity deals with the strength of the relationship between a scale and the construct it is designed to assess. As in the case of reliability, validity is a matter of degree and not an "all-or-none" property; the greater the nonrandom error in an instrument, the less indicators represent the theoretical concept that they are intended to, and the lower the validity.

To follow is a discussion of the measurement properties of each of the subjective instruments that has been utilized within the surface transportation domain to assess various aspects of operator status. The criteria for inclusion in this review were as follows: 1) the scale purported to measure a dimension of operator status, 2) some amount of information regarding the psychometric properties of the scale was available, and 3) at least one published study that used the scale was directly related to the task of an operator within surface transportation. Of the group of publications that met the above criteria for inclusion, research methodologies varied from simulator-based

laboratory investigations to driver field studies. Equally as divergent was the amount and level of detail in the descriptive and statistical information published for each measurement tool. A summary of the development of each measure is provided, including any available reliability and validity evidence in the subsequent sections.

2.1.1 Epworth Sleepiness Scale (ESS)

The Epworth Sleepiness Scale (Johns, 1992) is a fairly recently developed rating scale that has been tested on operators in surface transportation. This scale was developed as a simple means of subjectively measuring daytime sleepiness or sleep propensity in adults, by asking them to retrospectively characterize their behavior in a variety of common situations (Johns, 1992). It is intended to measure a general, persistent component of daytime sleepiness, independent of day-to-day and time-of-day fluctuations. The questionnaire is self-administered, and asks the individual to rate the chance that, over "recent times" he/she would have dozed in eight everyday situations, using a 0-3 scale (0 = "would never doze"; 3 = "high chance of dozing"). The ESS is scored as a sum over items, from 0-24, where higher scores represent greater levels of subjective sleepiness. The situations are the following:

(1) sitting and reading
(2) watching TV
(3) sitting, inactive in a public place
(4) as a passenger in a car for an hour without a break
(5) lying down to rest in the afternoon when circumstances permit
(6) sitting and talking to someone
(7) sitting quietly after a lunch without alcohol
(8) in a car, while stopped for a few minutes in the traffic.

As a part of the development of the ESS, two factor analyses were performed that independently confirmed the scale's unidimensionality (Johns, 1992). In further analyses, Johns (1992) assessed reliability levels using both the retest and internal consistency methods. For the retest method, 87 medical students were given the ESS on two occasions, five months apart. Their scores did not differ significantly from one administration to the next, as would be expected for a scale that measures a persistent component of sleepiness. In addition, 54 patients who suffered from obstructive sleep apnea syndrome (OSAS) and received medical treatment for their condition after the first ESS administration evidenced significantly lower scale scores the second time they completed the scale. As reductions in daytime sleepiness are to be expected after successful treatment for OSAS, the change in scores for the sleep-disordered participants was taken to indicate further support for the reliability of the ESS using the test-retest method of assessment.

Cronbach's alpha was also calculated for the ESS in order to evaluate the internal consistency of the measurement items. Again, two groups of participants were given the scale to complete: medical students and sleep-disordered patients. Alpha was calculated at .73 for the students and .88 for the patients. These results were believed to indicate a "reasonably high level of consistency" (Johns, 1992, p. 379) for the ESS, though the alpha for the student participants does not meet with the criterion (.80) that is outlined in Carmines and Zeller (1980). In addition to demonstrating reasonable levels of reliability for the ESS, there is also some initial evidence for construct validity that is provided by Johns (1992). In this paper, the ESS

was validated as a measure that is able to distinguish healthy, control subjects from various groups of sleep disordered patients suffering from ailments known to be associated with differing levels of daytime sleepiness. For example, evidence for construct validity was shown in sleep-disordered patients who underwent polysomnographic recordings of sleep latency (SL; the amount of time until sleep as measured using EEG). A significant negative correlation with the ESS was evidenced, as one would expect for individuals with illnesses associated with excessive daytime sleepiness. Also, patients who suffered from chronic insomnia, thus a low propensity for sleep, reported significantly lower levels of sleepiness than did the controls.

While informative, the above information does little to suggest that the ESS is appropriate for use on operators within the surface transportation industry. Indeed there are a number of studies in this literature that report use of the ESS in various ways and also help to elucidate its measurement properties. For instance, Philip et al. (1997) surveyed car drivers at rest areas, and additionally used a control group of non-drivers who were demographically similar and tested under the same conditions. As part of a sleep/wake diary that was used to gauge driving behavior and sleep habits over the previous year, scores on the ESS were reported to be "low" for both drivers and controls (Philip et al., 1997, p.387). ESS scores were also correlated with SL measures obtained during two daytime naps that were part of an objective assessment of daytime sleepiness. For the purposes of construct validation, a negative relationship between these measures would be expected, where lower scores on the ESS would be reflected in longer SL measurements, though a nonsignificant correlation was actually found. No explanation for this was provided, however, one might speculate that employing the ESS as a way to estimate levels of sleepiness over an entire year goes beyond the scale's intended timeframe (i.e., "recent times"). Moreover, a year is likely too long a period for individuals to be able to accurately aggregate over, with the increased chance that the persistent components of sleepiness that the ESS is designed to measure may have changed.

2.1.2 Karolinska Sleepiness Scale (KSS)

The Karolinska Sleepiness Scale (Akerstedt & Gillberg, 1990) is a somewhat less known subjective rating scale within the surface transportation literature. Though statistical evidence for unidimensionality has not been published, conceptually, the scale does appear to measure a single dimension of "sleepiness" that is differentiated over nine levels. In the KSS, individuals rate their sleepiness level for a single item on a nine-point scale. The scale labels are as follows, where steps in between are assigned values, but not verbal labels:

(1) extremely alert
(3) alert
(5) neither alert nor sleepy
(7) sleepy - but no difficulty remaining awake
(9) extremely sleepy - fighting sleep.

Evidence for the reliability of the KSS does not appear to be available. In a paper on unidimensional scaling, McIver and Carmines (1981) suggest that single-item scales, such as the KSS, rarely provide sufficient information to allow for the estimation of their measurement properties, including levels of both reliability and validity. Indeed, due to the limited detail that is available in published form about the development of the KSS, it is difficult to even suggest that there is evidence

for the face validity of this instrument. As was proposed earlier, it seems generally that this kind of support is more often assumed than generated.

Nevertheless, there is some preliminary evidence for construct validation of the KSS. Using eight male participants, Akerstedt and Gillberg (1990) showed a significant relationship between signs of sleepiness as displayed in EEG and electrooculogram (EOG) measures and ratings on the KSS, though it was noted that "considerable" (p. 35) levels of subjective sleepiness had to be reported before this relationship appeared. Some caution must be exercised regarding the findings of this study, however, due to the small number of participants used. In another investigation, Gillberg, Kecklund, and Akerstedt (1994) sought to validate the KSS against measures of performance, using six participants over two nights in a laboratory setting. Under the assumption that subjective reports of increases in sleepiness are related to performance decrements on laboratory tasks, such as vigilance and RT, it was shown that KSS ratings immediately before these tests were indeed predicting of performance. Despite the limited amount of information available regarding the development of the KSS, results from laboratory research does provide some support for the construct validity associated with this scale.

2.1.3 Pearson and Byars Fatigue Checklist

Use of one of the two forms of the Pearson and Byars Fatigue Checklist (also known as the Feeling-Tone Checklist, Pearson, 1957; Person & Byars, 1956) is reported in a small number of studies of simulated driving. During this instrument' s development, Guttman scaling techniques were used to determine that each of the item sets likely constitutes a unidimensional scale (Pearson, 1957). The two checklist versions consist of equivalent forms (Form A and Form B) of a 13-item list of terms that was selected to describe a "fatigue continuum" (Pearson & Byars, 1956, p. 1). Form A includes the following items:

(1) slightly tired
(2) like I"m bursting with energy
(3) extremely tired
(4) quite fresh
(5) slightly pooped
(6) extremely peppy
(7) somewhat fresh
(8) petered out
(9) very refreshed
(10) ready to drop
(11) fairly well pooped
(12) very lively
(13) very tired.

In the original version(s) of the checklist, individuals are asked to decide whether they feel better or worse than each of the statements, using a 0-2 scale (0 = "worse than", 1 = "same as", 2 = "better than").

Despite the considerable research that was undertaken in order to develop the Pearson and Byars Fatigue Checklist, its infrequent use outside of its original intent is not surprising, given the distinctly colloquial content of the items . This may be partially explained by the fact that the scale originated in Texas in the late 1950's. It also seems probable that the time period of its development is the reason why the analyses used to evaluate the measurement properties of the checklist are rather atypical by current standards. Nevertheless, a brief explanation and evaluation of these methods will be attempted.

During scale development, the internal consistency of the measure was assessed using a

Chi square test on all original items. In their paper, McIver and Carmines (1981) point to the "criterion of internal consistency" (p. 24) method as a means of item analysis that was less time consuming than hand computations of interitem correlations, as were (at the time) required for the Cronbach's alpha statistic. However, their discussion clearly describes this technique for use as a way to determine the appropriateness of individual scale items for use, while Pearson employed the entire scale in his analysis. Additionally, a Chi square statistic was used to evaluate group mean differences instead of a t test, as prescribed in McIver and Carmines (1981). Though these discrepancies leave some doubt as to the interpretation of the analysis, Pearson suggested that a "definite" (p. 187) trend in the data resulted. Items that were significant for a first administration of the scale (A.M. data) tended to be from the positive end of the fatigue continuum, and items that were significant for a second administration of the scale (P.M. data) tended to be from the negative end of the continuum. This was stated to demonstrate that items tended to be significant when they fell "within that part of the [fatigue] continuum which seems to be 'functioning' at the moment" (p. 187).

In further statistical procedures, the above findings were used as the basis for the selection of items for parallel forms of the checklist. After the finalization of the two 13-item versions, a more conventional means of estimating reliability was undertaken (Pearson, 1957). The alternative-form method was employed using two groups of participants, where the order of test versions was counterbalanced. The correlation between the two forms of the checklist was .92 for one group of subjects and .95 for the other group. These results were deemed sufficient evidence for the reliability of both checklist versions; indeed the relationships are quite strong. This evidence must be questioned, however. While the alternative-form method of reliability estimation is heralded for the reason that using parallel forms of a scale protects against inflated correlations due to item recall, this unfortunately did not apply for the current study. For reasons that are not elucidated by the authors, the data that they chose to correlate to estimate scale reliability were from the end of the experimental session and corresponded to the third administration of Form A and second administration of Form B of the checklist. It is impossible to know how much multiple completions of each version may have inflated the reliability of this scale as estimated using the alternative-form technique.

The validity of the Pearson and Byars Fatigue Checklist also comes under suspicion, based on an explanation of how scale items were judged valid based on the ability to discriminate between "fatigued" and "non-fatigued" (Pearson & Byars, 1956, p. 3) criterion groups. As previously detailed in this section, within the social sciences, criterion-related validity is the most difficult type of validity evidence to demonstrate because no consensus exists among researchers regarding the appropriate criteria against which to validate measures. With this in mind, it appears as though Pearson and Byars have oversimplified a complex phenomenon by suggesting that groups of "fatigued" and "non-fatigued" participants were formed based on testing that occurred both before and after a performance task called the Pursuit Test. However, in fairness, it should also be noted that the detail provided regarding this seemingly unconventional means of evidencing criterion-related validity was probably not sufficient to allow for a complete understanding or critique of the procedure.

There also exists a second validation study that was reported by Pearson (1957). In this effort, scores on each form of the checklist were evaluated over time, where an increase in subjective reports of fatigue was expected for an experimental group above and beyond a control group. The experimental group was given the Pursuit Test (believed to induce fatigue), while the controls were not. Results showed that both groups reported increases in subjective fatigue over time, however the increase for the experimental group over the control group was significant for both Form A and Form B of the checklist. As a caveat to this seemingly promising finding, however, it should be noted that there is no mention made within this study of controlling for time of day effects, which were likely present, as the protocol took almost 5 hours in total. Nevertheless, Pearson points to this result as further evidence for (construct) validity for the Pearson and Byars Fatigue Checklist.

2.1.4 Stanford Sleepiness Scale (SSS)

The Stanford Sleepiness Scale (Hoddes, 1973) is a well-known rating scale that has been used within the surface transportation industry. MacLean, Saskin, Fekken, and Knowles (1989) and others (Horne, 1991) suggest that the SSS is a multidimensional subjective scale measuring "sleepiness." A factor analysis performed by MacLean, et al. (1989), resulted in two major dimensions and a third weaker factor. In a paper by Horne (1991), it was suggested that the two major factors seem to relate to the dimensions of Vigor and Fatigue in the Profile of Mood States (McNair, Lorr, & Droppleman, 1981). The SSS is completed by choosing the one of the following statements that best describes the individual's state of sleepiness:

(1) feeling active and vital; alert; wide awake
(2) functioning at a high level, but not at peak; able to concentrate
(3) relaxed; awake; not at full alertness; responsive
(4) a little foggy; not at peak; let down
(5) fogginess; beginning to lose interest in remaining awake; slowed down
(6) sleepiness; prefer to be lying down; fighting sleep; woozy
(7) almost in reverie; sleep onset soon; lost struggle to remain awake.

During its development stages, a reliability estimate for ten subjects was reported as a correlation of .88 with an alternate form of the scale (Hoddes, Dement, & Zarcone, 1972). One must assume the integrity of this analysis, as no further information regarding this alternate SSS version is available, and no mention is made regarding whether the correlation was significant, especially given the relatively small number of subjects in the sample.

In subsequent research, accounts of various methods used to evidence validity for the SSS are provided. Unfortunately, some of the earliest statistical evidence, as in the case of the Pearson and Byars Fatigue Checklist, is difficult to interpret. For example, in one study (Hoddes, 1973), validity is inferred based on positive a correlation between SSS ratings and performance scores, however, the direction of this relationship counters logic, which would predict that reports of greater sleepiness would be associated with poorer performance - a negative relationship.

Two other studies (Harnish, Chard, & Orr, 1996; Johnson, Freeman, Spinweber, & Gomez, 1988) present their findings more clearly, however, the evidence actually points against a relationship

between scores on the SSS and objective measures of sleepiness. In their paper, Harnish, Chard, and Orr (1996) reported a nonsignificant relationship between the time to the onset of sleep and corresponding SSS scores (administered immediately before beginning the SL measurements), leading to speculation that the SSS measures a different dimension of sleepiness than do measurements of SL. In the Johnson (1988) study, support was found for a relationship between the SSS and a VAS that was also designed to measure sleepiness, but no association could be substantiated between the subjective SSS and a measure of SL. Again in this study, it was concluded that subjective and objective measures of sleepiness probably do not measure the same dimensions of state.

While the above research does not paint a very clear picture of support for the validity of the SSS, there does exist some evidence that the SSS is a valid measure of sleepiness as a result of total or partial sleep deprivation. In a study by Herscovitch and Broughton (1981) the sleep of participants was restricted by approximately three hours per night for five nights before they were tested on the SSS every 15 minutes during waking times. A significant increase in SSS ratings was found after the nights of sleep restriction, with a return to baseline levels following a night of recovery sleep. Additionally, research performed by Carskadon (1979) investigated the effects of total sleep loss on SSS scores, and similar to the above results, showed that there was a significant increase in reported sleepiness during deprivation, with a return to baseline values upon recovery. These studies evidence construct validity for the SSS, in that a consistent pattern of findings was demonstrated with respect to a hypothesized theoretical relationship between sleep restriction and subjective reports of sleepiness. It is difficult to make any further assessments of the validity of the SSS (e.g., face validity), because details of how the scale was constructed do not appear to be published. Its popularity among researchers over the years since its introduction would suggest, however, that there is evidence for the face validity of the SSS.

In a recent study where the SSS was used within the surface transportation literature, Wiley, Shultz, Miller, Mitler, and Mackie (1996) provided mixed evidence for the construct validation of the SSS. In this field investigation, 80 qualified CMV drivers participated in one of four work schedule conditions: 1) 10-hr baseline daytime route for each of 5 days, 2) 10-hr rotating route (starting 3 hr earlier each day) for 5 days, 3) 13-hr nighttime start, each night for 4 nights, or 4) 13-hr daytime start for 4 days. Physiological, performance, and alertness (e.g., SSS - five administrations over 24 hr) measures were collected during driving and sleep, where appropriate. Findings that evidence construct validity for the SSS revealed a positive correlation between self-ratings of sleepiness and both the number of hours of driving during a trip and the total number of trips made. Especially encouraging is the fact that this relationship held true for all conditions. Moreover, since two of the conditions required daytime driving and two required nighttime driving, there is less likelihood that time-of-day effects were a factor. Results of correlational analyses between the SSS ratings and performance tests (e.g. a tracking task), however, demonstrated either nonsignificant or very small negative relationships. Depending on how much value one puts in correlations that are significant at very low levels, the association between SSS scores and performance test results may or may not signify additional evidence for the construct validity of the SSS. As the theoretical relationship between sleepiness and performance is rather tenuous and

often questioned, it is difficult to know where to draw the line when evaluating validity in this case.

2.1.5 Stress-Arousal Checklist (SACL)

The Stress-Arousal Checklist (Mackay et al., 1978) has been used in surface transportation research and is the most recent incarnation of Thayer's classic Activation-Deactivation Checklist (Thayer, 1967). The difference between the two scales is that the Mackay et al. (1978) version substituted out adjectives that were "too American" for words that were more appropriate for use in the UK. The end result is comprised of 34 mood adjectives. Factor analyses determined that these items measure two dimensions: stress and arousal. The arousal factor is relevant for the purposes of this paper and includes the following 15 adjectives: active, energetic, vigorous, alert, lively, activated, stimulated, aroused, drowsy, tired, idle, sluggish, sleepy, somnolent, and passive. Individuals are instructed to indicate, for each word, how it describes their feelings at that moment, using the following response scale:

(++) = definitely feel
(+) = feel slightly
(?) = do not understand or cannot decide
(-) = definitely do not feel.

Split-half reliability estimates for the SACL are cited in Watts, Cox, and Robson (1983), but detailed in an unavailable publication (Cox, Mackay, & Page, 1982). The arousal factor was reported to have an alpha value of .82, while the stress factor resulted in an alpha of .80. These values meet or exceed the criterion for acceptability that is stated in Carmines and Zeller (1980), thus evidencing support for the reliability of the SACL for use in the UK. Although the Mackay et al. (1978) instrument was later modified by Cruickshank (1984) as a result of a possible response bias, the existing surface transportation research uses the original version of the SACL.

Evidence for the construct validity of the arousal dimension of the SACL was presented in conjunction with the reliability information that is not currently available; thus it cannot be reviewed in detail. However, based on an abstract for this publication (Cox et al., 1982), it is known that the SACL was used to measure mood in a repetitive work situation (loading and sorting tasks) under three different durations (30, 60, and 120 min). Results that are relevant for the purposes of this article indicated that arousal levels decreased across work periods, and as a function of the duration of the work period. This finding is theoretically consistent with the notion that time on task is related to concomitant reductions in states, such as arousal.

Only one research effort within the surface transportation literature that uses the SACL was discovered. In this field study, Raggatt and Morrissey (1997) recruited volunteer long-distance bus drivers who worked rotating day (n = 5) and night (n = 5) shifts. Each shift was approximately 12 hours in duration, including a 30-minute rest break after four and eight hours of driving. In most cases, both before and after the shift, as well as at each of the two rest breaks, physiological measures were taken and the SACL was administered. Baseline data were subsequently collected at four times on the second of two days off after a driver was on duty. Results for arousal ratings showed a significant interaction between day (on-duty vs. rest) and time of measurement, such that both day and night shift reports of arousal were initially elevated for on-duty times and then, at the end of the shift, dropped lower than ratings on rest days.

The drop in arousal at the end of each shift was mirrored in measures of deactivation in heart rate. The correspondence between these measures was tentatively speculated to indicate fatigue onset after long hours at the wheel, but the authors were cautious with this suggestion, given the small sample size used and because other factors could have confounded the results.

Nevertheless, the above findings are encouraging, and suggest avenues for future research using the SACL, if not some evidence for the construct validity of the measure. Because this version of the SACL contains UK-specific item choices, however, a thorough investigation of the measurement properties of this instrument is warranted, especially for use in other cultures. Additionally, with regard to the face validity of this scale, the relatively small amount of available research using the SACL makes it difficult to suggest that there is such evidence.

2.1.6 NPRU Mood Scale (NPRU)

The NPRU Mood Scale (Lubin, Moses, Johnson, & Naitoh, 1974) was developed by the Navy as an instrument that could be easily, briefly, and repeatedly administered to measure cumulative sleep loss and performance decrements (Moses, Lubin, Maitoh, & Johnson, 1974). There is one known publication within the surface transportation domain that reports results for the NPRU. This scale was derived from an early version of the Profile of Mood States (McNair et al., 1981), a mood scale that has not been utilized in surface transportation. However, items were further modified and eliminated in an effort to target maximal sensitivity to sleep loss.

In what would eventually become known as the NPRU, the Lubin et al. (1974) team decided a priori on items that reflected either positive (decrease during sleep loss) or negative (increase during sleep loss) effects of total sleep restriction and then tested their predictions. Though the analysis is not reported statistically, it seems that these results may reflect a confirmatory factor analysis, supporting two factors within the NPRU. It was reported that 19 of the 21 original items were deemed to be positive, and 21 of the 31 items were deemed as negative. Additionally, the modified list of positive items was found to be a better measure of one night of sleep loss than any weighted combination of positive and negative terms.

Scoring the NPRU is therefore performed separately for the positive and negative subscales (scored over items). The positive subscale includes the following 19 items: active, alert, carefree, cheerful, able to concentrate, considerate, dependable, efficient, friendly, full of pep, good-natured, happy, kind, lively, pleasant, relaxed, satisfied, able to think clearly, and able to work hard. The negative subscale is comprised of the following 10 items: annoyed, defiant, drowsy, dull, grouchy, jittery, sleepy, sluggish, tense, and tired. Instructions for the NPRU ask individuals to chose the answer that best describes how they feel "now." The rating scale consists of four points (0 = "not at all", 1 = "a little", 2 = "quite a bit", 3 = "extremely").

In one laboratory study that reports on the development of the NPRU, Moses et al. restricted the sleep of 14 naval recruits for three nights and followed with a night of total sleep loss. Having administered the NPRU positive subscale only, the authors found that mood scores were significantly lower after total sleep loss. In another investigation, participants were kept awake in a laboratory for 54 hours performing cognitive and communications tasks ("work"), but were given regular breaks ("rest," Angus & Heslegrave, 1985). Each hour, the NPRU and scales that measured fatigue and sleepiness

were completed. Results showed an overall decrease in positive mood and increase in negative mood over the duration of the experiment. In line with the theoretical notion that sleepiness/fatigue and mood are related, sleepiness and fatigue reports both dropped significantly over the 54 hours. Concomitant with the aforementioned changes in mood were performance decrements on a number of common experimental measures. Additionally, after 18 hours, participants were found to report significantly lower positive mood and higher negative mood ratings during work, as opposed to rest sessions. The results of this and the above investigation suggest evidence for the construct validity of the NPRU with regard to expected changes in mood over time and declining performance. Additionally, support was demonstrated for the transient nature of mood in the fluctuations that occurred in ratings between work and rest periods. Unfortunately, however, time-of-day effects were not considered in either of these studies, so results must be interpreted with caution.

2.1.7 Conclusions Regarding Reliability and Validity

By and large, the scales discussed above have demonstrated only tentative evidence for measurement properties that are a vital means of ensuring meaningful empirical results. Especially with regard to the surface transportation literature, this makes it difficult to fully comprehend the applicability and relevance of these measures. It appears as though oftentimes researchers are short sighted, and may use an instrument simply because it has been reported in similar investigations, without knowing if it is truly appropriate, reliable, or valid. Indeed, it is a large task to fully investigate a measure a priori, or to perform the research necessary for evaluation, if such evidence is not available. However, without knowledge of the basic statistical foundations of a scale, the greater challenge lies in discovering meaningfulness within the existing research.

2.2 Sensitivity and Appropriateness of Subjective Measures in Surface Transportation

With regard to subjective instruments, even if evidence for reliability and validity has been established, a measure is of limited use if not also demonstrably sensitive and appropriate for application within a particular domain. Clearly, the terms "sensitive" and "appropriate" can be taken to have a variety of meanings; within the context of this article, they are further elucidated with reference to the pinnacle of application for empirical findings.

There are existing subjective measures that could be used on operators in real-world job situations, as a means of identifying their preparedness for work, or "fitness for duty" (FFD). FFD is not a new concept in transportation, and in fact has been in use for years in the form of performance tests that are administered before the onset of a shift, in order to identify whether an individual is qualified to work at that time.

As a means of determining if a subjective measure is appropriate for use in FFD detection, three requirements, as proposed by Hartley, Arnold, Smythe, and Hansen (1994) with regard to their own research, have been adapted: 1) measurement instruments must be portable and capable of being deployed in a moving vehicle/vessel or during brief rest stops, 2) measures must be non-invasive and allow the operator as much freedom as possible to carry out normal operations (navigation, loading, eating, sleeping), and 3) tests must be acceptable to operators. Unlike some test apparatus, it is easy to make a case for the portability of a rating scale,

though for safety reasons, completion would be most appropriate during brief, non-navigating intervals. Additionally, verbal administration could be considered as a future possibility. With regard to invasiveness, questionnaires would be considered very low, as it is doubtful that they would interfere with operator tasks. Finally, in contrast to the majority of measurement devices, which are often criticized for being cumbersome, invasive, and time consuming to administer, it is likely that a brief, easily completed rating scale would be accepted by operators, if not welcomed. Subjective measures, therefore, clearly meet with the requirements for appropriateness that are outlined by Hartley et al. (1994).

Unlike traditional FFD measures, however, in the case of operator status, which has been shown to diminish over time, it would be important to be able to test an operator repeatedly over the duration of work. In the surface transportation industry, operators may regularly be faced with being on duty during odd hours, in addition to irregular, on-call, or extended shifts. Therefore, further extending Hartley et al.'s (1994) criteria, it is necessary that a measurement instrument demonstrate sensitivity to operator decrements both over time and at different times of day. Ideally, a scale would also be sensitive to changes in task performance, but as discussed previously, the legitimacy of this relationship is as often questioned as it is supported, so requiring this type of evidence would not be appropriate.

With these terms in mind, this section will further consider the instruments reviewed previously with regard to their reliability and validity for their likely sensitivity to the detection of impaired operator status and whether they are appropriate for repeat administration. A brief explanation of the degree to which each scale meets these criteria will be provided as a means of determining which are the most applicable for consideration as FFD measurement instruments. In the interest of redundancy, however, the reliability and validity information previously presented will not be reiterated.

2.2.1 Epworth Sleepiness Scale (ESS)

As previously detailed, the ESS was designed to assess a general, persistent component of daytime sleepiness that is independent of day-to-day and time-of-day fluctuations. This can be seen in the scale instructions, which ask the individual about "recent times," as well as in the items, which are representative of general, everyday situations. As noted by Mitler and Miller (1995), the ESS is therefore not appropriate for use on various occasions throughout the day, or even from day to day. Furthermore, they propose that this scale is also not suitable for the assessment of the effects of short-term conditions, such as acute sleep loss, which is a likely problem for on-call operators or others who occasionally work long or sustained hours.

Empirically, there is some evidence from the surface transportation literature that suggests that the ESS may be sensitive to general trends in elevated sleepiness for those who work long hours at night versus individuals with schedules that are not as severe (Hakkanen & Summala, 2000). However, this small amount of support cannot be weighted very strongly in light of other research findings that were inconclusive (Maycock, 1997; Philip et al., 1997) and the ways in which the ESS fails to demonstrate appropriateness for repeated use. As an FFD measurement tool, the ESS unfortunately does not fulfill the relevant criteria and should not be considered for this type of application within surface transportation, unless

only infrequent assessments of daytime sleepiness are desired. Because this instrument has shown itself to be able to distinguish between individuals who suffer from sleep disorders and healthy control subjects, it seems that clinical applications, such as one-time assessments of chronic sleep disorders, are really the most appropriate use for this scale.

2.2.2 Karolinska Sleepiness Scale (KSS)

There is not a great deal of available research that has utilized the KSS in ways that contribute to the evaluation of its appropriateness and sensitivity. Being a single-point rating scale, the KSS is certainly an instrument that would lend itself to rapid assessments of FFD. For this same reason, however, there should be concern regarding factors such as overestimation due to memory, or reactivity. These issues were introduced in the prior discussion of problems that can accompany reliability testing involving two administrations of the same instrument, but also apply here.

Since the purpose behind FFD testing is to make judgements about whether an individual should be permitted to work, it is likely that operators would be especially concerned with maintaining appropriate and consistent levels of alertness on the job. Given a simple rating scale with few anchors, such as the KSS, it would seem quite easy for an operator to recall how he/she responded in a previous test administration, and to base future responses on that rating without regard to his/her current state. Whether this happens or not probably lies in the motivation of the individual, as Gillberg, Kecklund, and Akerstedt (1994) demonstrated in a laboratory experiment where the KSS was administered 12 times over one night and resulted in "marked" (p. 239) increases in ratings over that time. For participants undergoing testing in a laboratory setting, there is no likely reason why one would want to appear more alert than he/she is, in which case the KSS seems to perform quite well under conditions of repeat administration. Nevertheless, because using the KSS in the laboratory is vastly different from using it in the field, the current recommendation would be against repeat administration of this scale in an FFD context.

Despite the above caveat, there is some encouraging evidence with regard to the sensitivity of the KSS as reported in studies within surface transportation. Results from two investigations suggest that this instrument reflects time-of-day and time-on-task (i.e., during driving) changes in sleepiness (Gillberg et al., 1996; Kecklund & Akerstedt, 1993). Moreover, it may also be sensitive to fluctuations in objective measures, such as sleep latency and performance, over time. It seems that perhaps the KSS should not be dismissed altogether for use in the field; however, it is certainly the case that without further research, scale administration should probably be limited to one-time or otherwise infrequent assessments of operator status.

2.2.3 Pearson and Byars Fatigue Checklist

The Pearson and Byars Fatigue Checklist requests individuals to record their current feelings in comparison to a list of 13 items. It exists in parallel forms and has the added complexity of a three-point rating scale that responses to each item are summed over. For these reasons, it does not seem as likely that the checklist would be subject to individuals easily recalling their responses from one administration to the next. Despite this, the checklist is brief and simple to complete, as long as one is not hindered by its colloquial nature. Interestingly, while this issue seems rather obvious, Kinsman and Weiser's (1976) critique of the checklist items was not noted in the later surface

transportation research that used this scale. Nevertheless, the dated nature of the terms appears undeniable, and should be considered a legitimate obstacle to the use of this scale, as words that are not understood or recognized could impact score accuracy and, as a result, operator FFD assessments.

Some amount of evidence in the surface transportation literature for the sensitivity of the Pearson and Byars Fatigue Checklist to time-on-task effects is derived from the results of Nilsson et al. (1997). In this study, a reanalysis of the data yielded a linear increase in fatigue reports over time, until subjects were unable to continue their participation in the simulated driving task. This small amount of support for the sensitivity of the checklist is promising, but suggesting its use as an FFD detection instrument is probably premature. Research beyond the initial studies on scale development first needs to be conducted to establish the appropriateness of the items in this instrument. In addition, further investigation of the checklist's sensitivity to additional factors, such as time of day, is probably warranted before recommending its use in the field. The most encouraging aspect of the Pearson and Byars Fatigue Checklist currently is that it appears that, unlike the ESS and KSS, it has promise as an instrument that is suitable for repeat administration.

2.2.4 Stanford Sleepiness Scale (SSS)

The Stanford Sleepiness Scale is reported by Mitler and Miller (1995) to be an instrument that can be administered on multiple occasions per day, though this statement is not supported by specific research findings. However, speculation might conclude that this reference was to a laboratory study conducted by Carskadon (1979), where participants completed the SSS every 15 minutes over all waking periods (0800 - 2200) for six days. Scores were later averaged for each hour and resulted in significant increases in sleepiness during deprivation (two days) and a return to baseline values upon recovery (two days). Moreover, a significant time-of-day effect was discovered, while, more generally, sleepiness was shown to increase over the sleep deprivation days and decrease during the recovery days. Within the laboratory, at least, the performance of the SSS over multiple administrations seems to be well documented. As in the case of the KSS, however, the ease with which an individual might recall ratings from one FFD test to another must be considered for this scale, as it consists of only seven points. For this reason, despite the laboratory findings that suggest that the SSS is sensitive to multiple administrations, only infrequent use of this scale to assess operator status seems appropriate.

In addition to the laboratory, the SSS has shown itself to be sensitive to certain factors that would affect operators within the surface transportation industry, though study findings are not entirely conclusive. In one truck simulator study where participants drove for two 8-hr sessions, the SSS was sensitive to the effects of long periods of driving (i.e., time-on-task), with concomitant performance decrements, though time-of-day could not be ruled out as a factor (Ranney et al., 1999). Similarly, in a field investigation of truck drivers, the SSS was shown to reflect expected increases in sleepiness as the number of hours driving increased for each of four work schedule conditions; however, the sensitivity of the SSS to performance measures was only slight (Wiley et al., 1996). Finally, the SSS demonstrated sensitivity to total sleep deprivation in a naval study by How et al. (1994), where performance on tests designed to simulate work tasks also decreased.

Recommendations for use of the SSS as an FFD test are tentative. It is suspected that the simplistic nature of the scale, while attractive on those grounds, would also make ratings simple to recall if used on a repeated basis. Furthermore, mixed support was found for the sensitivity of the SSS to performance decrements; though, it should be kept in mind that the relationship between sleepiness and performance is not firmly understood. Until additional research is able to clarify existing findings and also whether the SSS is sensitive to other factors, such as time-of-day, an overall recommendation for use of this instrument in an FFD context would be for situations where only infrequent assessments are desired.

2.2.5 Stress-Arousal Checklist (SACL)

The SACL is similar to the Pearson and Byars Fatigue Checklist, in that ratings are made for a number of adjectives regarding one's feelings "at the moment" and then summed for an overall score. For this reason, it is unlikely that individuals would be able to duplicate their mood state or successfully falsify one on multiple occasions. Evidence regarding the appropriateness of repeat administration of the SACL is not explicitly stated, but may be inferred from a study by Raggatt and Morrissey (1997). In this investigation of long-distance bus drivers, the SACL was administered several times over the duration of rotating day and night shifts, and then again at the same times on the second of two days off. Results over multiple ratings showed that arousal was initially elevated for both day and night shifts, but at the end of the work period dropped lower than on days off. Although the authors cautioned that their sample was small and that other factors might have influenced the results, these findings should nevertheless be considered as preliminary evidence for the appropriateness of using the SACL for repeated administration.

Additionally, support for the sensitivity of the SACL to the decrements associated with long periods of driving is suggested by the results of the Raggatt and Morrissey (1997) study. However, since the previous investigation is the only known to exist within surface transportation, it may be helpful to consider the results of a study that used the SACL in a simulated repetitive work situation and also evidenced sensitivity to the effects of time-on-task (Cox et al., 1982). Results demonstrated a decrease in reported levels of arousal across work periods of 30, 60, and 120 minutes, and also as a function of the duration of the work period. Before the SACL should be considered for use as an FFD instrument, however, further research into its appropriateness for multiple administrations and regarding its sensitivity to factors such as time-of-day would be recommended.

2.2.6 NPRU Mood Scale (NPRU)

The NPRU was developed with the intention of being an easily, briefly, and repeatedly administered mood measure that is sensitive to sleep loss and performance decrements (Monk, 1987). Indeed, it appears to be as easy to complete as it is to administer. Yet, as it consists of 29 items, where ratings are made on a four-point scale and then summed over the positive and negative subscales, the NPRU is suitably complex, so that individuals are not likely to be able to intentionally duplicate or falsify their scores.

In a laboratory study carried out by Angus and Heslegrave (1985), the appropriateness of the NPRU for repeated use was supported. This investigation involved total sleep deprivation over a 54-hour period, during which time participants engaged in

various tasks, were provided periodically with rest breaks, and received the NPRU on an hourly basis. Results for the repeated administration of the NPRU showed that it was sensitive to an overall decrease in positive mood and increase in negative mood over the duration of the experiment. This finding suggests that the NPRU may be appropriate for use in an FFD context. However, additional research looking at repeated scale administration in a surface transportation-related context is still necessary to determine whether this encouraging pattern of results would hold true.

With regard to the sensitivity of the NPRU to factors relevant for operators in surface transportation, the above results suggest that this scale is sensitive to sleep loss. Additionally, the authors (Angus & Heslegrave, 1985) noted performance decrements on a number of measures over the same time period, as well as increased reports of sleepiness and fatigue. Time-of-day fluctuations did not appear to be considered in this study, but nevertheless, results indicate at least some support for the sensitivity of the NPRU to operator-relevant issues, such as decrements over time-on-task, degraded performance, and fluctuations in sleepiness and fatigue. Clearly, the appropriateness of the NPRU as an FFD measure is suggested, however, further research within surface transportation is warranted before the NPRU can be seriously considered for such a use.

2.2.7 Conclusions Regarding Sensitivity and Appropriateness

Each of the subjective measures discussed above demonstrates varying degrees of sensitivity and appropriateness with regard to its use as an FFD instrument in the surface transportation industry. Though not entirely surprising given the fairly recent introduction of some of these scales, it is nevertheless unfortunate that more research does not exist within the surface transportation domain to clarify the relationships between instruments and the dimensions of operator status that they are used to assess. Clearly, it would not be wise to currently recommend any of these measures for use in an FFD context. More research is first necessary within surface transportation to clarify and extend existing findings, thereby more fully developing and characterizing the nature of each of the scales. Furthermore, the current investigations do not even begin to touch on the issue of an individual's ability to successfully continue to work at various levels of impairment, or where cutoffs for safe operation exist. If use as an FFD test is a future possibility for any of these measures, researchers must first endeavor to elucidate and flesh out existing relationships. Thereafter, the next step is the difficult task of establishing criteria for unacceptable/unsafe levels of operator status.

3. Concluding Recommendations for Subjective Measures

Of the six subjective rating scales that were reviewed, each uses a format that would easily lend itself to field assessments of operator status. They differ in that the Epworth Sleepiness Scale, the Pearson and Byars Fatigue Checklist, the Stress-Arousal Checklist, and the NPRU Mood Scale require responses to multiple items using various rating scales, whereas the Karolinska Sleepiness Scale and the Stanford Sleepiness Scale are both single-item measures. Because only one response is required to complete the KSS and the SSS, these instruments would be the most simple to complete, however, their nine- and seven-point scales, respectively, considerably limit the variance in ratings that is possible. Additionally, a single-item

response format would be more likely to be recalled from one test administration to the next, if an operator were looking to duplicate or falsify ratings. For the purposes of FFD assessment, the most desirable type of instrument would be one where a reasonable number of items are scored using a rating scale. This format would generally allow for sufficient variability in responses, a greater likelihood of acceptable levels of reliability, and little chance of individuals recalling their ratings from one test to the next.

With regard to the statistical properties of each of the measurement instruments, the evidence is mixed, but the general recommendation is not. Additional support for the reliability, validity, and sensitivity of each of these rating scales for use in the assessment of surface transportation operators is required, with one exception. The ESS, which was developed fairly extensively in clinical settings, seemed to be more well-suited overall for clinical applications than for surface transportation. This scale was developed to measure a general and persistent component of daytime sleepiness that is independent of time-of-day and day-to-day variation. As such, it is impossible to consider the ESS for routine FFD assessments, which are based on the premise that operator status fluctuates regularly over the day and between days.

Scale measurement properties for the KSS are difficult to assess, since little information on its development is available. As reliability and validity are especially difficult to evidence in single-item measures, this is not surprising, and at the same time indicates that such types of scales are probably not adequate for many measurement situations, especially FFD. Nevertheless, two studies performed in the surface transportation area found results that suggest that the KSS is sensitive to time-of-day, time-on-task, and concomitant decrements in objective measurements. However, with only a small sample size in one study and small numbers of participants per condition in the other, the need is highlighted for further research on the KSS if there is any possibility in using it further within the surface transportation industry.

The Pearson and Byars Fatigue Checklist is backed by a good amount of support for its reliability and validity, though the statistical era that the developmental studies came out of made a current interpretation rather difficult. Along the same lines, a re-evaluation of the items contained in the checklist is warranted, as the eccentricity of some of them by today's standards is apparent, and their effects on the scale's measurement properties is unknown. Although the checklist evidenced sensitivity to time-on-task decrements in a driving simulator study, further research using this scale for assessments of operator status is quite necessary, especially if use for FFD assessment is to be considered.

The popularity of the Stanford Sleepiness Scale suggests that it is well regarded by researchers, and is also likely considered face valid. Nevertheless, one critique that exists for this scale is that the terms found in each "statement" are not mutually exclusive from those in the remaining options . The evidence that was cited for this is that individuals often select more than one of the seven statements in the scale. However, since the participants that Lee was referring to were more than likely patients, it is difficult to generalize this finding to other domains. Within the surface transportation area, there has been more research using this scale than any of the others. A possible by-product of this is that findings for the sensitivity of the SSS are mixed. The explanation that has been offered is that the SSS may measure a different dimension of sleepiness than some

objective tests. At any rate, it is the case that additional research using the SSS in investigations of surface transportation operators is necessary to both further elucidate its measurement properties and explain inconsistencies in the existing findings.

The Stress-Arousal Checklist was derived from a popular American measure of mood and used in the UK, where it was found to demonstrate evidence in a laboratory experiment for suitable levels of reliability and some support for its construct validity. However, within surface transportation, the checklist was only found in one Australian field study. This investigation had very few participants, but offered tentative evidence for the SACL's sensitivity to long hours of driving. Especially since this mood measure was altered from a state of being "too American", research utilizing it in different cultures and within surface transportation to establish additional reliability and validity evidence is warranted. Moreover, the relationship between the SACL and relevant factors that are related to operator status needs to be further characterized if this measure is to be used subsequently in the surface transportation domain or possibly in an FFD context.

Another mood measure, the NPRU, was developed fairly extensively, but traditional statistical information for the scale's reliability was not provided. Laboratory investigations offered some evidence to indicate that the NPRU reflects time-on-task and performance decrements, as well as sleepiness/fatigue. However, within the surface transportation literature, only one relevant research effort was located. This study reported reductions in positive mood and elevations in negative mood on workdays compared to non-workdays, but did not break the relationship down further, as this was not the focus of the investigation. In order to be able to consider the NPRU as a potential measure of FFD, research should be conducted with the intention of demonstrating additional evidence for the reliability, validity, and sensitivity of this instrument in a surface transportation setting.

Finally, it is worth noting that the aforementioned measures are developed and administered in English. For this reason, it might not be easy to judge how suitable these inventories are for use in other cultures, as it is fairly obvious that some of the expressions may have suffered somewhat in translation. This issue remains a concern to be addressed in the further research.

References

Akerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, *52*(1-2), 29-37.

Angus, R. G., & Heslegrave, R. J. (1985). Effects of sleep loss on sustained cognitive performance during a command and control simulation. *Behavior Research Methods, Instruments and Computers*, *17*(1), 55-67.

Bartley, S. H., & Chute, E. (1947). *Fatigue and impairment in man.* New York: McGraw-Hill.

Brown, I. D. (1994). Driver fatigue. *Human Factors*, *36*(2), 298-314.

Carmines, E. G., & Zeller, R. A. (1980). *Reliability and validity assessment.* Thousand Oaks, CA: Sage Publications.

Carskadon, M. A. (1979). Effects of total sleep loss on sleep tendency. *Perceptual and Motor Skills*, *48*(2), 495-506.

Cox, T., Mackay, C., & Page, H. (1982). Simulated repetitive work and self-reported mood. *Journal of Occupational Behaviour*, *3*(3), 247-252.

Cruickshank, P. J. (1984). A stress and arousal mood scale for low vocabulary subjects: A

reworking of Mackay et al. (1978). *British Journal of Psychology*, *75*(1), 89-94.

Dalziel, J. R., & Job, R. F. S. (1997). Motor vehicle accidents, fatigue and optimism bias in taxi drivers. *Accident Analysis and Prevention*, *29*(4), 489-494.

Evans, G. W. (1994). Working on the hot seat: Urban bus operators. *Accident Analysis and Prevention*, *26*(2), 181-193.

Gillberg, M., Kecklund, G., & Akerstedt, T. (1994). Relations between performance and subjective ratings of sleepiness during a night awake. *Sleep*, *17*(3), 236-241.

Hakkanen, J., & Summala, H. (2000). Sleepiness at work among commercial truck drivers. *Sleep*, *23*(1), 49-57.

Harnish, M. J., Chard, S. R., & Orr, W. C. (1996). Relationships between measures of objective and subjective sleepiness. *Sleep Research*, *25*, 492.

Hartley, L. R., Arnold, P. K., Smythe, G., & Hansen, J. (1994). Indicators of fatigue in truck drivers. *Applied Ergonomics*, *25*(3), 143-156.

Herscovitch, J., & Broughton, R. (1981). Sensitivity of the Stanford Sleepiness Scale to the effects of cumulative partial sleep deprivation and recovery oversleeping. *Sleep*, *4*(1), 83-91.

Hoddes, E. (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, *10*(4), 431-436.

Hoddes, E., Dement, W. C., & Zarcone, V. (1972). The development and use of the Stanford Sleepiness Scale (SSS). *Psychophysiology*, *9*, 150.

Horne, J. A. (1991). Dimensions to sleepiness. In T. H. Monk (Ed.), *Sleep, sleepiness and performance* (pp. 169-196). New York: John Wiley & Sons.

Horne, J. A., & Reyner, L. A. (1995). Driver sleepiness. *Journal of Sleep Research*, *4*(S2), 23-29.

How, J. M., Foo, S. C., Low, E., Wong, T. M., Vijayan, A., Siew, M. G., & Kanapathy, R. (1994). Effects of sleep deprivation on performance of naval seamen: I. Total sleep deprivation on performance. *Annals of the Academy of Medicine, Singapore*, *23*(5), 669-675.

Johns, M. W. (1992). Reliability and factor analysis of the Epworth Sleepiness Scale. *Sleep*, *15*(4), 376-381.

Johnson, L. C., Freeman, C. R., Spinweber, C. L., & Gomez, S. A. (1988). *The relationship between subjective and objective measures of sleepiness.*. San Diego, CA: Naval Health Research Center.

Kecklund, G., & Åkerstedt, T. (1993). Sleepiness in long distance truck driving: An ambulatory EEG study of night driving. *Ergonomics*, *36*(9), 1007-1017.

Kinsman, R., & Weiser, P. (1976). Subjective symptomatology during work and fatigue. In E. Simonson & P. C. Weiser (Eds.), *Psychological aspects and physiological correlates of work and fatigue* (pp. 336-405). Springfield, IL: Charles C. Thomas.

Lee, K. A., Hicks, G., & Nino-Murcia, G. (1991). Validity and reliability of a scale to assess fatigue. *Psychiatry Research*, *36*(3), 291-298.

Lubin, A., Moses, J. M., Johnson, L. C., & Naitoh, P. (1974). The recuperative effects of REM sleep and stage 4 sleep on human performance after complete sleep loss: Experiment I. *Psychophysiology*, *11*(2), 133-146.

Mackay, C., Cox, T., Burrows, G., & Lazzerini, T. (1978). An inventory for the measurement of self-reported stress and arousal. *British Journal of Social and Clinical Psychology*, *17*(3),

283-284.

MacLean, A. W., Saskin, P., Fekken, G. C., & Knowles, J. B. (1989). Should the Stanford Sleepiness Scale be revised? *Sleep Research, 18*, 370.

Maycock, G. (1997). Sleepiness and driving: The experience of U.K. car drivers. *Accident Analysis and Prevention, 29*(4), 453-462.

McDonald, N. (1989). Fatigue and driving. *Alcohol, Drugs and Driving, 5*(3), 185-192.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling*. Newbury Park, CA: Sage Publications.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1981). *Manual of the Profile of Mood States*. San Diego: Educational and Industrial Testing Service.

Mitler, M. M., & Miller, J. C. (1995). Methods of testing for sleeplessness. *Behavioral Medicine, 21*(4), 171-183.

Monk, T. H. (1987). Subjective ratings of sleepiness: The underlying circadian mechanisms. *Sleep, 10*(4), 343-353.

Moses, J. M., Lubin, A., Naitoh, P., & Johnson, L. C. (1974). *Subjective evaluation of the effects of sleep loss: The NPRU Mood Scale* (74-25). San Diego, CA: Naval Health Research Center.

Nilsson, T., Nelson, T. M., & Carlson, D. (1997). Development of fatigue symptoms during simulated driving. *Accident Analysis and Prevention, 29*(4), 479-488.

Nowlis, V. (1965). Research with the Mood Adjective Check List. In S. S. Tomkins & C. E. Izard (Eds.), *Affect, cognition, and personality* (pp. 352-389). New York: Springer Publishing Company.

Pearson, R. G. (1957). Scale analysis of a fatigue checklist. Journal of Applied Psychology, 41(186-191).

Pearson, R. G., & Byars, G. E., Jr. (1956). *The development and validation of a check-list for measuring subjective fatigue* (No. 56-115): USAF School of Aviation Medicine.

Philip, P., Ghorayeb, I., Leger, D., Menny, J. C., Bioulac, B., Dabadie, P., & Guilleminault, C. (1997). Objective measurement of sleepiness in summer vacation long-distance drivers. *Electroencephalography and clinical neurophysiology, 102*(5), 383-389.

Pivik, R. T. (1991). The several qualities of sleepiness: Psychophysiological considerations. In T. H. Monk (Ed.), *Sleep, Sleepiness and Performance*. New York: John Wiley & Sons.

Raggatt, P. T., & Morrissey, S. A. (1997). A field study of stress and fatigue in long-distance bus drivers. *Behavioral Medicine, 23*(3), 122-129.

Ranney, T. A., Simmons, L. A., & Masalonis, A. J. (1999). Prolonged exposure to glare and driving time: Effects on performance in a driving simulator. *Accident Analysis and Prevention*, 31(6), 601-610.

Thayer, R. E. (1967). Measurement of activation through self-report. *Psychological Reports, 20*(2), 663-678.

Watts, C., Cox, T., & Robson, J. (1983). Morningness-eveningness and diurnal variations in self-reported mood. *Journal of Psychology, 113*, 251-256.

Wiley, C. D., Shultz, T., Miller, J. C., Mitler, M. M., & Mackie, R. R. (1996). *Commercial motor vehicle driver fatigue and alertness study: Project report* (FHWA-MC-97-001). Washington, D.C.: Federal Highway Administration.