Principles of Subjective Rating Scale Construction

B. G. HUTCHINSON, Department of Civil Engineering, University of Waterloo, Canada

This paper reviews the basic principles of subjective rating scale construction that have been developed in psychophysics and applied psychology. Particular emphasis is placed on the subjective measurement of pavement serviceability. The rationale underlying the AASHO Road Test pavement serviceability rating system is examined on the basis of these principles. Although some of the deficiencies of this serviceability measuring technique are illustrated, the full impact of these principles cannot be assessed until further experimental studies are undertaken.

•ONE of the most significant developments resulting from the recently completed AASHO Road Test was the formulation and definition of the concepts of serviceability and failure of highway pavements, reported by Carey and Irick (1). Although the concept of pavement serviceability has been used more or less intuitively for many years to gage the success of pavement designs, the significant contribution of this study (1) was to demonstrate that serviceability was quantifiable. Furthermore it was shown that serviceability is a psychological quantity or experience, and not a physical measurement derived from pavement surface roughness.

The technique developed for measuring both serviceability and failure at the Road Test is based on a subjective estimate procedure. Although the manner in which human beings gage serviceability is necessarily an empirical problem, the known facts of psychophysics, however, set certain valuable guidelines.

It is well established that psychological experiences are measurable. However, all psychophysical quantities are subject to potential bias and distorting factors. The fact that an observer can be influenced in what he reports does not mean that his psychological impressions are not quantifiable, but merely that the task of measurement is difficult. An observer is sensitive not only to the physical stimuli he is trying to measure, but also to a large number of other factors that can distort his judgment to varying degrees.

In view of the susceptibility of human observers to external influences in communicating their psychological impressions, most psychophysical investigations seek to establish a measurement scale of the psychological experience and to relate this to a scale of measurement of the physical stimulus. Routine estimates of a particular psychological magnitude are then made from measurements of the physical correlate. The pavement serviceability rating system described by Carey and Irick (1) was developed within this type of framework. However, it is apparent that this subjective measurement procedure was developed without full cognizance of the basic principles of subjective rating scale construction. It is the purpose of this paper to review the basic principles of subjective rating scale construction, and to examine the validity of the rationale underlying the Road Test pavement serviceability rating procedure.

Paper sponsored by Committee on Pavement Condition Evaluation.

MEASUREMENT OF SERVICEABILITY

Measurement, in general, is concerned with the rationale involved in the construction of a measuring scale, as well as the properties that can be attributed to measurements executed with a scale. The measurement of the majority of the properties of objects are expressed in the real number system. The real number system possesses certain fundamental properties of which the most important are order and additivity. The order of numbers is given by convention. Additivity refers to the fact that the operation of addition (used here in the completely general sense) gives results that are internally consistent. In other words, equal differences can be determined from the numbers, such as 7 - 5 = 4 - 2, as well as equal ratios, such as 8/4 = 6/3.

If it is possible to assign numbers to the properties of objects such that the properties of objects designated by the various numbers have the same characteristics as the number system (that is, if an isomorphism exists), then the number system may be used as a mathematical model of the properties of the object. It is, therefore, of great analytical advantage if this isomorphism between the properties of numbers and the properties of objects can be established. The principles and manipulations of mathematics applicable to the number system may then be used to manipulate the properties of the objects themselves.

Measurement exists in a variety of forms depending on the extent to which the properties of the number system are reflected in the scale of measurement. Measurement scales are classified into four basic types, and the classification proposed by Stevens (15), and given in Table 1, is generally accepted. Each of the four scale types given in Table 1 reduces the completely arbitrary element in the assignment of numbers to property magnitudes to a different degree. Stevens (17) suggests that a fundamental technique for evaluating scales of measurement is by using the concept of invariance of scale values under transformations of the scale. Table 1 contains a brief description of the empirical rule or operation invoked in the measurement operation, the transformations under which each scale type remains invariant, and an example of each scale type. Inasmuch as the arbitrary element in the assignment of numbers to properties is restricted to a different degree for each scale type, the characteristics of the numbers that are available for meaningful use as a model of object properties are likewise restricted.

The measurement problem with respect to pavement serviceability, therefore, resolves to one of first deciding the level of measurement required, and second, developing a procedure for scaling serviceability at this level of measurement. If the serviceability is to be used for establishing maintenance or resurfacing priorities for example, then all that would be required would be an ordinal scale of measurement; that is, an ordering of the pavement sections. However, if the measure is to be used to establish statistical relations between the serviceability measure and other factors, such as pavement strength and environment, then pavement serviceability must be measured on at least an interval scale. For example, consider a hypothetical pavement section that was rated at three periods throughout its life as possessing serviceability ratings of 4.0, 3.0 and 2.0. Unless the difference in serviceability between the first and second ratings is equal to the difference between the second and third ratings, it would be meaningless to attempt to relate these changes in serviceability to, say, differences in axle coverages.

TABLE 1

A CLASSIFICATION OF SCALES OF MEASUREMENT[a]

Scale	Basic Empirical Operations[b]	Allowable Transformations	Example
Nominal	Determination of equality	Any one to one substitution	"Numbering" of football players
Ordinal	Determination of greater or less	Any increasing monotonic function	Moh hardness scale of minerals
Interval	Determination of equality of intervals	Any linear transformation	Temperature (°F)
Ratio	Determination of equality of ratios	Any linear transformation retaining natural origin	Length, density, temp. (Kelvin)

[a] After Stevens (17).

[b] The basic operations needed to create a given scale are those listed down to and including the operation listed opposite the scale.

PSYCHOPHYSICAL MEASUREMENT

Psychophysics is concerned with the determination of quantitative relationships between physical stimuli and corresponding psychological or sensory events. The notions of a stimulus continuum, a response or sensory continuum, and a judgment continuum must be introduced to comprehend the principles of psychophysical measurement.

A stimulus or physical continuum refers to changes in some physical property such as the frequency of a sound wave, frequency of vibration, amplitude of vibration, or weight in pounds. Corresponding to these physical stimuli are certain sensory experiences or response continua such as pitch, perceived frequency of vibration, or subjective weight. It is not possible to measure directly quantities on the response continuum because these may only be estimated by observing an external verbal or symbolic response of an observer; that is, in the form of an externally communicated judgment by an observer. It is from these judgments that evidence concerning the response continuum must be derived. The introduction of a third continuum is, therefore, necessary for logical interpretation of response or sensory continua and their relationship to the corresponding physical continua. The exact manner in which human beings detect and respond to such physical stimuli as vibration, noise, etc., is not clearly understood. Goldman (5) and Hornick (10) have discussed the functions of some of the anatomical and biological systems that detect vibrations and motions, while Stevens (16) has investigated some of the factors concerned with the perception of noise.

Existing psychophysical theory, therefore, presupposes the existence of a judgment continuum paralleled by a response continuum, and through this relationship the judgment continuum is also related to the stimulus continuum. Common practice in psychophysics has been to assume a linear regression relating the judgment and response continua with perfect correlation. However, this correlation is not always perfect, and the nature of this correlation is discussed in more detail later in this paper.

Guilford (7) has provided an exhaustive review of the psychophysical scaling procedures that have been developed and used to establish relationships between these continua. Although they differ in detail, all psychophysical scaling methods may be considered as the combined effect on the sensory response of an observer, of an experimenter's operations of stimulation and instruction. Instruction refers to the response that is elicited from an observer, and the major difference in psychometric scaling methods is due to the nature of the response so obtained. Scaling methods are generally classified as judgment or response methods. With judgment methods the observer is instructed to assess the amount of a specified attribute possessed by a physical stimulus. With the response methods a specific attribute of the stimulus is not specified, and the observer is instructed only to indicate whether he agrees with or endorses a particular stimulus.

Rating scale methods are the most popular psychometric scaling procedures that depend on human judgment. Because of their widespread use, a significant body of principles governing their construction and use have been generated. They have been used in personnel evaluation, the reactions of individuals, aesthetic judgments, in the psychological evaluation of physical stimuli, etc.

PSYCHOLOGICAL MODEL OF RATING PROCEDURES

Several types of rating scales have been developed and widely used, but they are all essentially alike in that they require the assignment of objects by inspection, either along an unbroken continuum, or in ordered categories along the continuum. A numerical rating scale, an example being the Road Test serviceability scale, typically consists of a sequence of numbers defined by definitions or cues, and raters assign an appropriate number to each stimulus in line with these cues. Although other types of rating scales, such as graphic scales, have been developed and used, the numerical rating scale is the most appropriate scale type for pavement serviceability measurement if an interval scale of measurement is to be achieved.

On the basis of the previously outlined existing psychophysical principles which are discussed in detail by Guilford (7) and Torgerson (20), the following psychological model is proposed as the most appropriate model underlying the subjective determination of pavement serviceability.

1. Serviceability is a discriminable attribute of highway pavements and raters are

capable of making direct quantitative judgments of the amount of this attribute associated with any pavement section.

2. Each rater's judgment is considered to be a direct report of the level of serviceability of a pavement on a linear subjective continuum (interval scale) of this attribute. The origin and units in which the judgments are expressed may be arbitrary but they must remain constant.

3. Some variability in judgment with respect to the serviceability of any pavement may occur as is the case with any measurement procedure. This variability is treated as random error, and the individual estimates may be averaged to provide an estimate of the scale value of serviceability. It is implicitly assumed that the scale value estimate may be obtained from replications by an individual, or from judgments by a number of raters. That is, raters are assumed to be interchangeable.

Although not explicitly stated by Carey and Irick (1), this is essentially the psychological model assumed in the AASHO Road Test rating procedure. Guilford (7) however, has pointed out that several well known systematic errors occur in rating methods, and that these systematic errors must be removed from the raw judgments before psychological models of the aforementioned type may be considered to hold in actual rating studies. The most important of the recognized systematic rating errors occuring in ratings are, as follows:

1. The error of leniency which refers to the constant tendency of a rater to rate too high or too low for whatever reasons.

2. The halo effect which refers to the tendency of raters to force the rating of a

TABLE 2

SERVICEABILITY RATING MATRIX OF MINNESOTA AND INDIANA RIGID PAVEMENTS

Pvmt. No.	Serviceability Rating									Mean
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9	
201	1.8	1.0	2.4	2.0	0.7	1.2	0.7	1.0	0.9	1.3
202	1.9	1.1	1.7	2.4	2.7	2.1	1.5	1.2	1.4	1.8
203	2.4	1.8	3.1	2.7	1.7	2.4	1.5	1.7	1.5	2.1
204	4.4	4.1	4.5	4.0	3.8	3.8	3.6	4.2	4.1	4.1
205	4.4	3.9	4.5	3.5	3.5	3.8	3.1	3.8	4.0	3.8
206	3.6	2.4	3.7	2.6	3.2	3.2	2.9	2.9	2.4	3.0
207	3.5	2.4	4.2	2.4	3.0	3.6	2.5	3.2	2.6	3.0
208	3.4	2.4	4.1	2.2	2.7	3.2	2.8	2.8	2.3	2.9
209	2.9	1.8	3.8	2.4	2.9	2.3	2.2	2.5	2.3	2.6
210	1.9	1.2	2.0	1.9	1.6	1.6	0.8	3.0	1.5	1.7
211	4.9	4.7	4.8	4.2	4.7	4.3	4.1	4.3	4.2	4.5
212	4.9	4.2	5.0	3.8	4.5	3.8	4.3	4.3	4.3	4.3
213	4.3	3.4	4.2	3.8	3.8	3.6	3.1	3.7	3.5	3.7
214	4.2	2.9	3.2	3.5	3.1	3.8	3.3	4.0	4.0	3.5
215	4.6	3.4	4.8	3.3	4.6	4.3	3.7	4.5	4.0	4.1
216	4.5	3.2	4.7	3.6	4.4	4.0	3.0	4.3	3.5	3.9
217	1.7	1.0	1.2	1.8	1.5	1.6	0.7	0.8	1.0	1.3
218	1.8	1.0	1.6	1.8	0.9	1.6	0.8	0.7	1.0	1.2
219	3.6	1.8	3.8	2.9	2.7	3.2	2.9	3.7	2.0	3.0
220	4.8	4.2	5.0	4.4	4.6	4.1	3.9	4.4	4.3	4.4
401	4.0	3.8	4.5	4.3	3.8	3.2	4.1	3.9	4.0	4.0
402	3.9	3.2	4.8	3.4	3.7	3.7	4.1	4.0	3.8	3.8
403	3.7	3.3	4.7	3.6	3.9	3.4	3.5	3.5	2.5	3.6
404	3.3	2.1	4.2	3.7	2.9	3.0	3.6	3.2	2.5	3.2
405	3.0	1.8	3.5	2.5	2.6	2.3	3.1	2.2	2.2	2.6
406	3.0	2.5	3.2	3.1	2.8	2.7	3.1	2.7	2.4	2.8
407	2.8	1.6	1.8	1.5	2.4	2.0	2.5	1.5	1.0	1.9
408	2.6	1.5	1.8	0.8	2.8	2.4	2.5	2.0	1.0	1.8
409	3.1	1.9	2.7	1.5	2.2	1.8	3.0	2.1	1.0	2.1
410	3.0	2.0	2.2	1.8	2.7	2.2	3.1	2.1	1.3	2.3
411	2.5	1.5	1.7	0.8	2.3	1.5	2.3	2.0	1.3	1.8
412	4.0	1.8	3.1	2.3	2.5	2.7	3.1	3.0	2.3	2.8
413	4.1	4.3	4.9	4.1	4.6	3.4	4.3	4.5	4.0	4.2
414	4.1	4.7	4.7	4.4	4.4	3.5	4.1	4.6	4.1	4.3
415	4.0	4.6	4.9	4.3	4.5	3.4	4.7	4.2	4.2	4.3
416	2.1	0.6	0.4	0.4	2.0	1.7	0.6	1.0	1.0	1.1
417	3.0	1.7	3.0	1.7	2.2	1.8	2.6	2.5	1.0	2.2
418	4.4	4.3	4.9	4.3	4.3	4.4	3.9	4.6	4.0	4.3
419	3.4	2.0	3.6	2.6	2.9	3.4	3.1	2.5	1.4	2.8
420	3.2	2.3	3.0	2.2	3.0	2.9	2.7	2.3	2.0	2.6

particular attribute in the direction of the overall impression of the object rated.

3. The error of central tendency which refers to the fact that raters hesitate to give extreme judgments of stimuli and tend to displace individual ratings toward the mean of the group.

Table 2 contains the individual ratings of nine raters of the 40 Minnesota and Indiana rigid pavement sections surveyed in the rating studied reported by Carey and Irick (1). This rating matrix is examined in the following section for the presence of systematic errors of the aforementioned type, and the techniques that have been developed for removing these errors are outlined.

SYSTEMATIC RATING ERRORS

An analysis of variance of the rating matrix of Table 2 is given in Table 3. Both sources of variation, that of between raters and between pavements, are shown to be significant at the 1 percent level of significance. The previously described psychometric model of the rating procedure requires that raters be interchangeable, but the data in Table 3 illustrate that this requirement is violated in that the differences in ratings between raters are significant. Table 4 gives the mean rating and the standard deviation of ratings for each rater. This source of variation in ratings between ratings may be removed by transforming each rater's ratings to a distribution with mean and dispersion equal to the grand mean rating and the mean standard deviation.

The deviation of each rater's average rating for all pavement sections from the grand mean rating will indicate the magnitude of a rater's relative leniency error. The relative leniency errors, ΔR, are given in Table 4, and inspection of Table 2 reveals that, in general, this relative leniency error is constant for a given rater. The transformation of each rater's dispersion of ratings to a constant standard deviation is necessary because the contribution to the total variance of all ratings of a rater's ratings is proportional to the magnitude of the standard deviation of his ratings.

Guilford (7) has stated that a positive leniency error has been found to be the most common type of leniency error, but with the present data it is not possible to estimate the absolute leniency error. Guilford has further suggested that the descriptive cues may be adjusted to counteract this type of error by giving most of the scale range to degrees of favorable report. Evidently, raters anticipate a mean rating scale value somewhere near the cue good, or its equivalent, and a distribution symmetrical about that point.

The second type of systematic error common to ratings is known as the halo effect. The halo effect is considered to be a constant type error, and has been previously defined as the tendency of raters to force the rating of a particular trait in the direction of the overall impression of the object rated, and to that extent to make the ratings of some traits less valid. Symonds (18) suggests that the halo effect is more prevalent in ratings when a trait is not easily observable, or when the trait is not clearly defined. A relative halo effect would be manifested in significant interaction terms in an analysis of variance of a rating matrix in which the level of all attributes influencing a rater's ratings were systematically and quantitatively recorded. The available rating data cannot be analyzed for this error because the pavement traits were not systematically recorded.

The third type of error, known as the

TABLE 3

ANALYSIS OF VARIANCE OF SERVICEABILITY RATING MATRIX

Source	Sum of Squares	D.F.	Var.	F Ratio	P
Between raters	386	8	9.9	55	Significant at 0.01 level
Between pavement	12	39	1.5	8.3	
Remainder	57	312	0.18		
Total	455	359			

TABLE 4

MEANS AND STANDARD DEVIATIONS OF RATINGS BY RATERS

Rater No.	Mean Rating	R	Std. Dev.
1	3.40	+ 0.44	0.87
2	2.58	- 0.38	1.20
3	3.50	+ 0.54	1.25
4	2.81	- 0.15	1.39
5	3.05	+ 0.09	1.07
6	2.92	- 0.04	0.98
7	2.89	- 0.07	1.09
8	2.98	+ 0.02	1.21
9	2.54	- 0.42	1.25
Mean	2.96		1.15

central tendency error, has been defined as the tendency of raters to judge stimuli in the direction of the average stimulus. One factor contributing to this error is that raters tend to displace ratings towards the mean of the group. Johnson (12) has explained this error from a statistical viewpoint, in terms of the regression towards the mean that always occurs when two variables are imperfectly correlated. In a judgment situation, this imperfect correlation results from imperfect discrimination of an attribute by an observer. The interpretation of this error, therefore, requires the introduction of the judgment continuum as distinct from the sensory continuum for logical explanation.

If a central tendency effect has occurred in a set of ratings, then the scale value estimates will have much less dispersion than the true scale values. The problem of removing this error resolves to one of establishing a relation between the true scale values, Tj, and the obtained scale values, Mj. The obtained scale values are the mean values of the individual estimates by raters. Guilford (7) has established this as a regression problem in which the obtained scale values, $\overline{M}j$, are predicted from the true scale values, Tj, and the dispersion of the single judgments, Aj, around these means represents the errors of prediction. Guilford assumes that the obtained and true scale values are perfectly correlated, and that the means of the two sets of values are equal. Further, it is assumed that the standard deviations of the true scale values and all single values are perfectly correlated. Guilford states that the justification of this assumption is that in the limiting case when the correlation between the values is perfect, Aj is perfectly predicted from Tj. Inasmuch as the correlation between Tj and Mj is perfect, all that is required is a linear transformation equation.

The standard deviation of all single ratings (transformed to equivalent distributions as previously described) is 1.18. The standard deviation of the obtained mean scale values is 1.04. Therefore, the transformation equation becomes

$$Tj = \frac{1.18}{1.04} (Mj - \overline{M}j) + \overline{M}t =$$

$$1.134\ Mj - 0.39 \qquad (1)$$

because $\overline{M}j = \overline{M}t = 2.94$.

Table 5 gives the relationship between the obtained scale values and the true scale values for the Indiana and Minnesota rigid pavements. The mean values of both sets of scale values (2.94), which are assumed to be equal, define the indifference point. The table shows that ratings below this point are overestimated, whereas ratings above this point tend to be underestimated. In general the greater the distance of a stimulus from the indifference point, the greater the error of estimation.

TABLE 5

COMPARISON OF OBTAINED AND TRUE RATING SCALE VALUES FOR INDIANA + MINNESOTA RIGID PAVEMENTS

Pavement Section	Scale Value		Pavement Section	Scale Value	
	True	Obtained		True	Obtained
201	1.3	0.9	401	4.0	4.0
202	1.8	1.6	402	3.8	4.0
203	2.1	1.9	403	3.6	3.6
204	4.1	4.2	404	3.2	3.2
205	3.8	3.9	405	2.6	2.1
206	3.0	3.1	406	2.8	3.0
207	3.0	3.1	407	1.9	1.8
208	2.9	2.9	408	1.8	1.7
209	2.6	2.4	409	2.1	2.1
210	1.7	1.4	410	2.3	2.2
211	4.5	4.7	411	1.8	1.6
212	4.3	4.6	412	2.8	2.8
213	3.7	3.8	413	4.2	4.4
214	3.5	3.7	414	4.3	4.4
215	4.1	4.4	415	4.3	4.4
216	3.9	4.1	416	1.1	0.9
217	1.3	1.0	417	2.2	2.2
218	1.2	1.0	418	4.3	4.5
219	3.0	2.9	419	2.8	2.6
220	4.4	4.6	420	2.6	2.6

The foregoing rationale assumed that the amount of under- or overestimation to be a linear function of the distance of the stimulus from the indifference point. Torgerson (20) has pointed out that a tendency exists for observers to force any series of stimuli into a normal distribution. If this were true, then scales constructed according to this rationale would not possess equal interval properties. However, the exact nature of the regression cannot be evaluated unless a corresponding physical continuum is available.

A second limitation of this rationale is concerned with the assumption that the discriminal dispersions at each scale value are equal. If these dispersions are not equal then the stimuli having greater dispersions may have regressed more toward the mean than stimuli having smaller dispersions. This would violate the assumption that perfect correlation exists between the true and obtained scale values. Inspection of the Road Test rating data suggests that the dispersions in ratings vary with the magnitude of the scale value; the dispersions at extreme scale values are less than the dispersions of the more central scale values. Insufficient data are available to properly evaluate this factor.

Newcomb (14) and Murray (13) have suggested that other significant errors, similar in nature to the halo effect, do frequently occur in subjective ratings. However, accepted methods for removing these errors from ratings have not been evolved due to an incomplete understanding of the precise nature of these errors.

Some of the quantifiable errors that frequently occur in ratings have been described and the following sections describe more general, but equally significant errors that distort ratings, and that must be minimized or removed from the ratings.

VALIDITY AND RELIABILITY OF RATINGS

The validity of ratings refers to the degree to which they are truly indicative of a psychological experience generated by a physical stimulus. The reliability of ratings refers to the consistency with which ratings are made, either by different raters, or by one rater at different times.

Rigorous validation of ratings is only possible through the comparison of ratings with more objective measures of the stimulus attribute. However, due to the complexity of many physical stimuli, precise measures of physical correlates are unknown. In such a case it is important to examine thoroughly the factors influencing the validity of ratings. Conditions may then be established which will be conducive to producing the highest possible validity in subjective estimates.

A most important factor affecting the validity of ratings is the definition of the attribute of an object that is to be rated. Guilford (7) has pointed out that many psychophysical investigations have demonstrated that an attribute name is primarily useful as a label, and used without adequate definition and without cues may become very misleading. Ghiselli and Brown (4) have pointed out that when personnel are rated on the basis of a general or overall trait there is greater probability of error, because different raters will base their judgments on different aspects of the performance included under the general trait name.

In view of the statements by Carey and Irick (1), Hveem (11), Housel (9), and Wilkins (21), there is some confusion concerning the exact nature of pavement serviceability. The terms pavement serviceability and pavement roughness have been used interchangeably. Pavement roughness refers to the distortion of a pavement surface from the geometry of the designed surface. The serviceability and failure of an engineering design can only be defined relative to the purpose for which a design has been provided. The purpose of a highway pavement is, as has frequently been stated, to provide a surface of adequate riding qualities throughout the life of a pavement. The riding quality afforded by a particular pavement section is a subjective experience and must be measured as such. The absolute riding quality is not a unique subjective characteristic but depends on the interrelationship of the pavement roughness, vehicle, and vehicle occupants. An absolute scale of riding quality would require the establishment of absolute levels of subjective experience that result from particular vibrational environments. A particular pavement section would, therefore, exhibit a wide range of riding qualities depending on the properties of the vehicular system using it. Consequently, pavement serviceability must be operationally defined in terms of the relative riding quality for each highway user. It is apparent that rating efforts at the Road Test were directed toward obtaining subjective estimates of the pavement distortion and deterioration, and not to obtaining estimates of the subjective experiences of riding quality.

Guilford (7) and Ghiselli and Brown (4) have provided further information of many of the other factors that are known to influence the validity of ratings.

The reliability of ratings is commonly defined operationally as the proportion of observed variance that is true variance. One technique for estimating the reliability of ratings is by re-rating a given set of physical stimuli and correlating the two sets of ratings. Guilford (7) has suggested that such a technique is susceptible to spurious correlation due to the memory of raters.

Ebel (1951) has described a method of estimating the reliability of ratings which is based on an analysis of variance of the ratings. The reliability of ratings for a single rater is given by

$$r = \frac{Vp - Ve}{Vp + (k - 1)\,Ve} \tag{2}$$

while the reliability of the mean ratings of the raters is given by

$$r = \frac{Vp - Ve}{Vp} \tag{3}$$

in which

r = reliability of ratings,
Vp = variance between pavements (or other stimulus),
Ve = variance of residuals, and
k = number of raters.

The reliability coefficient can be readily computed for any rating matrix. Although the coefficient is not particularly meaningful for a single matrix, it is invaluable in the evaluation of various scale formats as is pointed out later in the paper.

SCALE CONSTRUCTION AND FORMAT

To assist raters in arriving at quantitative judgments at an interval scale level of measurement, the attribute definition should be supplemented and reinforced by cues or descriptive phrases. Champney (2) after an extensive study has listed criteria that may be used as a guide to the systematic development of cues for rating studies. The most important of these recommendations are that cues should apply to a very short and particular range on the continuum to provide raters with definite anchors, and that the cues for each trait should be unique to that trait. In particular, cues of a very general character such as "excellent," "poor," etc., should be avoided. The determination of the optimum scale format for a particular rating situation is necessarily an empirical problem. The error of leniency and the central tendency effect may be minimized by judicious selection of cues. It was previously pointed out that a positive leniency error may be minimized by using only unfavorable cues, because raters anticipate a mean rating somewhere near the cue "good" or its equivalent. The error of central tendency may be counteracted by adjusting the strength of the descriptive phrases. Greater differences in meaning may be introduced between steps near the extremities of the scale than between steps near the central area.

A most important parallel problem concerns the number of steps or categories that should be employed in a rating scale. The Road Test scale uses five categories. Wilkins (21) has pointed out that a ten-category scale is used in the Canadian rating studies. If the steps in a rating scale are too coarse, the raters' powers of discrimination cannot be effectively used. However, loss of reliability may result from steps that are finer than the raters' discrimination abilities. Symonds (19) has suggested that optimum reliability in ratings will be obtained by seven categories. Other studies have shown that the optimum number of steps varies considerably with the nature and complexity of the trait being rated. Consequently, the optimum number of categories may only be determined by experimental evaluation, and one objective criterion for optimizing the number of categories is the reliability coefficient, defined in Eqs. 2 and 3.

INTERVAL SCALE PROPERTIES

The basic psychometric model previously described has proceeded under the assumption that raters are capable of judging stimuli on an equal interval scale. Although the previous sections have been devoted to describing various errors and distortions common to ratings, no explicit provision for testing this fundamental assumption was proposed. Torgerson (20) has pointed out that consistency of judgments, or reliability, is not an adequate criterion to evaluate this assumption. For example, a criterion of judgment consistency cannot distinguish between equal interval judgments and judgments of ordinal positions of stimuli. Both Torgerson (20) and Stevens (16) suggest that the best approach to this evaluation lies in an examination of the invariance characteristics that an interval scale must possess.

It was pointed out in Table 1 that an equal interval scale is one in which the numbers assigned to stimulus magnitudes must be determined within a linear transformation of the form $y = a + bx$. That is, ratios of differences between scale values must remain invariant upon transformation. Thus, for a subjective estimate scale without a physical correlate, the minimum requirement for an interval scale would be that the ratios of differences in scale values assigned to at least three stimuli should remain invariant when the stimuli are scaled under differing experimental conditions. That is, a linear relation should exist between the different sets of scale values. Such an evaluation is not feasible with the available Road Test rating matrices.

Guilford (7) has concluded from the limited number of studies carried out to evaluate the measurement status of rating scales, that they may be regarded as having the status of ordinal measurements and only approach the status of interval measurements. He further suggests by the various methods of scaling and correction that they can be more or less successfully transformed to interval scale measurements.

One factor important to achieving valid and reliable ratings is the notion of scale anchoring. The anchoring concept refers to those conditions that control the origin and unit of the subjective continuum in which raters will report their judgments of magnitude. Experimental studies of anchoring effects have demonstrated that the unit and origin in which judgments are expressed are not absolute, but are functions of a particular experimental situation. Raters adjust the origin and unit to the distribution of the particular set of stimuli rated and to the rating categories allowed. The majority of psychophysical investigations have achieved invariant units and origin by systematic training of raters.

The preliminary rating studies at the Road Test were carried out to establish a common unit and origin for the rating panel. However, the application of a subjective serviceability rating procedure to pavement rating on a national scale may result in significant discrepancies in ratings from area to area. Rating panels from regions in which a wide distribution of pavement serviceabilities exists might be expected to establish a different subjective unit of serviceability than panels from regions possessing a much narrower range of pavement serviceabilities. Similarly, the origin of ratings would be expected to be a function of the average serviceability level existing in a region.

PHYSICAL CORRELATE OF SERVICEABILITY

An immediate problem in the serviceability measurements of pavements is to establish a suitable and rational physical correlate of pavement serviceability. Measurement of such a correlate, along with the subjective measurements, would then allow pavement serviceability to be scaled by more rigorous techniques than the methods used for purely subjective estimates. These scaling methods are well described by Torgerson (20) and Guilford (7). Furthermore, available psychophysical evidence suggests that it is unreasonable to apply a subjective rating procedure to the routine measurement of pavement serviceability over a wide area.

It has been recognized for many years that the longitudinal distortion of pavement surfaces is a determining factor with respect to their riding qualities. The empirical relations developed at the Road Test between the subjective magnitudes of serviceability and certain physical measurements of the pavement surface have also demonstrated this point. The measurement of the variance of the pavement slope seems to be a most valuable measurement in this regard. Only limited success has been achieved in characterizing roughness profiles with the aid of such instruments as the BPR roughometer, and the Michigan profilometer. The observed randomness of pavement roughness indicates that some form of statistical characterization is required. Limited applications of spectral density techniques reported by Grimes (6) and Coleman and Hall (3) have met with some success. Further, spectral density functions of road roughness profiles can be fundamentally related to the vibrational environment produced in vehicles excited by the various pavements.

SUMMARY

1. Pavement serviceability is a subjective or psychological phenomenon and must be measured as such. It must be measured on a scale possessing at least interval scale status if the serviceability measures are to be used for statistical correlation with other pavement properties.
2. Existing psychophysical theory presupposes the existence of a judgment, a sensory, and a stimulus continuum. Knowledge of the sensory or psychological continuum can only be achieved by measuring judgments by observers. It is these judgments that are subject to bias and distortion by a variety of environmental factors, thus tending to invalidate the measurements of the psychological experiences.
3. The basic psychometric model underlying subjective rating procedures assumes that raters are capable of making direct quantitative judgments on a linear or interval scale of measurement. Several well known systematic errors are known to distort subjective ratings. These include the leniency error, the central tendency effect, and the halo effect. These errors must be removed from ratings before the basic psychometric model is valid.
4. An important factor influencing the validity of ratings is how well the attribute to be rated is defined. The Road Test definition is very general in nature and appears to have resulted in some confusion with respect to just what attribute is being rated. Pavement serviceability must be defined as the relative riding quality afforded each highway user.
5. A useful objective measure of reliability (reproducibility) of ratings is the reliability coefficient which is defined as the proportion of observed variance that is true variance.
6. A most important influence on the ability of raters to achieve interval scale status is the nature of the cues or descriptive phrases that are used to reinforce the definition of the subjective continuum. In addition to providing anchors, judicious arrangement of the cues may be used to counteract a positive leniency error and the central tendency effect.
7. An important factor concerning the scale format is the number of rating categories used. If insufficient categories are used, then the raters' powers of discrimination cannot be fully utilized. Loss of reliability may result from too many categories. A useful objective criterion of optimization of the number of categories, as well as the cue format, is the reliability coefficient.
8. No explicit provision for testing the ability of raters to achieve interval scale status is contained in subjective rating procedures. Consistency of judgment, or reliability, is an inadequate criterion. The invariance characteristics of scales may be used to test this assumption in the absence of known and measurable physical correlates.
9. Anchoring refers to those conditions that control the origin and unit of the subjective continuum in which raters will report their judgments of magnitude. Experimental studies of this phenomenon have revealed that the unit and origin are not constant but functions of a particular experimental situation. These observations have important implications with respect to subjective serviceability rating at a national level.
10. It is generally considered that subjective rating procedures achieve the measurement status of ordinal scales, and only approach the status of interval scales. The

various methods of scaling and correction allow transformation to interval scale measurements.

11. Although some of the more common distortions and biases to which ratings are vulnerable have been shown to be present in the Road Test ratings, a complete evaluation of the ratings is impossible. The need exists for the design of suitable experiments to evaluate some of the factors concerning scale format, anchoring, etc., that have been described.

12. A more rigorous scale of pavement serviceability cannot be established until a suitable physical correlate of pavement serviceability is established.

REFERENCES

1. Carey, W.N., and Irick, P.E., "The Pavement Serviceability-Performance Concept." HRB Bull. 250, 40-58 (1960).
2. Champney, H., "The Measurement of Parent Behaviour." Child Development, 12: 131-166 (1941).
3. Coleman, T.L., and Hall, A.W., "Implications of Recent Investigations on Runway Roughness Criteria." Meeting, Flight Mechanics Panel. Agard, Paris (Jan. 14-18, 1963).
4. Ghiselli, E.E., and Brown, C.W., "Personnel and Industrial Psychology." McGraw-Hill (1955).
5. Goldman, D.E., "Effects of Vibration on Man." Handbook of Noise Control, Ed. C.M. Harris, McGraw-Hill (1957).
6. Grimes, C.K., "Development of a Method and Instrumentation for Evaluation of Runway Roughness Effects on Military Aircraft." Agard Report 119, NATO (1957).
7. Guilford, J.P., "Psychometric Methods." McGraw-Hill (1954).
8. Helson, H., "Adaptation Level as a Basis for a Quantitative Theory of Frames of Reference." Psychol. Rev. 55: 297-313 (1948).
9. Housel, W.S., "The Michigan Pavement Performance Study for Design Control and Serviceability Rating." Univ. of Michigan, International Conf. on the Structural Design of Asphalt Pavements (1963).
10. Hornick, R.J., "Effects of Whole-Body Vibration in Three Direction Upon Human Performance." Jour. of Engineering Psychology, V. 1: 3 (1962).
11. Hveem, F.N., "Devices for Recording and Evaluating Pavement Roughness." HRB Bull. 264, 1-26 (1960).
12. Johnson, D.M., "The Central Tendency of Judgment as a Regression Phenomenon." American Psychologist, 7: 281 (1952).
13. Murray, H.A., "Explorations in Personality." Oxford Univ. Press (1938).
14. Newcomb, T., "An Experiment Designed to Test the Validity of the Rating Technique." Jour., Educational Psychology, 22: 279-289 (1931).
15. Stevens, S.S., "On the Theory of Scales and Measurement." Science, 103: 667-680 (1946).
16. Stevens, S.S., "Calculation of the Loudness of Complex Noise."Jour., Acoustical Society of America, 28: 5, 807-829 (1956).
17. Stevens, S.S., "Measurement, Psychophysics, and Utility." Measurement: Definitions and Theories, Edited by C. Churchman and P. Ratoosh, Wiley (1959).
18. Symonds, P.M., "Notes on Rating." Jour., Applied Psychology, 9: 188-195 (1925).
19. Symonds, P.M., "Diagnosing Personality and Conduct." Appleton-Century-Crofts (1931).
20. Torgerson, W.S., "Theory and Methods of Scaling." Wiley (1960).
21. Wilkins, E.B., "Pavement Evaluation Studies in Canada." Univ. of Michigan, International Conf. on the Structural Design of Asphalt Pavements (1963).