Quantifying Pavement Serviceability as It Is Judged by Highway Users

Robert J. Weaver, New York State Department of Transportation, Albany

A method for applying scientific psychophysical principles to quantification of the attribute of pavement serviceability is presented. The method has been used in New York to develop, and to serve as the calibrating-recalibrating standard for, the Pavement Serviceability System, which completed a fourth consecutive network survey in the fall of 1978.

In 1963, shortly after the AASHO Road Test, the rationale that had been used there to rate pavement serviceability was critically examined by Hutchinson (1). The concept of serviceability and performance defined for the AASHO Road Test (2) is, of course, beyond challenge; the greatest contribution of that study was to show that serviceability and performance had to be quantified, and a rating system was devised to do it. Hutchinson's critique asserted that, for proper measurement of the attribute of serviceability, the panel-rating procedures used should have incorporated well-established principles and methods of psychophysics and applied psychology. Hutchinson discussed the variation between those principles and what was done, with concern for the potentially serious impacts on the validity of the conclusions about pavement serviceability drawn from Road Test data. He concluded that subjective estimation procedures, typified by Road Test panel ratings, were inappropriate for the task and, further, that they tended to measure pavement distortion and deterioration rather than riding quality, which is the essence of serviceability. He pointed out, however, that much experimental work was necessary to disclose the severity of these impacts and to develop a practical methodology for the measurement of serviceability. Moreover, conduct of that work involved yet another problem—the lack of a precise, measurable physical correlate that varies nearly one-to-one with serviceability. Hutchinson very accurately sized up the task, mapped the direction, and stated the dilemma that has inhibited progress for many years: The first (the method) cannot be found with assurance until one has the second (the statistic), and the second cannot be found with assurance until one has the first.

In 1969, pressure was building among officials of the New York State Department of Transportation (DOT) to find objective, reproducible means of measuring pavement condition and deterioration and to implement the AASHO Road Test serviceability-performance concept in some manner. The objectives were concerned first with pavement design but then grew to include project selection and programming as well. Two years spent in testing all conceivable approaches to measuring serviceability versus physical characteristics did not produce working correlations that were capable of sustained statewide or long-term serviceability measurement as a decision-making data source. Trends were present, but there was intolerable point scatter. Regression only masked the imprecision. The results tended to prove that either (a) people make astonishingly poor "measuring instruments" or (b) highway users are not served by highway pavements in any simple or uniform way according to the definition of serviceability. There was a third possibility—that, as Hutchinson suggested, the panel-rating procedure used over the years was not an effective means of finding out how people are affected by pavement conditions. This is now known to be the case.

The scope of this paper is limited to determining the characteristics of serviceability of pavement test sections as an independent variable. The experimental work required is a panel-rating procedure that somewhat resembles present serviceability rating (PSR) procedures, the principal differences being the larger number of raters required and their instructions. The method of analysis, although considerably more complex than that for PSRs, has vast compensatory benefits. Its principal benefits are that it provides (a) a serviceability scale of great significance to highway users and agencies and (b) the ability to measure serviceability directly with greater accuracy and reproducibility than can be achieved by using the objective measurements generally held to be the physical correlates of serviceability. The method outlined here is the foundation and primary calibration standard for New York's Pavement Serviceability System, which completed its fourth consecutive annual network survey in 1978.

SERVICEABILITY

Serviceability has been defined by Carey and Irick as "the ability of a pavement to serve the highway user" (2). In 1972, Carey expressed concern that "the technology of profile measurement frequently seems to reflect a lack of concern for the basic issue" (3). Hutchinson states, "the serviceability and failure of an engineering design can only be defined relative to the purpose for which a design has been provided " (1).

The purpose of pavements is to provide adequate traveling surfaces for highway users; this, then, must be the object of measuring serviceability as a characteristic of pavements. As Carey and Irick defined serviceability, it was to be a measure of the one aspect that had no prior measure—the effect of present pavement deterioration on the immediate level of service the pavement provides to the highway user. Their concept held that this variable had to be measured because the changing value of serviceability with time is the only means of properly quantifying performance. Moreover, it was held that serviceability was a subjective variable that had to be measured by the human instrument.

It is a simple fact that any physical measurement of profile distress generally increases as visual deterioration of the pavement proceeds and that serviceability decreases on a roughly parallel course. Measurements are not needed to prove the obvious but rather to distinguish exact relations that are not obvious. As Carey (3) states, performance rests on serviceability, serviceability depends on surface profiles, and this is the basis for pavement evaluation. Thus, pavement evaluation begins with independent measurement of serviceability, and that is possible only by applying psychophysical principles to discover how pavement condition affects pavement users.

PSYCHOPHYSICS

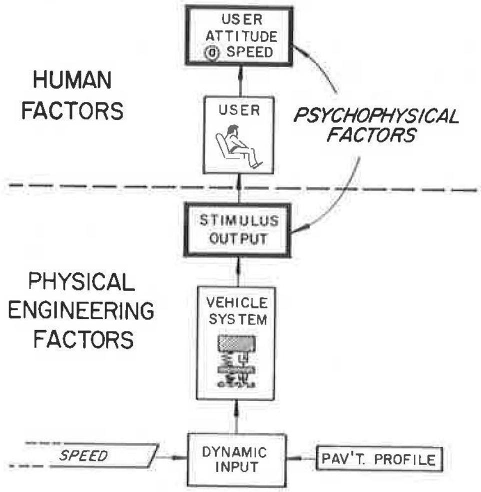
Psychophysics is the science of uncovering fundamental quantitative relations between human responses and applied physical stimuli. The techniques are quite powerful and widely used in experimental psychology, marketing research, and product development. How pavement condition affects the general highway user is exactly the kind of situation in which psychophysics has application.

In summary, we have learned that, although human sensory systems reacting to such stimuli as sound, taste, smell, feeling, and vibration are indeed marvelous "instruments", the ratings obtained from the person about what he or she felt, heard, saw, or tasted make very poor readouts. To give a rating, we must construct it in our conscious minds. While we are doing so, our concentration varies, we have trouble remembering what was sensed, distractions intervene, and random thoughts intrude. It is typical for ratings by any panel of raters to range over half to three-fifths of a rating scale. This is not something peculiar to PSR results; it is the way all human beings function in rating any physical experience. Psychophysical techniques provide many ways of stripping away this superficial variability to get at the true response of all people for all magnitudes of a stimulus.

Many engineers are familiar with how an electronic signal can be buried in random noise during transmission but then extracted from the noise unchanged after being recorded. The techniques of signal analysis used to extract the signal from the random electrical noise make a fine analogy to psychophysical analyses as they are used to extract the true human response from what can be thought of as random "rater noise". In both cases, a record is needed of all possible values of (signal + noise) before analysis is possible. Taking the mean of (signal + noise) at any time and place can give one only a mean of the signal and noise combined, at that time and place. That is what a PSR represents. Consequently, if one wishes to measure serviceability, there are two stringent experimental requirements, neither of which is met in conventional PSR practice:

1. A wide range of test sections, encompassing the full range of serviceability to be experienced, to differentiate small differences adequately as well as to define a complete value scale of serviceability; and
2. Approximately 60-80 persons functioning as raters for adequate definition of the rater-noise factor across the entire experiment.

Figure 1. Serviceability model.



The true serviceability value of a given test section cannot be found from human ratings in any other way.

PSYCHOPHYSICS APPLIED TO SERVICEABILITY

Road Test Ratings

At the AASHO Road Test, variability, or noise, in human ratings was treated as abnormal. Because the whole experiment depended on measuring serviceability, a crash effort went into making PSRs reproducible through training, discussions, and continued trials. Those who participated soon learned to converge to a "right" rating arrived at by consensus. This unfortunate development ran exactly counter to psychophysical principles and introduced an irreversible bias that seemed to Hutchinson the most serious deficiency in the conclusions. Under such conditions, the individual rater is almost completely distracted from stimulus response, drawing into his or her evaluation any and all attributes of the pavement and even guessing what other raters are going to do. In psychophysical terms, these types of bias are, respectively, (a) autokinetic or central tendency and (b) the halo effect. The PSRs of the AASHO Road Test were made more numerically reproducible but were biased even farther away from serviceability.

Serviceability Model

For either psychophysical scaling of serviceability or evaluating which physical measurements of pavements relate to serviceability, one needs a comprehensive model. The model developed shows how the important human and physical factors interact (see Figure 1).

The stimulus output from the vehicle is the interface between these two groups of factors. Travel over a test section provides a definite and reproducible stimulus to the rater or user that varies with travel speed. When road users slow down because of bad pavement conditions, they are unconsciously using the speed relation to raise serviceability to a more tolerable level. In the new methodology, rater instruction provides a clear concept of serviceability as a "travel experience, on this pavement at this speed". A fixed travel speed must be established for all raters for each test section in an experiment. The stimulus is a complex function of travel speed and vehicle characteristics. One might conclude that the obvious differences in the ride provided by various types of vehicles and the resulting sensitivity to vehicle characteristics produce prohibitive complications. In the absolute sense—input-output transfer functions—this is the case. However, this is where human psychology enters in. Both raters and users judge the speed-profile input from a single vehicle—the one in which they are riding. After a period of acclimation to a strange vehicle, raters' discrimination of pavement serviceability is easily separated in their own minds from that portion of the ride experience affected by vehicle type and quality.

The model shown in Figure 1 depicts human perception of pavement serviceability as a vibratory experience in which travel speed is a major variable. A given profile will thus be perceived differently in terms of serviceability as travel speed varies. This speed does not enter into the analysis of ratings but is important in determining physical correlates of serviceability.

Physical Correlates of Serviceability

The model is equally useful in evaluating physical measurements that are likely to vary closely with serviceability. Although that subject is beyond the scope of this paper, a few comments are necessary.

Historically, PSR-based serviceability has focused on measures of pavement distress rather than on how that distress affects the serviceability provided by the pavement at various travel speeds. Consequently, when the output of response-type devices has been found to be sensitive to speed or vehicle characteristics, there has been no way to treat those facts except as shortcomings of the device or problems that required standardization. The serviceability model shows rather clearly that any technique of profile characterization that is not vehicle mounted and speed sensitive is bound to correlate very poorly with the highway user's perception of pavement quality and hence with psychophysically scaled serviceability.

BASIC DIFFERENCES IN RATING METHODS

The new rating methods were evolved deliberately to follow the original AASHO Road Test scheme as closely as possible. For instance, raters use a 0-5 scale with five main categories and express their judgment by a mark on this scale. The mark is then scaled as a number and referred to as the individual serviceability rating (ISR). In the psychophysical application, however, each ISR is only a tool that is used to determine how all raters rated the serviceability of each section in relation to that of all other sections. PSR, on the other hand, takes each ISR as an absolute number without regard to any question of relativity among raters or sections. In fact, in one of the classical psychophysical methods—paired comparisons—each rater merely determines which of a pair of items is better or worse. Each rater, however, must make n(n - 1)/2 pair comparisons for each of n stimuli included in the experiment, which results in an astronomical number of total judgments.

Pavements cannot, of course, be presented in pairs for simultaneous comparison. But if they could be, the final scale value of each section determined by this procedure would be identical to the scale value determined by using the successive categories. In psychophysics, this has been established to be the case whenever a reproducible physical stimulus is involved. The difference between these two psychophysical methods for 88 test sections and 79 raters is merely one of experimental efficiency and analytical simplicity—6952 judgments versus 302 412. Thus, although the numerical rating was made to look like an AASHO Road Test ISR, it was not obtained in the same way, and the two are interpreted entirely differently to arrive at the serviceability of a test section.

RATING EXPERIMENTS

The overall rating experiment must completely embrace all the major variables of pavement profile condition, travel speed, and pavement type (rigid, flexible, and flexible-over-rigid overlays). Properly correlating these variables with collateral physical measurements requires a minimum of about 90 test sections. No fewer than 60 raters should be considered, and the more the better up to about 100. Crisp, efficient execution of such a large experiment requires a great deal of careful preparation and attention to detail. Detailed methods and procedures are available elsewhere (4) to amplify this outline of major considerations.

Selection of Test Sections

Important psychological factors in rating experiments are as follows:

1. The standard maximum length of test sections is 0.8 km (0.5 mile); the minimum length is 396 m (1300 ft).
2. Homogeneity in the nature of the stimulus over the section is desirable.
3. Rapid transition from one section to the next is required; circuit layout should average 10 or more sections/h/team.
4. Atypical surroundings that could distract the rater should be avoided.

Selection of Raters

Raters should be selected for their ability to make sincere, independent judgments and to follow simple instructions. Nondrivers and engineers who are accustomed to studying pavement distress as a means of judging serviceability should not be selected.

Rater Instruction and Scale Anchoring

The objectives of rater instruction are to explain the attribute of serviceability that is to be rated and the scale for doing so, to discuss the process of rating, and to clear up other details. No attempt should be made to train raters to the point of actually rating specific test sections, and raters should understand that their ratings will not be compared directly with those of any other rater. All instruction lectures should start by clearing up the normal concerns of most persons about the schedule, such as stops, lunch, and return time, and anything else even remotely likely to be of concern. Since many teams of raters are necessary, the entire lecture should be given uniformly to all raters from a prepared text. Raters must be told that all ratings are confidential and that they are not permitted to discuss ratings or pavements or even to comment on a past rating experience among themselves during their tour.

"Anchoring" the rating scale is an important part of instruction. For every relation between human response and a physical stimulus, magnitudes of the stimulus exist beyond which a change in stimulus has no proportional change in response. These two points on the stimulus scale are the liminal points, or "limens", of the relation and, of course, the ends of the rating scale. It is to these two limens that scale anchoring is directed, and this is done by describing a liminal serviceability experience to the raters to correspond to the end labels, as follows:

1. Perfect—"At the travel speed, this experience is so good that you doubt whether you could detect any improvement even if the pavement were made smoother."
2. Impassable—"At the travel speed, this experience is so bad that you feared you or the vehicle would not make it in one piece to the end of the test section."

Without this anchoring, defined relative to the limens of the serviceability attribute, the labels alone remain vague, floating references for each rater's interpretation.

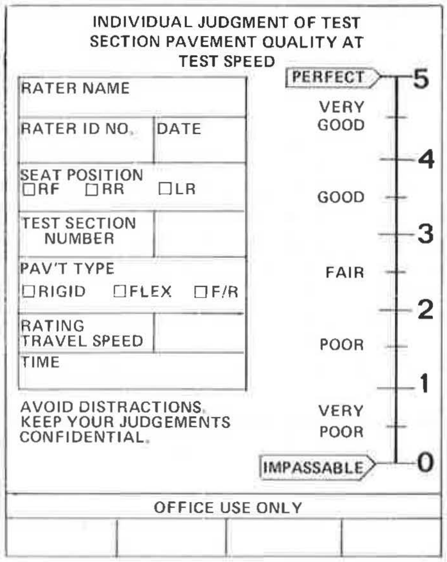
Raters must believe their judgments to be exact. The instructions must ask for exactness. To arrive at an exact judgment, raters generally take three iterative mental steps. They contemplate the two scale ends and the indifference point (midscale), which readily places their judgment to within a one-third portion of the scale. Then they consider the category descriptions—good, fair, and so on—and this places their judgment to within a one-fifth or smaller portion of the scale. Raters usually need help in making the third and final iteration. Left alone, they will obtain help unconsciously by letting their attention drift toward other attributes (distractions). To keep raters concentrating on serviceability, the instructions should provide them with a mirror image of serviceability—the nature and intensity of their feelings about the need for rehabilitation. They should be invited to consider how far from indifferent they rate the highway agency's performance in providing adequate traveling surfaces, as if the test section were a standard.

The rating form used, which is shown in Figure 2, is similar to a PSR form, but the question of "acceptability" that is asked separately on conventional PSR forms is noticeably absent. This is a logical second question for raters in rating the attribute of pavement distress and distortion, but serviceability, properly measured, is acceptability.

Teams and Vehicles

Rating vehicles must be driven by a nonrating staff member who knows the exact itinerary and test speeds and can supervise the rating procedure. A team of three raters is assigned as a group. To avoid possible distractions, males and females are not teamed together. The team and seating position remain fixed on all rating days. On different days, a different vehicle and driver are assigned to each team, but no such changes are permissible in a single day. There are no special requirements for vehicles other than good mechanical condition and reasonable comfort.

Figure 2. Rating form.



INDIVIDUAL JUDGMENT OF TEST SECTION PAVEMENT QUALITY AT TEST SPEED

RATER NAME

RATER ID NO. DATE

SEAT POSITION
□RF □RR □LR

TEST SECTION NUMBER

PAV'T TYPE
□RIGID □FLEX □F/R

RATING TRAVEL SPEED

TIME

AVOID DISTRACTIONS. KEEP YOUR JUDGEMENTS CONFIDENTIAL.

PERFECT 5

VERY GOOD

4

GOOD

3

FAIR

2

POOR

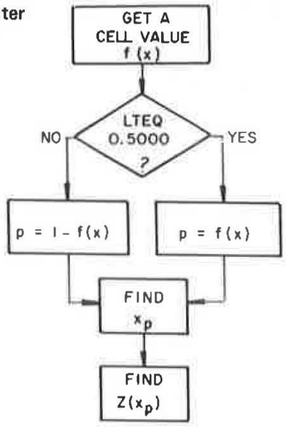
1

VERY POOR

IMPASSABLE 0

OFFICE USE ONLY

Figure 3. Flowchart of computer program basis for Z(x) matrix.



PSYCHOPHYSICAL ANALYSIS OF RATING RESULTS

The analysis for serviceability scale values (SSVs), the final result, starts by taking all raters and pavements in a single matrix. Given 90 test sections and 80 raters, the resulting 7200 ratings are placed in a 90x80 matrix and then transformed into a 90x10 matrix for 10 class intervals on each test section. The 7200 ratings are thus reduced to 900 observations of class-interval frequency and cumulative frequency. Individual raters or ratings are no longer of interest. The only other figure carried forward from the rating experiment is the total number of raters. The calculations involve conversions into five more successive 90x10 matrices and finally to five 20-element tables before the SSV is found for each test section.

The analytical procedure is a straightforward but lengthy process. It cannot be described by an equation, but it has been reduced to a purely step-by-step clerical task, as follows:

1. Assemble a two-dimensional matrix (matrix 1) of individual ratings by test section number and rater number. Scale the rater's marks to the nearest 0.01 division and enter.
2. Select class intervals to be used in converting the original matrix to a matrix of test section versus frequency in class interval. Use the smallest class interval possible, but they must generally (not necessarily completely) be large enough to avoid zero-frequency intervals embedded between those with substantial frequency counts. The size of class interval permitted becomes smaller as the number of raters becomes larger. Prepare this matrix (matrix 2). A class interval of 0.50 is adequate for 60-80 raters (divide the rating scale into 10 intervals).
3. Fill the blank matrix 2 with a fraction in which the numerator is taken as the frequency of rating in each class interval and the denominator as the cumulative frequency up to and including that class interval (reckoned from the low to the high end of the rating scale).
4. Prepare a new matrix (matrix 3) of test section versus class interval. Fill the cells of this matrix by dividing the cumulative frequency (the denominator in the matrix 2 values) by the total number of raters. The result is a decimal value of cumulative proportions.
5. Prepare a new matrix (matrix 4) of test section versus class interval. Convert the cumulative proportions in each cell of matrix 3 to the normal probability function Z(x) for each cell of this matrix. This requires a statistical table that gives Z(x) in terms of P(x) and Q(x) (5) or a computational algorithm such as that described below (see Figure 3): To find x_p, the following rational approximation for $Q(x_p) = p$ is used. The maximum error is $|\epsilon(p)| < 4.5 \times 10^{-4}$, which is quite adequate for the purpose. The calculation involves the following steps:

Find $t = \sqrt{\ln (1/p^2)}$.
Find $x_p = t - [c_0 + c_1 t + c_2 t^2)/(1 + d_1 t + d_2 t^2 + d_3 t^3)]$, with

the following values for constants: $c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, and $d_3 = 0.001308$.
Take the result x_p and find $Z(x_p) = (1/\sqrt{2\pi})\, e^{-x^2/2}$, where $e = 2.718282$.

Test values for the check of the module are given below:

f(x)	t	x_p	$Z(x_p)$
0.2000	1.794 12	0.841 45	0.279 96
0.4500	1.263 73	0.125 38	0.395 80
0.5500	1.263 73	0.125 38	0.395 80
0.8400	1.914 46	0.994 42	0.243 31

6. Prepare a new matrix of test section versus class interval (matrix 5). The cell value for this matrix is determined by the Z(x) value in the corresponding cell of the completed matrix 4, which is subtracted from the cell value immediately to the left in that matrix. The result, including the algebraic sign, is entered in the new matrix.

7. Prepare a new matrix of test section versus class interval (matrix 6) to be the matrix of frequency proportion. Use the numerator in matrix 2, divide it by the total number of raters, and enter the result in this new matrix.

8. Prepare a new matrix of test section versus class interval (matrix 7). Obtain cell values by dividing the cell values of matrix 5 by the corresponding cell values in matrix 6.

9. Prepare a table (referred to here as Table A) with class-interval columns and two rows. Label the rows A and B. The value for each element is obtained from matrix 7: for row A, the sum of all elements that have nonzero values in cells immediately to their right, and for row B, the sum of all elements that have nonzero values in cells immediately to their left. (In computing this element, keep count of the number of values used for each entry for use in step 11.)

10. Prepare a new table (Table B) by class-interval columns with a single row of elements. The elements to be computed for this table are the resultant values (with sign) of adjoining intervals in Table A that have nonzero entries in both rows A and B (see Figure 4).

11. Prepare a new table (Table C) that is similar to Table B. This is the mean of differences between intervals. Divide the value in Table B by the number of cell values used to obtain the value in line B in Table A.

12. Prepare a new table (Table D) that is similar to Tables A and B. This will be used to compute cumulative interval values. Proceeding from the lowest interval, the value to be entered in each cell is the summation of the cell value in Table C with all cell values to the left.

13. Prepare a new table (Table E) that is similar to Tables A, B, C, and D. The entries into this table are obtained from Table D. Take the value in Table D for the interval that contains the original midpoint of the rating scale, subtract that value from each interval value in Table D, and enter the result (with sign) in Table E. This is the final result of the analysis. This table of values and the matrix of frequency proportion (matrix 6) are the tools to use in computing the scale value of each test section in the experiment.

The scale value of a test section is computed as the sum of products of each cell in matrix 6 multiplied by the corresponding value in Table E. The serviceability of each test section is computed in this way. These scale values have zero as a point of indifference, which is the most important feature of the scale. Positive values indicate increasingly positive highway user attitudes toward pavement serviceability at the rating speed. Negative values conversely indicate increasingly negative attitudes.

The positive-zero-negative scale values can be used as is for a serviceability scale or transformed to any other linear scale. For instance, 0 to 100 scale ends, 0.00 to 5.00, -250 to +250, and so on. The technique easily reproduces a 500-unit scale but not a 1000-unit scale.

The object of this procedure is to analyze frequency distributions, frequency proportions, and probability functions to establish the scale of user response and finally to determine where each test section belongs on it. The ISRs and mean ISRs shown in the remaining figures in this paper (Figures 5-10) were computed from the original ratings especially to show comparisons with SSVs. We chose to transform the positive-negative scaling back into a 0.00-5.00 scale to emulate the AASHO Road Test (this has been the practice in New York). On the transformed scale, the indifference point is at 2.40; experience has shown that such a transformation is best performed by adding 2.40 to all values on the psychophysical positive-zero-negative scale.

RESULTS

The validity of the new rating methods is best examined independently from all physical factors to show how ISRs, test sections, SSVs, and time relate. The evidence was acquired by computing mean ISRs on test sections of full-scale experiments. The main data set was taken from a May 1977 experiment to recalibrate New York's Pavement Serviceability System. This experiment contained 6952 judgments (ISRs) by 79 raters on 88 test sections.

The results shown in Figures 5-10 are in no way comparable to PSR although individual raters' ratings and mean ratings of panels of various sizes are shown in addition to the single output of the new method—the underlying true serviceability scale value found for each test section included in a proper experiment.

An ISR from this experiment is not the same as the judgment of a single rater making a PSR because PSR rater instructions psychologically incapacitate the rater for judging serviceability. The PSR rater believed he was rating a pavement and either gravitated toward or was pushed into a subjective evaluation of a different

Figure 4. Formulation of Table B in analytical procedure.

Portion of Table B

0-0.49	0.50-0.99	1.00-1.49
-	0	0.932

Portion of Table A

	0-0.49	0.50-0.99	1.00-1.49
A	0	-2.072	-5.059
B	-	0	-1.140

(-1.140)-(2.072) = 0.932

Figure 5. Mean ISR versus SSV for 79-member panel.

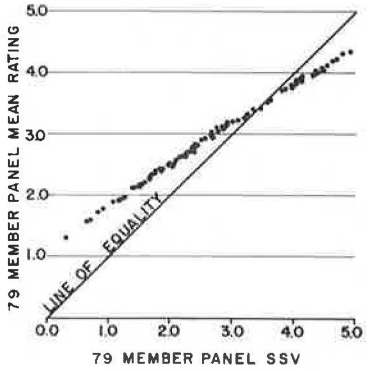


Figure 6. Mean ISR versus SSV from 1972 experiment.

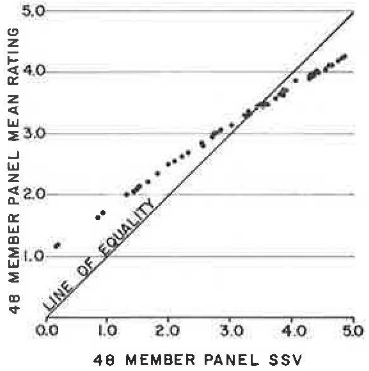
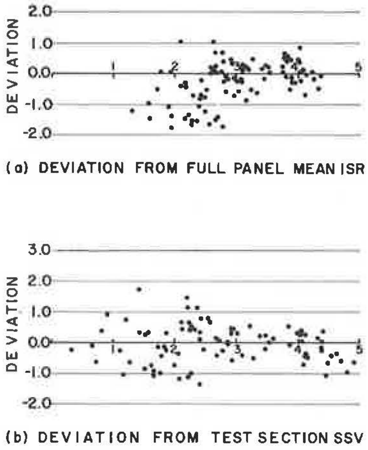


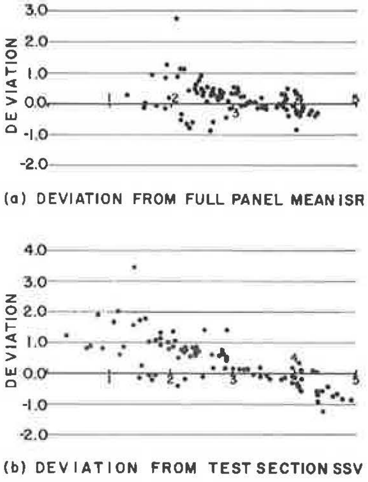
Figure 7. Functioning of typical rater as one of 79-member panel.



(a) DEVIATION FROM FULL PANEL MEAN ISR

(b) DEVIATION FROM TEST SECTION SSV

Figure 8. Functioning of single "expert" rater as one of 79-member panel showing bias toward rating the visual condition of the pavement.



(a) DEVIATION FROM FULL PANEL MEAN ISR

(b) DEVIATION FROM TEST SECTION SSV

characteristic of a pavement—i.e., surface distress and distortion. Further, he knew his ratings would be compared and he could be "odd man out", so he avoided extreme ratings. His rating scale of 0-5 or 0-10 did not have its ends anchored, i.e., defined relative to either the attribute of serviceability or surface distress. Speed was not a controlled variable on a test section

Figure 9. Small-panel versus full-panel mean ISR of test section.

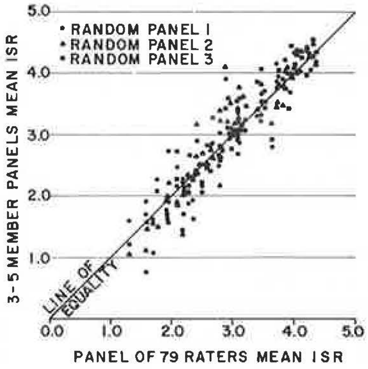
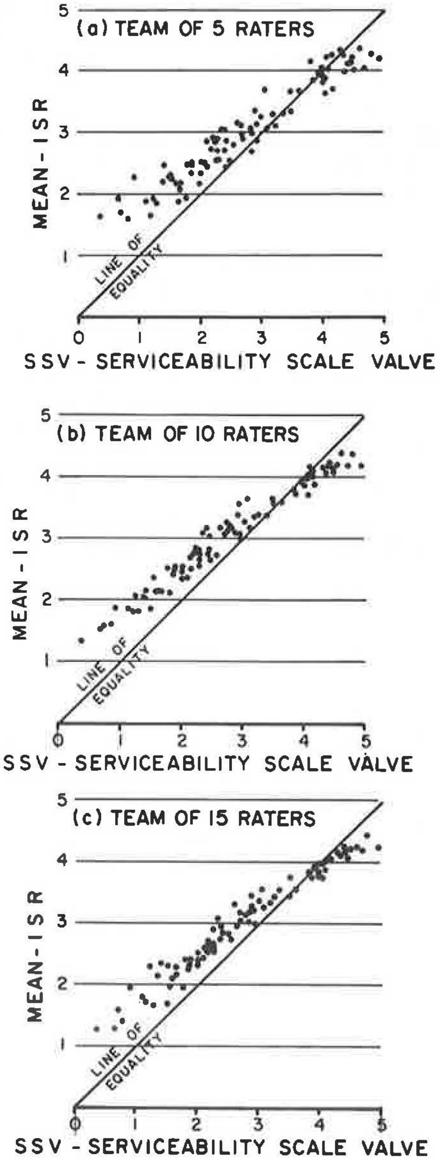


Figure 10. Effect of team size on mean ISR versus SSV of test sections.



for all raters. These are a few of the differences that have led to the strange variability encountered in PSR work. They are also reasons why one cannot retrofit the new analysis method to old PSR data. Those data are psychologically and psychophysically invalidated.

Mean ISR Versus SSV for Each Test Section

Data from the May 1977 psychophysical experiment can show how mean ISR relates to serviceability and provide some insight into the characteristics of ISRs as rater noise. Figure 5 shows that, given enough rater noise (as in this experiment), the mean ISR converges to a curve that has a clear relation with the value scale of serviceability. The mean ISR is only the mean of the signal plus rater noise on a test section, and the SSV is the signal. An SSV of 2.40 is the indifference point. The highway user becomes angry at an SSV under 1.00. If the user were to maintain speed at an SSV of zero, his or her vehicle would probably become an accident statistic because zero indicates impassable at the travel speed. One can see here how misleading the mean ISR is because of rater noise operating to obscure the true human response signal. One can also see that there is a clear—almost perfect—relation. The equation used is that of a second-order polynomial, as follows:

$$\mathrm{SSV} = c + b\bar{x} + c\bar{x}^2 \tag{1}$$

where $\bar{x}$ = the mean ISR of a large panel of n raters. The correlation statistics for the 1977 experiment are given below:

Statistic	1977
Number of test sections	88
Number of raters	79
c-intercept	−1.3369
b	
Coefficient	1.1974
Standard error	0.0364
a	
Coefficient	0.0523
Standard error	0.0060
F-value	52 946

It can be seen that the mean ISR and the corresponding SSV are far from numerically equal. One reason is that the rating procedure has a scale with fixed ends but the underlying human-value scale has no pre-established ends. When one senses an experience, one's psychological response will be found in the "white noise" that surrounds the underlying (and as yet unknown) scale value found by the analysis. The essence of psychophysical analyses is that the noise has a Gaussian distribution around the underlying scale value, not the mean rating scale value. The constraint introduced by the fixed-end rating scale (where the subject communicates his or her rating) is that obviously only a midscale experience has even the opportunity to display a Gaussian distribution. The end points act as sources of rater bias, causing a distortion in the distribution as the serviceability of a test section approaches the end points. As Figure 5 shows, even with the relatively bias-free nature of the ISRs in this experiment, no section mean can clear the bounds of the 1.5-4.4 range typical of PSR.

Reproducibility

The ultimate test of reproducibility would appear to be that the exact relation (Figure 5) be invariant across separate experiments: a different set of test sections, a different group of raters, and preferably a different point in time. For such a test, data from the earliest psychophysical experiment, which was conducted in July 1972, were analyzed for the relation between mean ISR and SSV. Although that particular experiment used only 41 raters and 48 test sections, the same relation resulted (see Figure 6). The regression results for the 1972 experiment are given below:

Statistic	1972
Number of test sections	48
Number of raters	41
c-intercept	−1.1297
b	
Coefficient	1.0822
Standard error	0.0430
a	
Coefficient	0.0755
Standard error	0.0075
F-value	27 509

Both regressions show the relation to be that of a second-order polynomial. The following equation was computed from the full set of 1977 data:

$$\mathrm{SSV} = -1.3369 + 1.1974\bar{x} + 0.0523\bar{x}^2 \tag{2}$$

where $\bar{x}$ is the mean ISR of a sufficiently large number of raters.

The primary usefulness of the equation appears to be as a means of direct comparison of the results of psychophysical experiments used to calibrate systems at different places and times. The equation can also be used as an analytical shortcut should one for some reason require a serviceability measurement at a predetermined travel speed on a few pavements. Such use would still require, however, that a large enough number of persons (60-80) be used for a mean ISR and that rater selection, instruction, and other details be adhered to without exception.

Rater Randomness

In Figure 7(a), the data of a typical rater show interesting trends relative to both mean ISR and SSV. The rater shows typical randomness with respect to the panel mean and a slight tendency to rate more negatively. When the final SSV is determined, however, the same rater's randomness is around the SSV [Figure 7(b)].

It was intended that the panel would not include any "experts" who would be biased by cracks and the appearance of the pavement. Fortunately for this study, however, one professional—a materials engineer who had had experience in conventional pavement evaluation—did join the panel as a last-minute substitute, undetected. His ratings, which are shown in Figure 8, indicate a bias not seen in the ratings of other raters. The trends are those to be expected from one who, consciously or unconsciously, considers cracks and visual condition. His skew from zero SSV shows that an excellent pavement with cracks is downgraded, and any bad pavement is often severely overrated because it appears better than it feels. If this rater had been blindfolded, he undoubtedly would have functioned like the remaining 78.

It would be interesting to compare SSVs from a panel of 79 experts with those for this panel since this type of bias is one of the disturbing consequences of AASHO Road Test PSRs detected by Hutchinson. But one can only speculate, for there is no way to retrofit the AASHO Road Test PSRs. Nevertheless, it is clear that the attribute of pavements called "cracks" is not the attribute defined as serviceability any more than are lane width, skid resistance, roughness, curvature, cross-slope, and sight distance.

Mean of ISRs with Small Panels

It has been shown how enough raters will cause the mean rating to converge to a fixed number on each test section. It is illuminating to show how well small panels can reproduce the mean of a 79-member panel on each test section. Random numbers were used to select three 5-member teams from the panel of 79, and the mean of ISRs on each test section was computed. The results are shown in Figure 9. The central tendency toward the equality line is clear, but the full-panel mean for any test section could not possibly be identified from any combination of data from these 15 raters.

To examine this question further, the three panels are used to make 5-, 10-, and 15-member panels. These mean ISRs versus the SSV determined from the full panel of 79 are shown in Figure 10. It can be seen that each addition of 5 new members draws the results slightly closer to a linear relation identical to those shown in Figures 5 and 6. At the same time, it is also clear that substantial convergence will not occur until a very large number of raters are included.

CONCLUSIONS

The serviceability of a pavement profile as perceived by the highway user can be precisely measured. This involves finding basic human responses to the range of physical stimuli generated by travel speed and pavement conditions. The principles of psychophysics have provided a solution to the measurement problem. Panel ratings have long been regarded as the most nebulous, controversial, and irreproducible means of evaluating pavement serviceability. However, when they are properly devised, conducted, and analyzed, panel ratings can measure serviceability with more precision and reproducibility than the common devices used to measure the pavement profile.

The method presented in this paper produces test-section SSVs that are shown to be independent of time, place, and differences in rating panels. This method differs in many ways from the classical AASHO Road Test PSR procedures. Unfortunately, the nature of PSR methods precludes any retrofit of AASHO Road Test data.

Requirements for application of the new methodology are as follows:

1. The raters should be 60-80 ordinary highway users, each rating all test sections
2. Rater instruction, which differs considerably from the PSR type of instruction, must firmly anchor the end points of the rating scale and fully describe the attribute of serviceability.
3. Travel speed is a vital serviceability variable. Each test section must maintain a defined travel speed for all raters, thus constituting not merely a profile but a "travel experience" whose serviceability is rated. The SSV that is computed is thus the serviceability of a test-section profile at its rating speed.
4. The method of analysis encompasses the entire set of judgments and test sections from its outset.
5. Combinations of test section and speed must provide a full range of serviceability experiences as well as profile conditions.

This method was developed to obtain a precise measure of serviceability on test sections at any time or place. It is then possible to calibrate (or recalibrate) vehicle-mounted profile-measuring systems whose output is to be correlated with scalar serviceability at the posted travel speed. Obviously, this method is also of great value in determining the merits of various profile-measuring systems.

The interplay of profile, speed, and serviceability variables has also been found to be sensitive to three different types of pavement: rigid, flexible, and flexible-over-rigid overlay. For these reasons, approximately 90 test sections (six serviceability levels × three pavement types × five speeds) are required in a single experimental plan. Independently of the ratings, the profile-measuring systems obtain outputs at all five speeds on all test sections (only the ratings are dealt with in this paper).

Application of this methodology admittedly involves a great deal of work, but it has the advantage of turning the serviceability-performance concept from a somewhat abstract idea into a powerful reality with precise results. A standard of serviceability exists that can be mobilized as needed by the psychophysical method presented here. The results measure serviceability exactly as it was originally defined—how the pavement serves the highway user.

ACKNOWLEDGMENT

Without the encouragement of W. R. Hudson of the University of Texas at Austin to reexamine available psychophysical data and find a way to present this subject to others, this paper would not have been prepared. I am also indebted to Robert O. Clark, Jr., and William F. Rever, whose dedicated and painstaking efforts over the years were the critical difference between ideas and their fruition, not only on this subject but also in many other developments that have led to the successful operation of New York's Pavement Serviceability System.

REFERENCES

1. B. G. Hutchinson. Principles of Subjective Rating Scale Construction. HRB, Highway Research Record 46, 1963, pp. 60-70.
2. W. N. Carey, Jr., and P. E. Irick. The Pavement Serviceability-Performance Concept. HRB, Bull. 250, 1960, pp. 40-58.
3. W. N. Carey, Jr. Uses of Surface Profile Measurements. HRB, Special Rept. 133, 1973, pp. 5-7.
4. R. J. Weaver and R. O. Clark, Jr. Psychophysical Scaling of Pavement Serviceability. Soil Mechanics Bureau, New York State Department of Transportation, Albany, Manual SEM-9, May 1977.
5. Handbook of Mathematical Functions. National Bureau of Standards, U.S. Department of Commerce, Applied Mathematics Series 55, 1970, p. 975.

Publication of this paper sponsored by Committee on Pavement Condition Evaluation.