Establishing Relationships Between Pavement Roughness and Perceptions of Acceptability

ARUN GARG, ALAN HOROWITZ, AND FRED ROSS

A psychological scaling experiment was conducted in Wisconsin to establish relationships between pavement roughness and users' perceived need to improve the road. A total of 32 road segments were selected for user evaluation. Except for their surface, they had very similar characteristics (speed limit, length, terrain, traffic volumes, scenery, etc.). Physical roughness was measured with both a response-type instrument (roadmeter) and a profilometer. Fifty paid subjects were selected randomly from the general population. They were asked to rate ride quality on both the traditional Weaver/AASHO categorical scale and on a newly designed magnitude estimation scale. In addition, subjects were asked, using a Likert scale, about their willingness to resurface and were asked to estimate the amount of extra time they would be willing to spend to avoid a particular segment, considering its roughness. The experiment yielded several useful mathematical relations between physical roughness and users' willingness to resurface. It was found that the magnitude estimation scale was preferable to the Weaver/AASHO scale for measuring subjective roughness. Surprisingly, the roadmeter was better than the profilometer for measuring physical roughness.

For the most part state DOTs assign dollars for pavement resurfacing on the basis of the statistical distribution of roughness across the highway system, political considerations, and budgetary limitations, rather than on rigorous consideration of highway users' satisfaction. There is a consensus of previous studies that the definition of a roughness standard should reasonably be guided by the degree of user satisfaction or dissatisfaction that can be expected at any particular level of roughness (*1–5*). Indeed, several studies have established a relationship between mechanical measures of roughness and the percentage of users saying that the road should be resurfaced (*1–3, 6, 7*). From the public's viewpoint pavement roughness, more than structural adequacy, drives the desire for pavement improvement.

In Wisconsin, the present serviceability index (PSI) is used to establish a standard for pavement roughness, called a "terminal" roughness level. The terminal roughness level is defined as the roughness level (expressed in PSI) at which a pavement is considered to be deficient and hence in need of improvement. The terminal levels in Wisconsin are 2.5 for the Interstate system, 2.25 on principal arteries, and 2.0 on other roads. PSI, in Wisconsin, is determined by converting the output of a roadmeter (a response-type instrument that yields inches per mile) to a 0 to 5 scale.

The objectives of this study were to establish more precise relationships between pavement roughness, user satisfaction with ride quality, and the perceived need to improve the road (willingness to incur costs to make the pavement smoother). In order to meet these objectives, a psychological scaling experiment was conducted. Fifty Wisconsin drivers, selected randomly, were asked to rate 32 road segments. Rather than being representative of all Wisconsin roads, most test segments were selected to have PSIs of 1.0 to 4.0. The segments were chosen to be similar in length, speed limit, terrain, traffic volume, and scenery. Subjects rated several different aspects of ride quality on both traditional scales (such as the Weaver/AASHO scale) and on scales specifically designed to achieve a better understanding of terminal roughness.

PSYCHOLOGICAL SCALING ISSUES

Of most interest here are automobile users' perceptions of pavement roughness. Although the manner in which human beings rate pavement roughness is necessarily an empirical problem, the known facts of psychophysics set certain valuable guidelines. An observer is sensitive not only to the physical stimuli he or she is trying to measure, but also to a large number of other factors that can distort judgment to varying degrees (*5*). This makes the task of subjectively measuring ride quality more difficult, although it is still quantifiable. Since human observers are susceptible to external influences in communicating their psychological impressions, most psychophysical studies use a scale to measure psychological experience and relate this to physical measurement. Then, psychological measurements can be estimated from measurements of the physical correlate. The original PSI (*7*) was developed within this type of framework. However, it appears that this subjective measurement procedure was developed without full cognizance of the basic principles of subjective rating scale construction (*5*).

Although some attempts have been made to correlate ride quality (a subjective measure) with pavement roughness (an objective measure), little information exists to define this relationship. Nearly all studies (*2, 3, 6, 7*) have used a category rating scale, usually the Weaver/AASHO scale, which was the original basis for PSI. Most of these scales suffer from

A. Garg and A. Horowitz, College of Engineering and Applied Science, University of Wisconsin—Milwaukee, P.O. Box 784, Milwaukee, Wis. 53201. F. Ross, Division of Highways and Transportation Services, Wisconsin Department of Transportation, 718 Clairemont Avenue, Eau Claire, Wis.

the error of leniency, the halo effect, and the error of central tendency (*5*). The Weaver/AASHO scale uses five categories, and there is additional concern that the scale does not effectively use a subject's power of discrimination. Hutchinson (*5*) suggests that cues of a very general character such as "excellent," "poor," etc., should be avoided. Further, if the scale is to be manipulated mathematically (as in ride quality studies), one must be able to measure psychological ratios (ratio scale) or at least differences (interval scale). In other words, a linear relationship should exist between the different sets of scale values. Hutchinson (*5*) suggests that such a relationship is not possible with the Weaver/AASHO scale. Based on a study of different scales, Holbrook (*1*) suggested that for predicting ride quality from physical measures of pavement roughness, magnitude estimation scales are preferred over categorical scales. This suggestion has not, as yet, been rigorously tested.

In a classical magnitude estimation experiment subjects rate a series of comparative stimuli as a fraction or multiple of a given rating for a single, standard stimulus (*8*). For example, a subject may be asked to rate the brightness of lights. The subject is first presented a standard amount of light and is told that this amount of light has a rating of 10. The subject is then presented a comparative stimulus. If the subject thinks that the light is one-half as bright, it would be rated "5." Conversely, if the subject thinks that the light is twice as bright, the rating should be "20." Thus, a magnitude estimation scale has a minimum value of zero and a maximum value of infinity. Extensive tests of magnitude estimation scales have demonstrated that they possess the ratio property.

Magnitude estimation is best implemented in a laboratory where both standard and comparative stimuli can be alternately presented to subjects in rapid succession. This procedure cannot be implemented on a road course because it is not possible to find a sufficient number of identical road segments to serve as standard stimuli. An alternative procedure, adapted for this study, is to first train the subjects about characteristics of the standard stimulus by repeated exposure. Then, subjects are presented several comparative stimuli, again presented the standard stimulus, again presented several more comparative stimuli, etc.

In order to implement this procedure, the comparative segments (stimuli) were organized into loops with all loops originating and terminating at the standard segment. Figure 1 shows the three loops used in this study. Since it is important that the standard segment remain distinct in the subjects' minds, one of the rougher segments (PSI = 1.6) was chosen. The standard segment was given an arbitrary value of 10.

EXPERIMENTAL DESIGN

The unusual requirements of this experiment necessitated the adoption of different procedures than have been used in previous studies. These procedures are briefly reviewed here.

Road Segment Selection

The study was conducted in rural Sheboygan County near Plymouth, Wisconsin. As shown in Figure 1, the course for the study was divided into three loops (A, B, and C). The loops A, B, and C included, respectively, 9, 12, and 11 comparative segments. In addition, a "standard segment" was located near the center of the course; thus, there were a total of 33 segments. All segments were 0.5 mi long. Of the 32 comparative segments, 11 were portland cement and 21 were bituminous concrete. Only six comparative segments had four lanes; the remaining twenty-six had two lanes. The standard segment was portland cement and had two lanes.

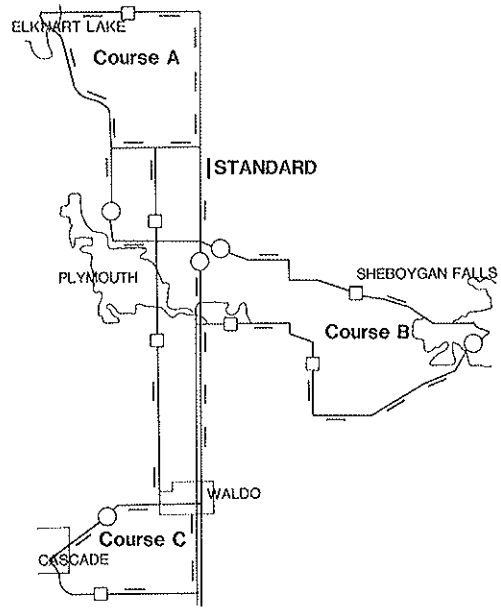

FIGURE 1 Geographical location of road segments.

The comparative segments were selected to represent different levels of pavement roughness, with PSIs ranging from 1.1 to 3.9. Successive segments were close to each other so as to minimize driving time and associated physical and mental fatigue to the subjects. All segments had a posted speed limit of 55 miles per hour. None of the segments was on freeways.

Vehicle Selection and Operation

A single, mid-sized car was selected because it is most representative of a typical car driven on Wisconsin highways. The car speed was maintained at 50 mph on test segments using cruise control. Windows were always rolled up to minimize road noise and wind effect on the subjects. Air conditioning was used as needed. Subjects were passengers and were required to wear seat belts. No smoking was allowed in the car.

Subject Selection

It was deemed critical to this study that the panel be representative of Wisconsin road users. Potential subjects were recruited at a nearby driver's license office and were asked several background questions. This background information was later used to select fifty subjects, twenty-five males and twenty-five females. The subjects represented a variety of age groups, family income groups, occupations, and places of res-

idence (rural versus urban). Subjects were paid to participate in the study.

Data Collection Forms

Three different forms were filled out by the subjects. Information about each subject was collected on the Background Information Form. A Road Segment Evaluation Form was used to rate subjective ride quality and related issues for each comparative segment. A Course Evaluation Form was used to rate passenger comfort, personal well-being, weather conditions, and driver skill. Thus, for each subject the Background Information Form was completed once, the Road Segment Evaluation Form was completed 32 times, and the Course Evaluation Form was completed three times, once for each loop. All three forms and instructions were assembled in looseleaf notebooks.

The Background Information Form was used to obtain information on subjects' backgrounds (age, sex, formal education, household income, etc.), driving habits (type of car, years of driving, miles driven/week, etc.), overall impression of Wisconsin highways, and relative importance of a number of variables related to subjects' satisfaction with an automobile ride. A total of 25 different pieces of information were collected on subjects' backgrounds.

Two different psychophysical scales were used to rate pavement ride quality. One was the traditional five-point Weaver/AASHO scale, the other a newly designed magnitude estimation scale where the standard segment was arbitrarily assigned a ride quality rating of 10.

The acceptability of pavement ride quality was determined using two different scales. The first question asked subjects to agree or disagree with this statement: "State and/or county money should be allocated within the next year to resurface or reconstruct this road in order to improve its ride quality." The second question determined how much extra time subjects would be willing to spend to avoid this type of pavement over a 50-min trip.

Previous studies had used a two-point (yes, no) scale for determining the acceptability of pavements, whereas this study has adopted a five-point Likert scale ("strongly agree" to "strongly disagree"). A five-point scale allows subjects to avoid definitive statements when the ride quality is neither very smooth nor very rough. Furthermore, a Likert scale can better provide a statistical relation between acceptability rating and pavement roughness.

In addition to ride quality and acceptability, the subjects were asked to rate other aspects of the ride (amount of traffic, appearance of the road surface, scenery, safety of the road and importance of the road) on five-point semantic differential scales. It was believed that some of these factors, along with subjects' background, could have some effect on ride quality and acceptability ratings, in spite of efforts to minimize such effects.

The purpose of the Course Evaluation Form was to determine if the subjects felt well, if the seats were comfortable, if the seating room was sufficient, and if they were satisfied with the driver. These ratings were collected in order to determine if adverse conditions influenced subjects' ratings of ride quality.

Data Collection Procedures

Each subject was randomly assigned a passenger seating position in the car. A research assistant drove over the standard segment six times, so that subjects were fully aware of the ride quality offered by that segment. Then the research assistant started evaluation of one of the three loops (A, B, or C). The order of the loops was randomly preselected. After each segment, the research assistant made a safe stop so that subjects could fill out the Road Segment Evaluation Form. The research assistant drove the car to a nearby rest area for a short break; then the procedure was repeated for the remaining loops.

Course Evaluation

In general, seating room, comfort of seats, interaction with other passengers, skill of the driver, personal well-being and weather conditions received very favorable responses from the subjects. Mean ratings for the six variables and the three courses ranged from 1.5 to 1.9, where "1" was most desirable and "5" was most undesirable. Thus, it would seem that none of these six variables had an adverse effect on ride quality and acceptability ratings.

Physical Road Roughness Measurement

Physical roughness of each segment was measured three times with the Wisconsin DOT response-type instrument—a roadmeter. The roadmeter yielded inches per mile, which was transformed into PSI. In addition each segment was measured by Michigan DOT with its profilometer. The profile was converted into a 0 to 100 scale of roughness called the Ride Quality Index (RQI). Michigan actually uses two versions of RQI; this study used the newer, more sophisticated version, so it is referred to here as "new RQI." No attempt was made to find an optimal transfer function between road profile and subjective evaluation of roughness.

MAJOR FINDINGS

Physical and Subjective Measures of Road Roughness

In the past, studies on road roughness have used averages (or mean values) for each segment to determine correlation coefficients and to perform regression analysis. However, statistics based on raw data are more desirable and powerful. All the statistics reported in this study are based on raw data unless otherwise stated. Correlations based on average values from the subjects will, in general, be substantially higher than those based on raw data. For example, Table 1 compares correlation coefficients based on raw data with those based on mean values.

Throughout this discussion, "ride quality rating" refers to the results of the magnitude estimation experiment.

Besides those correlations found in Table 1, extensive correlation analysis was performed on all variables from the Road Segment Evaluation Form. Most interestingly, road surface appearance was found to be highly correlated with all mea-

TABLE 1 COMPARISON OF CORRELATIONS BASED ON RAW DATA WITH THOSE BASED ON MEANS FROM THE FIFTY SUBJECTS

Variable No.	Variable	Correlation on Raw Data, Variable No. 1	2	3	4	5	Correlation on Means, Variable No. 1	2	3	4	5
1	Ride Quality Rating	1	-0.71	-0.58	0.48	0.51	1	-0.96	-0.85	0.70	-0.93
2	Weaver/AASHO Rating		1	0.53	-0.46	0.68		1	0.80	-0.68	0.93
3	PSI			1	0.85*	0.49			1	0.85*	0.92
4	New RQI				1	-0.42				1	-0.77
5	Money Allocated					1					1

*Independent of subject variability

sures of road roughness (subjective, physical, and acceptability). The measure of road roughness with the highest correlation with road surface appearance was the Weaver/AASHO scale (0.72). A lower correlation (0.59) was found between the ride quality rating and road surface appearance. It is understandable that the Weaver/AASHO scale would be highly correlated with the road surface appearance because of the general nature of the cues; the scale tends to capture other aspects of the road in addition to physical roughness. However, the ride quality rating does not have general cues, so it is less sensitive to other aspects of the road, such as road surface appearance, safety features, etc.

A typical objective of a ride quality study is to determine a quantitative relationship between subjective perceptions and physical measures of pavement roughness in order to make routine psychological estimates from measurement of the physical correlate. Least-squares regression analysis resulted in the following equations:

$$\text{Ride Quality Rating} = 13.76 - 2.33\ \text{PSI} \qquad (r = 0.58\ SE = 2.56) \quad (1)$$

$$\text{Ride Quality Rating} = -4.31 + 0.187\ \text{New RQI} \qquad (r = 0.48\ SE = 2.78) \quad (2)$$

$$\text{Weaver/AASHO Rating} = 1.48 + 0.637\ \text{PSI} \qquad (r = 0.503\ SE = 0.78) \quad (3)$$

$$\text{Weaver/AASHO Rating} = 6.44 - 0.051\ \text{New RQI} \qquad (r = 0.46\ SE = 0.82) \quad (4)$$

These relationships are illustrated in Figures 2–5. Note that

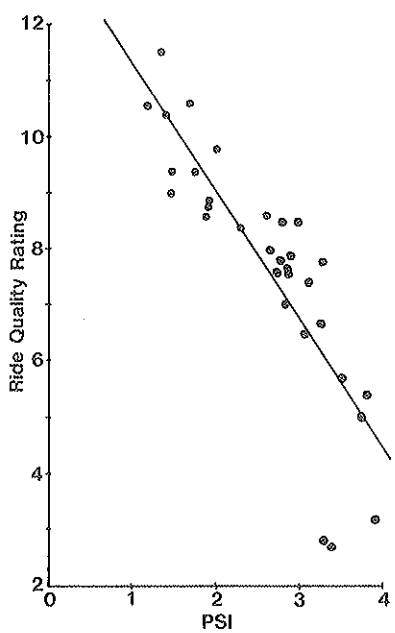

FIGURE 2 Relationship between mean subjective ride quality rating and PSI. (The regression shown is on the mean ride quality rating from the 50 subjects; $r = 0.85$.)

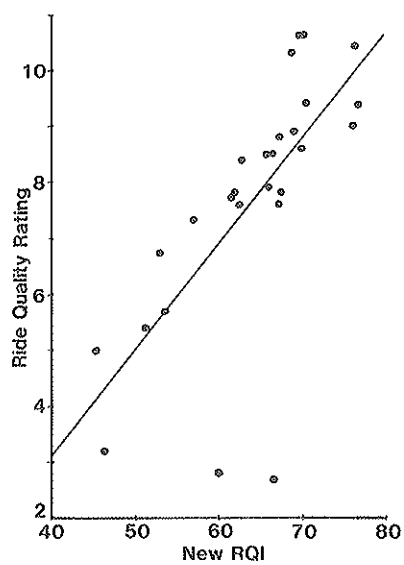

FIGURE 3 Relationship between mean subjective ride quality rating and new RQI. (The regression shown is on the mean ride quality rating from the 50 subjects; $r = 0.70$.)

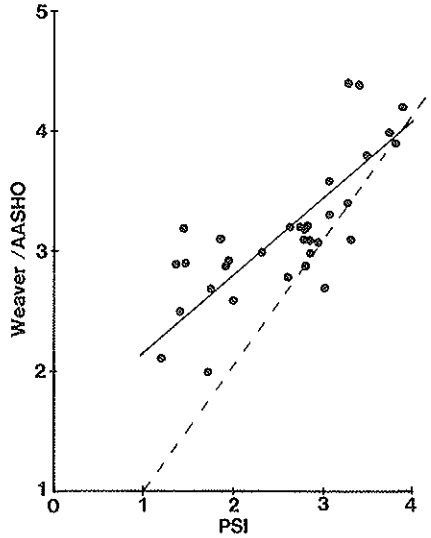

FIGURE 4 Relationship between mean Weaver/AASHO scale and PSI. (The regression shown is on the mean Weaver/AASHO rating from the 50 subjects; r = 0.80.)

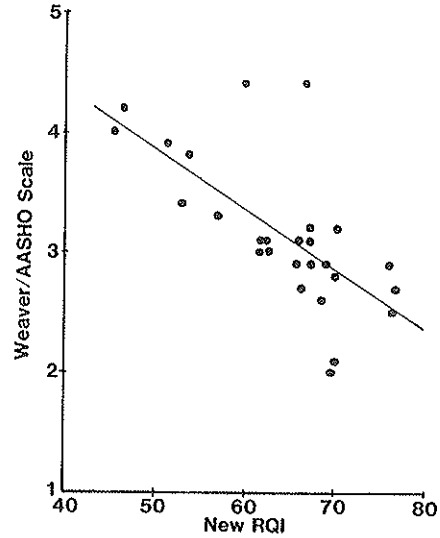

FIGURE 5 Relationship between mean ride quality rating on Weaver/AASHO scale and PSI. (The regression is shown on the mean Weaver/AASHO rating from the 50 subjects; r = 0.68.)

TABLE 2 RESULTS FOR STEPWISE FORWARD REGRESSIONS OF SUBJECTIVE MEASURE OF RIDE QUALITY AGAINST PSI AND OTHER ASPECTS OF THE RIDE

Dependent Variable	Independent Variables						
	Constant	Road Surface Appearance	PSI	Safety	Amount of Traffic	Correlation Coefficient (r)	Standard Error
Ride Quality Rating	2.44	1.92	-	-	-	0.59	2.54
	8.08	1.29	-1.52	-	-	0.67	2.33
	7.81	0.96	-1.69	0.62	-	0.69	2.27
	8.55	0.99	-1.72	0.59	-0.35	0.70	2.25
Weaver/AASHO Scale Rating	5.05	-0.70	-	-	-	0.72	0.64
	4.06	-0.59	0.27	-	-	0.75	0.61
	4.14	-0.49	0.32	-0.19	-	0.77	0.59
	4.04	-0.49	0.32	-0.18	0.05	0.77	0.59

only the mean values from the fifty subjects for the ride quality and Weaver/AASHO ratings are plotted in Figures 2–5; the regression lines in these figures were found from mean values and differ slightly from Equations 1–4.

Based on the above linear regressions, it appears that road roughness is best estimated by using the magnitude estimation type of scale for the psychophysical measure (ride quality rating) and PSI for the physical measure. The above relationship resulted in a correlation coefficient of 0.58 on raw data and 0.85 on mean values from the fifty subjects. The linear relationships between the subjective measures (ride quality and Weaver/AASHO ratings) and new RQI was found to be good, but not as strong as between the subjective measures and PSI.

Several other regressions were attempted to improve the relationship between ride quality rating and PSI. These included log transformation, polynomial equations, and others. However, none of those equations resulted in a significant improvement.

It is particularly interesting to note that PSI does not correspond closely to the Weaver/AASHO scale, as it should. Given the origins and interpretation of PSI, the y-intercept of equation 3 should have been about 0.0, and the slope should have been about 1.0 (dotted line on Figure 4). Either Wisconsin's method of measuring PSI no longer replicates PSI from the AASHO Road Test or the type of panel used in this study (a random sample of road users) differs substantially from the type of panel selected for the AASHO Road Test.

On balance, subjects were far less willing than the Wisconsin roadmeter to rate a road as "poor."

Since appearance of the road surface, scenery, amount of traffic, safety of the road, and importance of the road had significant correlations with the subjective road roughness ratings, forward stepwise linear regressions were performed between the subjective measures, these variables, and PSI. A level of significance of 0.01 was used for inclusion of an independent variable. The resulting regression equations are given in Table 2. It is clear from Table 2 that appearance of the road surface has a strong effect on the ride quality rating and an even stronger effect on the Weaver/AASHO scale rating. Road surface appearance and PSI explain 45 percent and 56 percent of the variation in ride quality rating and Weaver/AASHO rating, respectively. Thus, while PSI is important in determining subjective measures of ride quality, appearance of the road surface plays an equally important role.

The effect of road surface appearance on ride quality ratings is readily seen in Figure 2. There are three comparative segments that received particularly favorable ride quality ratings (about 3.0) but had PSI values of less than 4. These segments deviate substantially from the regression line. All three segments had new surfaces (two portland cement and one bituminous concrete) at the time of the experiment. There were no cracks, patches, discoloration, or any other evidence of deterioration. Objectively, however, there was some roughness to the surface.

The ride quality ratings show that subjects were significantly influenced by the excellent appearance of the road. The effect of appearance is even more pronounced when the ride quality rating is compared with new RQI (Figure 3).

Acceptability Measures

A series of linear regressions was performed to relate physical measures of road roughness to acceptability measures (for example, "money should be allocated"). Those regressions included a number of subject background variables such as seating comfort, use of seat belts, helpful road signs, percentage of miles driven on highways, privacy in vehicle, years of formal education, sex, weather, lack of construction, etc. However, the contribution of a single background variable was very small, and the number of variables was very large. By and large, little was learned from these regressions beyond the information contained in Table 1. That is, both PSI and new RQI are strongly related to the acceptability measures, with PSI providing a somewhat better fit. The fit could not be substantially improved by inclusion of information about the subjects' backgrounds. This result is important because it means that the results of this study (and similar studies) are likely to be insensitive to the location from which the sample is drawn.

Terminal Roughness

The major objective of this study was the determination of a terminal roughness, in other words, a value of PSI, that causes a predictable percentage of drivers to become dissatisfied with the road surface. In the past, this dissatisfaction has been measured by asking the subjects to answer yes or no to a question about spending tax dollars to improve the ride quality of a given road. Some studies have used a third category of undecided. In this study, however, a Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree) was used.

It is first necessary to establish a single criterion for dissatisfaction; in other words, which value on the Likert scale should be used to represent the "money-should-be-allocated" level of dissatisfaction with road roughness. For example, the value can range from 1 (strong agreement with the statement that "money should be allocated") to 3 (neutral point). Ideally, the cutoff value for "money should be allocated" should depend upon several factors, such as percentage of drivers dissatisfied, class of road (freeway versus local road), volume of traffic served, importance of the road to the travelling public, resources available, and so on. For the following discussion the criterion for dissatisfaction is set at 2.5. A value of 3.0 must be interpreted as ambivalent, and a value of 2.0 is too strict, considering the likelihood of central tendency in the scale. A value of 2.5 can be interpreted as being in slight to moderate agreement with the idea of spending money to resurface the road.

The percentage of subjects with a rating of 2.5 or less for "money should be allocated" is plotted against PSI in Figure 6. The linear correlation coefficient between the percentage of subjects and PSI was 0.89. In addition to a linear regression, curvilinear regressions were also tried. The linear regression resulted in a slightly better fit. The following equations can be used to estimate percent dissatisfied (PD) from PSI for a value of 2.5 for "money should be allocated."

$$PD = 72.49 - 19.74\ \text{PSI} \qquad (r = 0.89,\ SE = 8.2) \qquad (5)$$

$$PD = -19.81 + 95.48/\text{PSI} \qquad (r = 0.87,\ SE = 8.7) \qquad (6)$$

Equation 6 is recommended for determining the relationship between PSI and the percent of dissatisfied subjects. Equation 6 is recommended over equation 5 only because

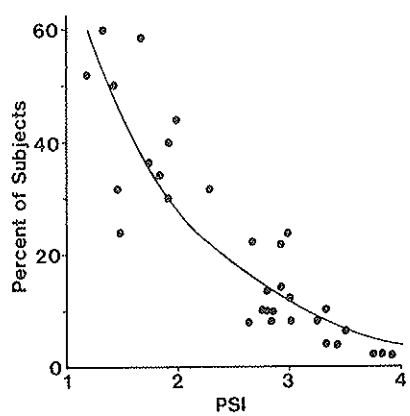

FIGURE 6 Relationship between percentage of subjects dissatisfied with ride quality based on a rating of 2.5 or less for "money should be allocated" and PSI.

TABLE 3 PSIs AS A FUNCTION OF PERCENTAGE OF RATERS DISSATISFIED AND LEVEL OF DISSATISFACTION

Percent of Raters Dissatisfied	PSI for "money should be allocated" rating of ≤ 2.5
50	1.36
45	1.47
40	1.60
35	1.74
30	1.91
25	2.13
20	2.39

it appears to fit better at lower values of PSI (between 1 and 2).

If desired, terminal roughness can easily be determined for any percentage of subjects dissatisfied from the graph, from Equations 5 and 6, or from Table 3. For example, the terminal roughness is 1.36 (PSI) for 50 percent of the Wisconsin drivers to be dissatisfied (as estimated by a value of 2.5 or less on the scale used for "money should be allocated"). This terminal roughness is considerably lower than those currently used by the Wisconsin DOT (2.0 to 2.5).

Willingness To Spend Extra Time

Subjects were asked to estimate the amount of time they would be willing to spend to avoid each segment, assuming they were to make a 50-min trip. The results are summarized in Figure 7. In addition, extra time may be computed from the following:

$$\text{Extra time} = 8.68 - 2.13\ \text{PSI} \qquad (r = 0.32,\ SE = 4.85) \qquad (7)$$

It is seen that the amount of extra time is small, but significant. For example, Equation 7 shows that subjects were willing to spend 5.5 minutes (or 11 percent) more time to avoid a road with a PSI of 1.5.

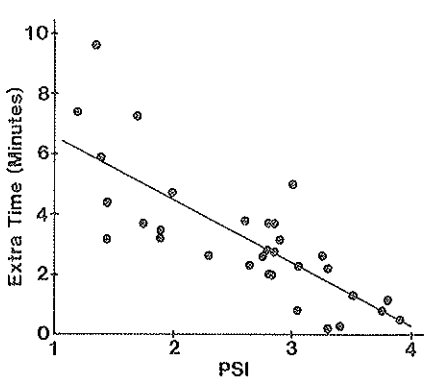

FIGURE 7 Relationship between extra time, in minutes, for a 50 min road trip and PSI.

These extra-time evaluations are analogous to time savings benefits. They can be converted to monetary units by using accepted values of time. Knowing the traffic volumes and speeds on a road, and the rate of pavement deterioration, it would be possible to compute total benefits of a new surface and to compare them with project costs.

CONCLUSIONS

Analysis of subjects' evaluations of ride quality leads to the following conclusions.

A number of subjects' personal and background variables had statistically significant effects on both the subjective measures of road roughness and the acceptability measures. However, these variables were found not to be of practical significance as they explained very little variation in either road roughness or acceptability measures. This suggests that the results of this study are generalizable; that is, they are not dependent on the location from which the sample is drawn.

Even though magnitude estimation is most easily accomplished in a laboratory, it is possible to successfully use this technique for measuring ride quality.

PSI has a higher correlation with the magnitude estimation scale (ride quality rating) than with the Weaver/AASHO scale. The magnitude estimation scale appears to have a greater power of discrimination between various roads, and it is less influenced by road surface appearance. Magnitude estimation is the preferred method of measuring subjective ride quality.

The best physical measure of road roughness appears to be PSI. PSI has the highest correlations with both the ride quality rating and the Weaver/AASHO scale. In a completely hands-off comparison, the less expensive roadmeter was found to be superior to the profilometer in predicting both subjective road roughness and the acceptability of the road. However, both instruments failed to relate with subjective measures of ride quality for those segments judged to be very good by the subjects.

Appearance of the road surface is extremely important to subjects rating ride quality.

People are willing to spend approximately 11 percent (5.5 minutes on a 50-min road trip) more time to take a better ride quality road and avoid a road with PSI of 1.5. This represents a significant resource expenditure. It is recommended that such time savings benefits be incorporated into evaluations of road resurfacing plans.

PSI, as measured in Wisconsin, does not closely approximate the Weaver/AASHO scale. There can be two explanations:

1. The measurement of PSI has somehow changed since it was first established, or
2. The subjects in this experiment differ considerably from the subjects in the AASHO Road Test.

If the latter explanation is true, then it is likely that PSI, regardless of where it is currently being measured, is not properly reflective of the opinions of road users.

A panel can provide detailed information about the need for road resurfacing. In order to properly use this information, it is necessary to establish both a criterion level of dissatisfaction and a percentage of road users who would be so dissatisfied.

ACKNOWLEDGMENTS

A number of people contributed both time and effort to bring this study to completion. The authors are grateful to Karl Dunn and his staff at WisDOT for their encouragement, suggestions, cooperation, and assistance. A special note of thanks is extended to the subjects without whose enthusiasm and cooperation this project would never have been completed.

Professor Edward Beimborn provided advice and encouragement. Professor Umesh Saxena made valuable suggestions in the statistical design and analysis of the experimental data.

We are also appreciative of the efforts of the staffs of the Human Performance Laboratory and the Center for Urban Transportation Studies of the University of Wisconsin at Milwaukee, among them: David Sincere and Richard Clark who collected, coded, and stored the data in the computer; Randal Zakowski who provided graphical illustrations; and Tarun Gupta who performed all statistical analyses.

REFERENCES

1. L. F. Holbrook. Prediction of Subjective Response to Road Roughness by Use of the Rapid Travel Profilometer. In *Highway Research Record 69*, HRB, National Research Council, Washington, D.C., 1969, pp. 212–226.
2. S. Mohan. *Development of a System for the Evaluation of Pavements in Indiana*. Joint Highway Research Project No. C-36-63-G, Purdue University, Ind., 1978.
3. S. K. Nair and W. R. Hudson. Serviceability Prediction from User-Based Evaluations of Pavement Ride Quality. Presented at the 65th Annual Meeting of the Transportation Research Board, Washington, D.C., 1986.
4. W. R. Hudson. Road Roughness: Its Elements and Measurements. In *Transportation Research Record 836*, TRB, National Research Council, Washington, D.C., 1981, pp. 1–7.
5. B. G. Hutchinson. Principles of Subjective Rating Scale Construction. In *Highway Research Record 46*, HRB, National Research Council, Washington, D.C., 1964, pp. 60–70.
6. M. S. Janoff and J. B. Nick. *Evaluation and Quantifying User Perception of Pavement Serviceability*. Report for Pennsylvania Department of Transportation, Pennsylvania, 1983.
7. W. N. Carey and P. E. Irick. The Pavement Serviceability Performance Concept. *Bulletin 250*, HRB, National Research Council, Washington, D.C., 1960, pp 40–50.
8. S. S. Stevens. On the Operation Known as Judgement. *American Scientist*. Vol. 54, 1966, pp. 385–401.

DISCUSSION

MICHAEL S. JANOFF
JMJ Research, P.O. Box 144, Newtown, Pa. 18940.

This is a well written paper that presents potentially valuable information. However, because of some omissions, misinterpretations, and possible mistakes, the results are not readily applicable. The authors have not considered the results of recent work by NCHRP and by Ohio DOT and have incorrectly interpreted a number of their references.

1. The authors state that little information is available to define the relationship between ride quality and roughness. This is false. Studies conducted in Texas, Ohio, Louisiana, New Mexico, Michigan, New Jersey, and Pennsylvania define such relationships quite well (*1–5*). The data from Ohio, New Jersey, Michigan, New Mexico, and Louisiana also reveal that the relationship between roughness measured with a profilometer and subjective ratings of ride quality are nearly identical for each of the five different states.

2. Perhaps the greatest problem in this paper is the relatively poor correlations that the authors have found between physical measures of pavement roughness and subjective ride quality ratings. For profile-derived roughness measures the authors have found correlations of at most .7 with (mean) ride quality ratings, which are far lower than the correlations found between profile measures of roughness and ride quality ratings in Ohio, New Jersey, Michigan, New Mexico, Louisiana, and Texas. For four different models of profile-derived roughness used in these other states, correlations between the roughness and ride quality ratings were always in excess of .9, and for the data from five states combined they were better than .93. A number of explanations for the authors' low correlations can be suggested.

First, the instrument used to collect profiles was the older one-wheelpath unit of Michigan DOT, which is not as accurate in measuring profiles as the more common two-wheelpath versions used in the states mentioned above. Although the profile from one wheelpath can be accurately used to predict rideability (*3*) the instrument used to collect this one-wheelpath profile must provide accurate data; it is suspected that the instrument used by Wisconsin DOT was not. In tests conducted as part of the NCHRP research (*3, 6*) the older one-wheelpath profilometer yielded correlations between RQI and ride quality of .85, while for the two-wheelpath profilometer these correlations between RQI and ride quality rose to .93.

Second, the magnitude estimation scale, which the authors believe is preferred to the similar interval scale, may not be yielding accurate and consistent rating data. As the authors point out, magnitude estimation is best suited to laboratory

research where the "standard" level can be presented to the test subjects along with each test condition. Was this scale pretested in any way to determine if it was suitable for such ratings of ride quality? (It appears that the raw rating data obtained with the magnitude estimation scale, as illustrated in their final report, appeared to have rather high rater variability, indicating a lack of consistency in the rating procedure.)

Both Holbrook (*7*) and Janoff and Nick (*5*) have shown that the choice of rating scale has no effect on the quality of the subjective ratings. The analysis by Hutchinson that the authors use to support their selection of the magnitude estimation scale as the preferred one was shown to be unimportant in panel ratings of ride quality (*5*). Although I believe that Hutchinson is correct in theory, in application drivers can use almost any scale to accurately and consistently rate ride quality, as demonstrated by Holbrook and by Janoff and Nick.

The instructions to the raters could also have had a significant effect on the quality of the ride quality ratings. In the past research for Pennsylvania DOT (*5*) three different rating scales were tested, and it was disclosed that with proper instructions to the raters either one could be used to derive accurate and consistent ratings of ride quality. This same result was also found by Holbrook. Did the authors test their instructions in any way?

The authors' finding that surface appearance is affecting the ratings of ride quality may also be a result of the instructions; Holbrook showed that the use of blindfolds had no effect on such subjective ratings, hence with proper instructions appearance should have no effect on ride quality (*7*).

3. PSI is not a physical measure of roughness but a value derived from a regression equation that relates a physical measure of roughness (typically an RTRRMS index) with a subjective rating of ride quality (such as mean panel ratings or PSR). That the authors found a correlation of .35 between their magnitude estimation scale and PSI and a correlation coefficient of .9 between PSI and the Weaver/AASHO rating is not unexpected. When RTRRMS data from New Jersey, Ohio, and Louisiana were compared to ride quality ratings a correlation of .79 was found for the Weaver/AASHO scale, almost identical to the authors' results. Unless the surfaces are separately analyzed for each type (BC, PCC, composite) no better results can be expected.

The problem is that RTRRMS instruments do not provide roughness data that are highly correlated with ride quality ratings except for BC surfaces (actually such instruments fail to respond to all of the roughness frequencies present in PCC or composite surfaces). On surfaces other than BC, and when the range of ride quality is great, the correlations fall to very low values. The authors should have analyzed the surface types individually to disclose the effect of surface type on the correlations.

The authors' conclusion that PSI, therefore, no longer approximates the Weaver/AASHO scale is conjecture only. A better explanation is that when RTRRMS are used, correlations of .8 to .85 are typical; to increase the correlation it is necessary to use profiles to measure roughness.

4. The Weaver/AASHO scale used by the authors (and illustrated in their final report) is not the same scale that was used by either Carey and Irick (*8*), Weaver (*9*), Janoff and Nick (*5*), or NCHRP (*2*, *3*); it is a completely different type of scale. The Weaver/AASHO scale is a vertical line, typically 5-in long, with major subdivisions at 1-in intervals (labelled 0, 1, 2, 3, 4, and 5) and minor subdivisions at ½-in points labelled "very poor" (at ½-in) "poor" (at 1½-in), "fair" (at 2½-in), "good" (at 3½-in), "very good" (at 4½-in), "perfect" (at the top) and "impassable" (at the bottom). The authors' scale includes the word and number cues but not the vertical line. In addition, the instructions to the raters used by the authors do not explain how to even mark this scale. In past uses of the Weaver/AASHO scale the raters were told to place a horizontal mark across the scale (in other words, across the vertical line) at the point that they feel best describes the ride quality of the test section. It appears that on the authors' scale the subject is instead placing a mark into one of the categories defined by the ½-in divisions. This is completely different from past applications and hence the data (that the authors refer to as Weaver/AASHO ratings) may be incomparable to past applications of this scale.

5. The authors use a five-point Likert scale instead of a two-point yes versus no rating to indicate need for improvement, claiming that such a scale is better for the intended use. Again no pretest or pilot results are shown, and it is suspected that the results may suffer the same problems as those related to the magnitude estimation scale.

6. In summary, the authors conclude that magnitude estimation is the preferred rating procedure and that PSI (or actually RTRRMS indexes) are preferred to profile-derived roughness for predicting ride quality. The opinion of this writer is that these conclusions are unsupported. The correlations between the roughness and rating data are too low in comparison to other recent ride quality research to draw such conclusions and the authors' lack of attention to and misinterpretation of past research sheds serious doubts on the validity of their results.

REFERENCES

1. K. N. Sukumar and W. R. Hudson. Serviceability Prediction from User-Based Evaluations of Pavement Ride Quality. In *Transportation Research Record 1084*, TRB, National Research Council, Washington, D.C., 1986.
2. M. S. Janoff et al. *Pavement Roughness and Rideability*. NCHRP Report 275, 1985.
3. M. S. Janoff. *Pavement Roughness and Rideability Field Experiment*. NCHRP Project 1-23(2), Final Report, Dec. 1987.
4. J. McQuirk and E. Spangler. *Use of the Inertial Profilometer in the Ohio DOT Pavement Management System*. ASTM Committee E-17 Symposium, Columbus, Ohio, June 1985.
5. M. S. Janoff and J. B. Nick. *Evaluating and Quantifying User Perception of Pavement Serviceability*. Penn-DOT, 1983.
6. M. S. Janoff. *Quarterly Report on NCHRP Project 1-23(2)*. Dec. 1986.
7. L. F. Holbrook. Prediction of Subjective Response to Road Roughness by Use of the Rapid Travel Profilometer. In *Highway Research Record 69*, HRB, National Research Council, Washington, D.C., 1963.
8. W. N. Carey and P. E. Irick. The Pavement Serviceability Performance Concept. In *Highway Research Record 250*, HRB, National Research Council, Washington, D.C., 1968.
9. R. J. Weaver. *Quantifying Pavement Serviceability as It Is Judged by Highway Users*. NYSDOT, 1978.

AUTHORS' CLOSURE

The basis of Mr. Janoff's criticism is that our results are in disagreement with his own work. Such criticism is a two-edged sword; it can raise doubts about the validity of Mr. Janoff's earlier efforts in measuring ride quality. Indeed, the principal reason for our conducting a ride quality experiment was disappointment with the experimental design of the NCHRP study to which Mr. Janoff refers. We chose not to make an issue of Mr. Janoff's work in our paper. However, we were particularly disturbed by the subject selection procedures, instructions to subjects, the way stimuli were presented, and the rating scales in the NCHRP study. That the results of the two studies are different is unremarkable; we attempted to correct problems in the NCHRP study design that we believe could have skewed the results.

We believe the low correlations between subjective measures of ride quality and physical measures of ride quality stem from the "hands-off" nature of our study design. We chose to take physical measures from other sources so that their transferability could be evaluated. The physical measure from the response-type instrument (PSI in our case) behaved as expected. The performance of the profilometer (new RQI) was much poorer than expected. Based on many earlier studies (including Mr. Janoff's) we do not feel that the problem stemmed from the quality of either equipment or data analysis. One-wheelpath profilometers have served quite nicely in the past. Instead, it appears that complex transformations of profilometer data are not applicable when there are significant variations in study design (for example, location, subject selection, or stimuli presentation). A major lesson of these comparisons is that engineers cannot take any physical measure of ride quality at face value; they must fully understand the conditions under which panel data were assembled.

Magnitude estimation is one of the most respected psychophysical measurement techniques. Its use in ride quality measurement has been suggested by many authors, and an abbreviated version was tried by Holbrook with very encouraging results. Mr. Janoff has seriously misrepresented Holbrook's conclusions. Holbrook states, "The equations predicting ride from physical input by means of magnitude estimation are to be preferred" (Janoff's reference 7, page 255). A magnitude estimation scale is quite different from any of the three categorical scales tested by Janoff and Nick. Their conclusions are not relevant.

Our experience indicates Hutchinson's criticism of the Weaver/AASHO scale is correct in practice as well as in theory. It is unfortunate that Mr. Janoff decided to disregard Hutchinson's advice.

We found intersubject variability to be about the same for magnitude estimation and for the Weaver/AASHO scales. We did not address this issue in our paper because it is uninteresting. Mr. Janoff's erred during his casual inspection of summary data from our full report.

PSI is a physical measure of ride quality; its peculiar historical origins cannot alter the fact that only mechanical data are used to compute it.

Our conclusion that Wisconsin's PSI no longer approximates the Weaver/AASHO scale is based on Figure 4 in our paper. Regardless of surface type, the regression line is plainly at the wrong angle.

Mr. Janoff has apparently received incorrect information about our depiction of the Weaver/AASHO scale. It was purposefully identical to the one used in his NCHRP study. In regard to another scaling issue, we are perplexed why Mr. Janoff believes an ad hoc 2-point scale of acceptability is inherently superior to a Likert scale—a standard tool of psychometrics.

We appreciate Mr. Janoff's interest in our research. However, we do not feel that his discussion is helpful for either understanding or extending our work.

Publication of this paper sponsored by Committee on Surface Properties—Vehicle Interaction.