Evaluation of Panel Rating Methods for Assessing Pavement Ride Quality

J.B. NICK AND M.S. JANOFF

The results of a pilot study that attempted to determine the preferred psychophysical scaling method for obtaining panel ratings of pavement ride quality are reported. Five principal tasks were undertaken: site selection, scale selection, panel selection, design and conduct of the experiment, and analysis and interpretation of the results. Thirty-three flexible road sections covering a wide range of roughness (uniform within each site) were selected, and a 92-mile route that could be traversed, at normal operating speeds, in 3.5 hr was established. Three candidate scales were selected for evaluation: the original AASHO rating scale, a scale developed by Holbrook that uses precisely defined and positioned word cues, and a nonsegmented scale with word cues only at the end points. Fifty-four average drivers were divided into 3 panels and rated the 33 sites in groups of 3. Each panel member used only one of the 3 scales and rated each section only once. A Mays ride meter was used to obtain roughness measurements for each section. The analysis was designed to (a) determine which scaling method resulted in the greatest agreement between raters and (b) determine which of the three scaling methods resulted in the best correlation between subjective and objective measures of road roughness. The results indicated that either subjective scale, if carefully used (i.e., with exact instructions and precise control of conditions), can provide high agreement between raters and exceptionally high R^2 values.

In this paper, the results of a pilot study conducted to evaluate three psychophysical scaling methods for obtaining panel ratings of pavement ride quality are summarized. The pilot study was part of a larger project conducted by Ketron, Inc., for the Pennsylvania Department of Transportation (PennDOT) for the purpose of investigating the predictive relationship between subjective ratings of pavement ride quality and Mays ride meter (MRM) measures of pavement roughness. The overall goal of the entire project was to develop regression equations between the subjective and objective measures--one for each of the three types of surfaces (flexible, rigid, and flexible over rigid)--so that MRM measurements could be used as surrogates of subjective responses to road roughness.

As a first step in meeting the objectives of the project, a pilot study was designed to evaluate several candidate rating scales for assessing pavement ride quality. The pilot study consisted of five interrelated tasks:

1. Selection of candidate scaling methods,
2. Selection of sites to be rated,
3. Selection of the rating panels,
4. Design and implementation of the experiment, and
5. Analyses and interpretation of the results.

The results of these tasks are summarized in this paper.

SELECTION OF RATING METHODS

Direct Versus Indirect Scaling Methods

Because the overall goal of this study was to establish the quantitative relationship between physical measurements of pavement roughness and subjective perceptions of ride quality, it was necessary to measure the psychological experience on at least an interval scale. That is, to establish a functional relationship between two measures, certain mathematical operations would be required that can only be meaningfully conducted when both quantities are measured on at least an interval (or ratio) scale. A problem arises, however, in that externally observable judgments by human subjects must be relied on to obtain the ratings of ride quality. Thus, certain assumptions must be made about the ability of the subjects to judge pavement ridability at the desired level of measurement.

There are two basic methods for obtaining ratings on an interval scale: direct and indirect. The fundamental difference between these methods is the assumption about the ability of the subject to describe preferences or sensations at the intended level of measurement. Direct scaling refers to methods of obtaining judgments of psychological quantities directly on an interval (or ratio) scale. In these methods, the desired quantitative level of judgment can be obtained either by carefully designing the scaling instrument or by the instructions the experimenter gives the individual raters at the beginning of the experiment. If it is assumed that the subjects have been able to carry out the task as intended, then the scale values of the stimuli (roads in this case) on an interval scale are given directly by their ratings and the scaling problem becomes one of merely averaging the results.

For indirect scaling methods, a subject's response, whether it be a statement of preference or an intensity of sensation, is only considered to be an indirect index of a mechanism mediating between the hypothetical psychological responses and the magnitudes of the external stimuli. Thus, in these methods an investigator assumes that, if a number of subjects were asked to judge each of several stimuli as belonging in one of a limited number of categories, only a rank ordering (i.e., an ordinal level of measurement) of preferences of road sections would be obtained, and additional statistical methods must be used to obtain the final ratings on an interval level of measurement.

Three scaling methods were selected for evaluation in the pilot study. One method was intended to yield indirect scale values and the other two were intended to yield scale values directly (although each was eventually analyzed as if it yielded both direct and indirect values).

All of these methods involve the use of a graphic rating-scale technique to record subjects' responses. This technique is a special type of category scaling that, in one form or another, has been used in almost all serviceability studies. Depending on the underlying assumptions, graphic rating techniques such as the one used here allow one to treat the responses as either lying in one of a limited number of categories (indirect methods) or lying along a linear continuum of unlimited categories (direct methods).

An entirely different approach, and one recommended by Holbrook (1), would be to ask the subjects to match a number directly to the perceived magnitude of ride quality and eliminate all intervening categories, orderings, cues, comparisons, and theories. However, magnitude estimation, as this method is called, was eliminated from consideration because

there was some question as to how well the subjects would be able to make the successive ratio judgments required by this method over the 100 roads in two days that the main study would involve.

WEAVER-AASHO SCALE

The first scale considered for evaluation was what is referred to here as the Weaver-AASHO scale. This scale, shown in Figure 1a, was originally used for the AASHO Road Test by Carey and Irick (2). Because this scale and similar versions of it have been used in a number of studies, it was selected for evaluation primarily as a baseline instrument with which the other scales could be compared.

This scale was also intended to be the one used to obtain indirect scale values by using an analysis method adopted by Weaver. After the AASHO Road Test, reviews such as that by Hutchinson (3) noted that the method used by Carey and Irick (2) may have violated several principles of psychometric methods. The most notable violations were the following:

1. The systematic errors of leniency (the tendency of a subject to rate too high or too low, for whatever reasons), the halo effect (contamination of subjective response caused by stimulus attributes other than those under consideration), and central tendency (the tendency of subjects to hesitate in giving extreme ratings) were not compensated for or removed from the raw data.
2. The raters were not provided with descriptive cues, or anchors, that corresponded closely enough to the trait being measured (i.e., ride quality).
3. It was assumed that the raters were good instruments of quantitative observation and yielded ratings directly on an interval scale.

There is considerable debate in the literature about the last issue. Many researchers believe that, in this situation, subjects are only capable of rendering judgments on an ordinal level and not directly on an interval level; that is, because it is possible that the judgments are only rank ordered, it is possible to determine whether one road was judged better or worse than another but not by what magnitude. Because of this mistrust of the ability of the subjects to make their judgments at the desired level of measurement, Weaver (4) of the New York State Department of Transportation (NYSDOT) adopted an indirect scaling method called the method of successive categories to analyze the data obtained by this scale. This method is based on Thurstone's law of comparative judgment and was developed by Guilford (5).

Although the method of successive categories relies heavily on untestable, hypothetical models of human judgment and requires a complicated analysis procedure, it was selected as a candidate method for evaluation in the pilot study for two reasons:

1. No study had yet been conducted that compared the results obtained by the more sophisticated direct scaling methods with the results from an indirect method.
2. Because Weaver had been using this method in New York for several years, apparently with success, a potential comparative data base was provided.

However, even with the implementation of the method of successive categories to transform ordinal judgments into interval-level scale values, this scale still potentially violates the other two principles of psychometric methods mentioned earlier by not providing more descriptive cues and by not compensating for the presence of possible systematic errors in the data.

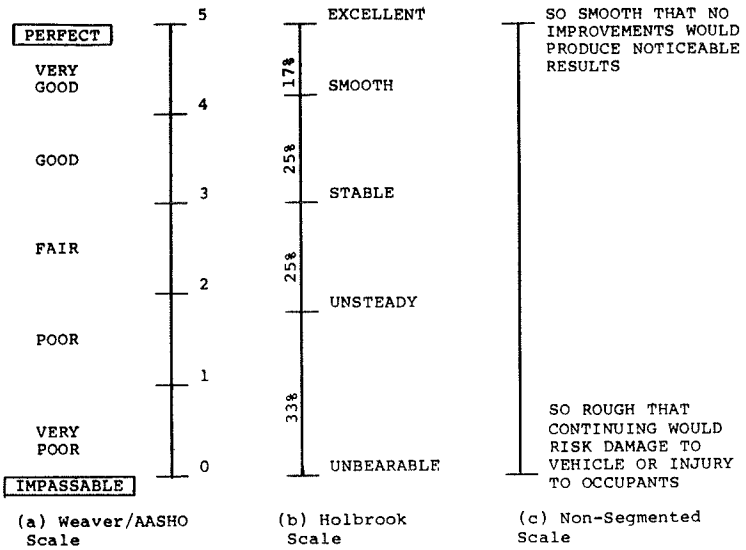
Holbrook's Scale

It is theoretically possible both to overcome the potential problems inherent in the Weaver-AASHO scale and to create a scale that yields interval-scaled ratings directly through more careful selection and placement of relevant cue words along a continuum. Holbrook (1) devised such a scale after considerable experimental work in selecting and placing the cue words (see Figure 1b).

The cue words were selected by having 80 subjects rate 92 words that could be used to describe varying degrees of ride quality along an 11-point roughness continuum. The 5 words shown on the scale were selected because of their narrow distribution in this preliminary scaling task. The median values for these 5 words along the 11-point roughness scale determined the location of the final scale.

This scale is presumably an improvement over the Weaver-AASHO scale because the cue words, or an-

Figure 1. Scales tested in pilot study.



(a) Weaver/AASHO Scale (b) Holbrook Scale (c) Non-Segmented Scale

chors, are not only relevant to the task of evaluating ride quality but are also placed along the scale so that the intervals between them represent similarly spaced intervals along the psychological roughness continuum. In addition, because of these characteristics, the ratings obtained with this scale should be relatively free of the systematic errors of leniency and central tendency. Both types of errors are generally counteracted by anticipating their magnitude and direction and then adjusting the position of the cue words along the scale to bias the raters so that they will not make these errors. However, because the strength and the position of the descriptive adjectives for this scale have been empirically predetermined, the compensating adjustments should have been made automatically. (The halo effect must still be dealt with by giving specific instructions to the raters.)

Nonsegmented Scale

Even though Holbrook's scale seems to represent a considerable improvement over the Weaver-AASHO scale, there could still be problems with the use of intermediate cue words. As Torgerson (6) notes, "In general, when we attempt to construct a scale for a set of stimuli using one of the subjective estimate methods, we are free to specify two and only two points on the continuum." If more points are to be defined, one must be sure that the placement of these points and their associated cues will represent comparably spaced intervals along the psychological scale for all (or almost all) of the raters. Although Holbrook was careful in selecting and placing his cues, the meaning of the intermediate cue words--"smooth," "stable," and "unsteady"--could still be vastly different for many subjects.

To avoid the connotative problems associated with the intermediate cue words, a better scale may be one in which only the ends are anchored and the subjects are allowed to place their rating marks on the scale unaided, as Torgerson has recommended. Hence, the last scale to be evaluated in the pilot study was what is referred to here as a nonsegmented scale (see Figure 1c). This scale was intended to yield direct interval values by virtue of the instructions given to the subjects. The subjects were instructed to place their marks on the scale so that the ratios of the differences in the distance between their marks reflected similar differences in the ride quality of the roads.

SITE SELECTION

The original plan for site selection required that 24 flexible-pavement sites be used in the pilot study. These 24 sites were to represent (a) four of the five maintenance functional classifications (MFCs) into which PennDOT stratifies the roads in the state (MFC-A was not represented because there are few flexible pavements in this class), (b) two different topographies (flat and rolling or hilly), and (c) three degrees of roughness spanning a broad range within each MFC.

However, in addition to the MFC, topography, and roughness requirements, there were other site characteristics that were considered essential to a successful study and eventually became important controlling factors in the evolution of the final course. In order of descending importance, these characteristics were the following:

1. The sites were to be as close as possible to each other to minimize the travel time between sites and through the entire course. Because it was anticipated that two runs a day would be made, a maximum total travel time of 3 hr was considered essential.

2. The length of the sites was to be adjusted so that travel time through each site would be equal and raters would thus have the same amount of time to evaluate the ride quality of each site. Thus, the length of each site depended on the expected speed to be used in driving through the site.

3. Each site was to have a uniformity of roughness over its entire length. In addition, approximately 500 ft of roadway before and after each site was to exhibit the same roughness characteristics. It was believed that this would help the raters avoid any uncertainty as to what the rating should be if an unrepresentative surface or bump were encountered.

4. The sites were to be selected only from rural areas in order to control the surrounding environment and minimize the possibility of being stopped within a site because of congestion or traffic lights.

5. Ideally, the sites were to be as straight as possible; if they were curved, the curves were to be gentle enough to allow speed to be maintained.

Initial field reconnaissance of potential sites on which PennDOT had previously used the Mays ride meter revealed that a significant number of the sites did not meet these criteria or their tabulated roughness had changed as a result of resurfacing. Consequently, the site selection strategy consisted of mapping a convoluted route linking the few usable sites and then selecting sites along this route that appeared, by visual inspection, to meet the requirements. In addition to the minimum of 24 sites required by the initial design, to prevent rater boredom other sites were selected along the route so that the time between sites was not excessive.

After the candidate sites and the route were approved, MRM measurements were made by PennDOT. All sites and almost all of the intersite roads were measured at a nominal speed of 40 mph. The PennDOT standard operating procedure is to measure all roads at 40 mph (or 25 mph if absolutely necessary) because the PennDOT MRMs have been calibrated at these speeds and correlation factors are available for comparison of the results obtained when either of these two speeds is used.

The resulting roughness distribution was found to be fairly rectangular and ranged from a low of 94 in./mile to a high of 752 in./mile. Furthermore, when linked, these sites formed a 92-mile route that took a manageable 2.75 hrs to drive at the posted speed limits. The median time between sites was 3.5 min.

The characteristics of the sites are given in Table 1. It should be noted that the ultimate distribution of exposure times ranged from 18 to 34 sec and the median time was 27.5 sec. Ideally, the length of a site was to be the empirically determined distance that yielded the desired exposure time, which, according to the original design, could have been anywhere between 23 and 29 sec. Frequently, however, the actual end point was picked to coincide with the most convenient vertical object nearest to the ideal end point that could be painted so as to be seen easily by the driver. Hence, the distribution was somewhat broader than desired. Nevertheless, no evidence was found to suggest that the differences in exposure time had a significant impact on the outcome of the study.

Table 1. Summary statistics for pilot study sites.

Site	MFC	Test Speed (mph)	Length		Travel Time (sec)	Axle Displacement	
			Miles	Feet		Total (in.)	Inches per Mile
1	B	45	0.308	1,626	25	131	424
2	C	45	0.426	2,249	34	76	17
3	D	30	0.291	1,536	35	219	752
4	D	35	0.285	1,505	29	161	56
5	E	30	0.245	1,204	29	102	418
6	E	35	0.200	1,056	21	60	301
7	D	40	0.400	1,440	25	95	238
8	D	55	0.390	2,059	25	42	108
9	E	35	0.300	1,584	31	108	358
10	D	45	0.432	2,281	33	52	120
11	E	40	0.372	1,964	34	100	268
12	B	45	0.364	1,922	29	40	109
13	B	45	0.272	1,436	22	33	121
14	D	40	0.252	1,331	23	28	109
15	C	45	0.299	1,579	23	39	130
16	C	55	0.400	2,112	26	38	94
17	D	35	0.324	1,711	33	209	645
18	E	35	0.335	1,769	34	97	290
19	E	40	0.249	1,315	22	51	203
20	E	35	0.209	1,104	22	45	216
21	E	30	0.208	1,098	25	61	295
22	E	35	0.285	1,505	29	111	390
23	D	40	0.274	1,447	25	63	231
24	D	30	0.257	1,357	31	158	615
25	D	45	0.228	1,204	18	30	133
26	D	40	0.331	1,748	30	85	257
27	E	30	0.204	1,077	24	149	733
28	D	45	0.294	1,552	24	50	170
29	B	40	0.292	1,542	26	72	248
30	D	40	0.275	1,452	25	133	485
31	D	40	0.272	1,436	24	116	428
32	C	45	0.389	2,054	31	56	145
33	B	45	0.369	1,948	30	76	207

SELECTION OF PANEL MEMBERS

Panel Size

In the pilot study three panels of 18 observers were used, one panel to evaluate each of the three scaling methods. This panel size was selected as a conservative compromise between the requirement of 20, to keep the maximum error in estimating the true population mean rating for any given site at or below 0.4 scale unit on the assumption that the sampling distribution was not normal, and the requirement of 9 observers for the same error if the sampling distribution was assumed to be normal. Ordinarily, smaller panel sizes would be acceptable because it would not be unreasonable to assume that the distribution of sample means approaches normality. However, there were at least two reasons for using more conservative estimates. First, because it could not be guaranteed a priori that the obtained ratings would be on an interval level of measurement, the estimated standard deviation of 0.6 (estimated from a subset of Weaver's data) used to arrive at a sample size may have been wrong or even meaningless. Second, in implementing the indirect scaling analysis method, the relative proportion of time a road is placed in a particular rating category by the subjects in the panel should be as similar as possible to the proportions that would be obtained from the population. This could only be ensured by using larger sample sizes. In addition, it was felt that the increased experimental control during the pilot study as well as its primary objective (i.e., to select a preferred rating method rather than to develop the final statistical relationship between objective and subjective data) reduced the requirement for a larger panel. A panel of 36 was ultimately used to derive the functional relationship between subjective ride quality and MRM data in the main study.

Panel Composition

Because balancing the composition of the panel by sex and age was the primary concern in selecting subjects, no explicit attempt was made to seek a balance based on other demographic variables. The panel was poststratified by miles driven per year, years of driving experience, vehicle size normally driven, seat position (front right, rear right, or rear left), scale used (one of three), and starting point (three starting points were used). The last three variables as well as the variable of sex were evenly divided among the panel.

DESIGN AND IMPLEMENTATION OF EXPERIMENT

Before they took the trip through the course, each group of three subjects was assigned to a particular combination of scale type and starting point (three alternative starting points were used to distribute the possible effects of learning and fatigue). The groups were then given their instructions.

Panel Instructions for Weaver-AASHO Scale

The panel instructions covered the purpose of the study, how the subjects were to use the rating scale, the procedures to be used in making their ratings, and other general topics. The instructions were outlined on a chalk board to ensure that all subjects received the same information each time. The instructions for the Weaver-AASHO scale are presented here. Instructions for the other scales are presented elsewhere (7).

Highway Improvement Study

Purpose: To survey typical Pennsylvania drivers in order to determine what they think of the quality of the ride provided by the roads in the Commonwealth. PennDOT will use this information to decide which roads to fix first given limited funds.

Object of Study: To obtain your personal opinion of how rough or smooth a ride is provided by roads in the area which represent the condition of various roads throughout the Commonwealth.

How to Make Your Ratings (A facsimile of the rating scale to be used was on the board for this section.)

Object: To place a mark across the vertical line which you think best describes the ride provided by the roads.

Definition of End Points

Impassable: A road which is so bad that you doubt that you or the car will make it to the end at the speed you are traveling--like driving down railroad tracks along the ties.

Perfect: So smooth that at the speed you are traveling you would hardly know the road was there. You doubt that if someone made the surface smoother the ride would be detectably nicer.

All the roads which you drive over today will be between these two extremes. That is, since these roads probably do not exist you will probably not consider any road to be worse than impassable or better than perfect. In order to help you place your mark on the line, we have included a number of words along the scale which could be used to describe how the riding sensation seems to you.

For example, if you should encounter a road for which you could describe the ride as FAIR but not quite GOOD, place your mark just below the line labeled "3" (illustrated). On the other hand, if you think the next road is still fair, but somewhat worse than the previous road, place your mark at a point which you think is the appropriate distance down in the FAIR category. To indicate small differences between the ride quality provided by the roads, you may place your mark anywhere you like along the scale.

NOTE: We are not asking you to place roads into one of five categories! You should use small differences in the position of your marks to indicate small differences between the ride quality provided by the roads. You may place your mark anywhere you like along the scale.

Procedure We Will Use Today

1. We will drive over a predetermined course in an ordinary passenger car.
2. The trip will take about 3 hours depending on traffic conditions.
3. We will ask you to rate 33 road sections; you will not be rating an entire road.
4. It will only take about 30 seconds to drive over each section.
5. As you approach each site, the driver will call out the number of the site. Be sure you have the proper form.
6. When the driver says START, begin concentrating on what the rating should be based on how the ride feels to you.
7. Maintain your concentration until the driver says STOP.
8. At that point, place your mark on the scale and pass the forms to the person sitting in the front right seat.
9. Many sites are only 3-4 minutes apart, so make your ratings as quickly as you can.
10. This procedure will be repeated for each site.
11. There are planned rest stops but if anyone would like to stop sooner, tell the driver.

Special Instructions

1. Do not consider any of the road before or after a test section. We are only interested in a rating for a small section of road.
2. Concentrate only on the ride quality provided by the roads. Don't let the appearance of the road surface influence your ratings. Judge only how the road feels!
3. Don't be distracted by conversations in the car or by pretty scenery.
4. Don't reveal your ratings to the other raters. There are no right or wrong answers, so don't "cheat." We are interested only in your opinion which is as valid as anyone else's.
5. Be critical about the ride quality provided by the roads. If they are not absolutely perfect as far as you are concerned, be sure to give it a rating on the scale which you think best reflects the diminished quality of the ride.
6. Be aware that there are many ways that the ride could be considered less than PERFECT. The road could (a) be so bumpy that it rattles your bones and makes your teeth chatter, (b) have bumps or undulations which make the car heave up and down as if it were a boat in high seas, or (c) have other imperfections in the surface which you think detract from the ride quality.

Driving the Course

The trip to the beginning of the course always started with a drive through the parking lot of a nearby shopping center, which was generally considered to be representative of a very rough road. This served two purposes: it acted as a trial run through a site so that the subjects could practice making their ratings, and it demonstrated to the subjects how the test car (a 1981 Chevrolet Citation four-door sedan) performed on a rough road in comparison with their own vehicles. The subjects were aware of the purposes of this trial run but were not told what the rating for the parking lot should be; that is, the ride quality of the parking lot was not meant to define any particular point along the scale. All of the runs were completed in 13 (working) days.

ANALYSIS AND INTERPRETATION OF RESULTS

Data Reduction

The data were reduced by measuring, to the nearest 0.1 in., the distance between the raters' marks and the low end of the scale. Means and standard deviations were computed for each of the 33 sites for each of the three scales. In addition, the 13-step analysis procedure used by Weaver (4) was applied to derive indirect scale values. These data were transformed to provide all positive scale values. The data for the three scales are given in Tables 2 through 4.

Analysis of Variance by Rank

An analysis of variance by rank was conducted to determine which scaling method resulted in the great-

Table 2. Weaver-AASHO scale.

Site	Mean	Standard Deviation	Indirect Scale Value	Transformed Indirect Scale Value
1	1.81	0.85	-0.906	1.937
2	3.21	0.65	1.308	4.151
3	0.86	0.48	-2.367	0.476
4	1.29	0.57	-1.667	1.176
5	2.21	0.73	-0.134	2.709
6	2.62	0.71	0.538	3.381
7	3.33	0.54	1.552	4.395
8	3.89	0.57	2.355	5.198
9	2.82	0.85	0.759	3.602
10	3.84	0.63	2.233	5.076
11	3.04	0.73	1.015	3.858
12	3.82	0.50	2.145	4.988
13	3.51	0.66	1.811	4.654
14	3.59	0.69	1.892	4.735
15	1.42	0.46	1.433	4.276
16	3.53	0.53	1.766	4.609
17	1.42	0.46	-1.422	1.421
18	2.97	0.62	0.879	3.722
19	3.45	0.54	1.726	4.569
20	3.12	0.69	1.132	3.975
21	3.07	0.65	1.135	3.978
22	1.91	0.65	-0.611	2.232
23	3.14	0.66	1.264	4.107
24	0.67	0.37	-2.843	0
25	3.71	0.61	2.103	4.946
26	2.13	0.60	-0.301	2.542
27	1.39	0.65	-1.480	1.363
28	3.51	0.68	1.853	4.696
29	2.97	0.58	0.927	3.770
30	1.63	0.54	-1.101	1.742
31	2.21	0.49	-0.173	2.670
32	3.39	0.57	1.641	4.484
33	3.04	0.73	1.093	3.936

est agreement among the raters. This analysis yielded two statistics: a chi-square value and a coefficient of concordance. The chi-square value was used to test the null hypothesis that the subject yielded only random ratings on the test sections. As the data given in Table 5 indicate, the chi-square values for all three scales were quite significant because, as expected, the subjects easily detected a difference between at least two sections and rated them accordingly.

The coefficient of concordance is similar to a correlation coefficient and can be used to indicate which scaling method produces the greatest agreement among the subjects. The coefficient can range between zero and one; the closer it is to one, the better is the agreement between the subjects. Thus, this statistic can be used to discriminate between rating methods that produce significant chi-square values. The coefficients of concordance (and the related average intercorrelations) show that the three scaling methods are comparable to one another in producing reasonably good agreement among the subjects when they are rating sites. Although Holbrook's scale yielded the highest coefficient of concordance, it would be difficult to choose one method over another on the basis of these results.

Regression Analysis

Regression analyses were also conducted to get a preliminary indication of the form and strength of the relationship between the objective and subjective measures provided by each scaling method. Scatter diagrams of the mean ratings versus normalized axle displacements (inches per mile) are shown in Figures 2 through 4 and, in general, reveal linear relationships. Because a straight line accounts for 84 to 90 percent of the variance in these data, it was considered unlikely that another functional relationship could be found that would explain a significant amount of the remaining variance.

Table 3. Holbrook scale.

Site	Mean	Standard Deviation	Indirect Scale Value	Transformed Indirect Scale Value
1	2.37	0.65	0.132	2.576
2	3.93	0.43	2.415	4.859
3	1.06	0.40	-2.196	0.248
4	1.66	0.71	-1.008	1.436
5	2.70	0.69	0.542	2.986
6	3.26	0.57	1.250	3.694
7	3.89	0.50	2.421	4.865
8	4.45	0.25	3.275	5.719
9	3.06	0.67	1.184	3.628
10	4.34	0.41	3.290	5.734
11	3.79	0.50	2.187	4.631
12	4.44	0.35	3.331	5.775
13	4.16	0.58	2.917	5.561
14	4.08	0.54	2.707	5.151
15	4.17	0.46	2.958	5.402
16	3.91	0.54	2.379	4.823
17	1.41	0.53	-1.610	0.834
18	2.68	0.66	0.490	2.934
19	3.91	0.59	2.444	4.888
20	3.18	0.85	1.231	3.675
21	3.13	0.84	1.232	3.676
22	1.95	0.60	-0.727	1.717
23	3.56	0.75	1.926	4.370
24	0.85	0.45	-2.444	0
25	4.23	0.55	2.962	5.406
26	2.24	0.51	-0.085	2.529
27	1.53	0.57	-1.322	1.122
28	3.76	0.71	2.094	4.538
29	3.40	0.67	1.650	4.094
30	1.77	0.42	-0.966	1.478
31	2.35	0.64	0.046	2.490
32	3.92	0.53	2.528	4.972
33	3.70	0.65	2.124	4.568

Table 4. Nonsegmented scale.

Site	Mean	Standard Deviation	Indirect Scale Value	Transformed Indirect Scale Value
1	1.63	0.68	-1.057	1.868
2	3.47	0.79	1.606	4.531
3	0.74	0.46	-2.709	0.216
4	1.04	0.45	-2.125	0.800
5	1.84	0.75	-0.750	2.175
6	2.60	0.65	0.391	3.316
7	3.50	0.64	1.714	4.639
8	4.04	0.52	2.554	5.479
9	2.63	0.62	0.575	3.500
10	4.01	0.41	2.447	5.372
11	2.97	0.74	1.066	3.991
12	4.09	0.50	2.602	5.527
13	3.49	0.70	1.714	4.639
14	3.83	0.59	2.259	5.184
15	3.52	0.81	1.730	4.655
16	3.90	0.66	2.369	5.294
17	1.04	0.53	-2.091	0.834
18	2.73	0.68	0.615	3.540
19	3.51	0.54	1.826	4.751
20	3.33	0.77	1.459	4.384
21	3.11	0.65	1.089	4.014
22	1.82	0.79	-0.885	2.040
23	3.13	0.75	1.207	4.132
24	0.52	0.32	-2.925	0
25	4.23	0.41	3.049	5.974
26	2.31	0.76	-0.011	2.914
27	1.42	0.79	-1.547	1.378
28	3.76	0.74	2.136	5.061
29	2.78	0.62	0.653	3.578
30	1.53	0.40	-1.332	1.593
31	1.98	0.51	-0.519	2.406
32	3.29	0.68	1.395	4.320
33	3.23	1.01	1.386	4.311

Table 5. Results of analysis of variance by rank.

Scale	Chi-Square[a]	Degrees of Freedom	Coefficient of Concordance	Average Intercorrelation
Weaver-AASHO	439.95	32	0.764	0.750
Holbrook	461.61	32	0.801	0.790
Nonsegmented	427.66	32	0.742	0.727

[a] $P < 0.001$, crit $\chi^2(32) \simeq 51$.

The linear relationship found between the objective and subjective measures was somewhat unexpected. Previous research (4) had suggested that the relationship between these two measures would tend to obey Fechner's law; that is, the subjective ratings (R) should be a logarithmic function of the physical stimuli (S), or

$$R = -a \log S \quad (1)$$

where a is a scale constant.

This formulation implies that small increases in road roughness for smoother roads should cause a greater decrease in the ratings than the same amount of change for rougher roads. In other words, people should be more sensitive to smaller differences in the variation of road roughness on very smooth roads than they are on rougher roads. One possible reason that this relationship was not found is that the high end of the function is not well defined. Only five sites produced an axle displacement of 550 in./mile or more. Perhaps if a larger number of sites with roughnesses in this range had been included in the study, the expected trend would have been found.

This issue aside, however, comparison of the three correlation coefficients shows that, as in the

Figure 2. Mean rating versus axle displacement for Weaver-AASHO scale.

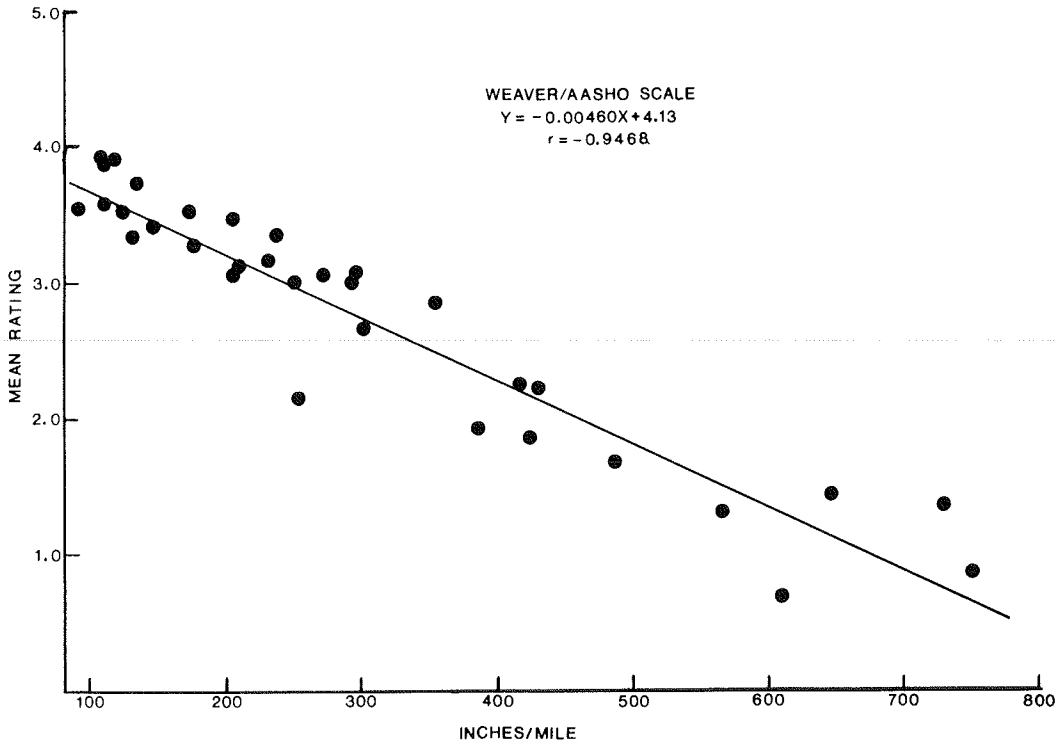
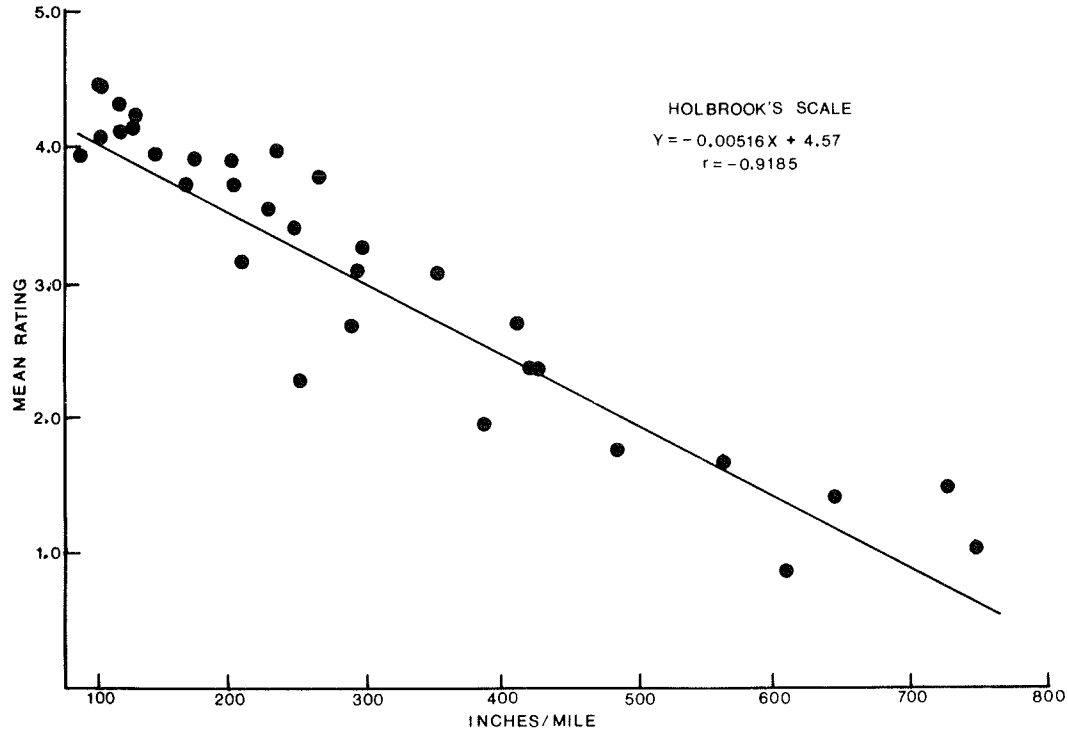
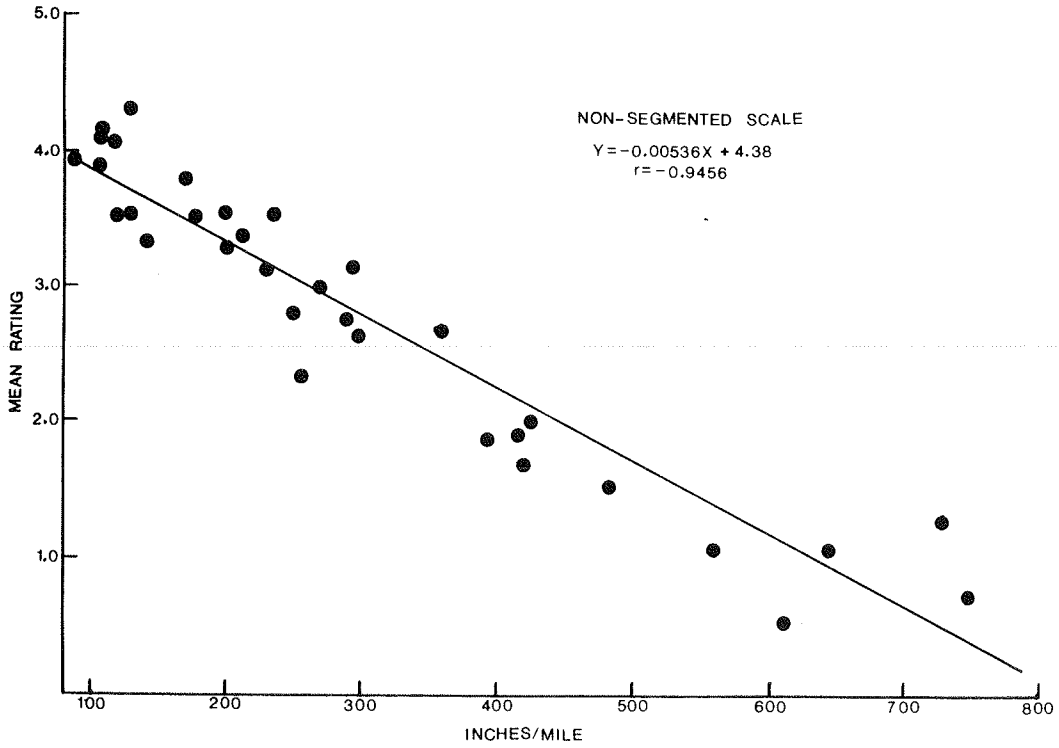


Figure 3. Mean rating versus axle displacement for Holbrook scale.



case of the results from the analysis of variance by rank, there is little reason for selecting one scaling method over the others on the basis of the correlations. The differences among the correlations can be attributed entirely to sampling variation. This conclusion is supported by the total overlap of the 95 percent confidence intervals (based on Fisher's Z-transform) for all three correlation coefficients. Therefore, there are no quantitative arguments for selecting one method over another. Consequently, for the sake of uniformity and the potential of comparing the data collected in this study with those collected over the years at NYSDOT, the Weaver-AASHO scale was selected for use in future work.

Figure 4. Mean rating versus axle displacement for nonsegmented scale.



Analysis of Direct Versus Indirect Scaling

When the transformed, indirect scale values (given in the last columns of Tables 2 through 4) were compared with the mean ratings (direct scale values), it was readily apparent that there was little difference between the two measures. A correlation of nearly 1.0 was found to exist between the two distributions (0.9981, 0.9986, and 0.9989 for the three scales). This suggests that the subjects are capable of making judgments directly on an interval scale (at least under these conditions). From a more practical standpoint, however, it appears that use of the method of successive categories is totally unnecessary in this situation. In a re-examination of Weaver's raw data, a correlation of 0.9917 was found between his direct scale values (i.e., averages) and indirect scale values (the method of successive categories), the same order of magnitude as those obtained in the research described here.

Other Analyses

The effects of starting point (learning and fatigue), seat position, sex, vehicle type normally driven, and average miles driven per year were found to be not significant. In addition, the results of a factor analysis showed that it was unlikely that the subjects rated the sites based on factors other than roughness.

One demographic variable that did prove to have a significant effect on ratings was the average number of years of driving experience (number of years licensed). Analysis showed that inexperienced drivers hesitated to give extreme ratings and were somewhat uncertain when they did rate the roads whereas more experienced drivers tended to show less variability and a willingness to give more extreme ratings. The implication of this is that care should be taken to avoid overrepresentation of inexperienced drivers on future panels.

CONCLUSIONS

The overall conclusion drawn from this phase of the study is that, if sites and panel members are carefully selected, subjects are properly instructed, and extraneous variables are controlled, quite similar results will be obtained with any of the scaling methods evaluated in this study. Thus, the choice of a scaling method can be and should be made on the basis of some attribute or attributes other than the ability of the subjects to use the scale, the correlation of the scale values with physical measures of road roughness, and other such factors.

Furthermore, no evidence was found to support the notion that the original method used by Carey and Irick violated good psychometric principles. The similarity of results among scales intended to remove systematic errors and the Weaver-AASHO scale appears to indicate that these errors do not play a significant role in the rating of ride quality (at least in a well-designed experiment). In addition, the strong correlation between direct scale values (averages) and indirectly obtained scale values also suggests that subjects are capable of making judgments directly at an interval level of measurement as Carey and Irick had implicitly assumed.

ACKNOWLEDGMENT

The research described in this paper was conducted by Ketron, Inc., for the Pennsylvania Department of Transportation. Michael S. Janoff was the project director for Ketron and Richard Cochrane was the program manager for PennDOT.

REFERENCES

1. L.F. Holbrook. Prediction of Subjective Response to Road Roughness by Use of the Rapid Travel Profilometer. HRB, Highway Research Record 291, 1969, pp. 212-226.

2. W.N. Carey and P.E. Irick. The Pavement Serviceability Performance Concept. HRB, Bull. 250, 1960.
3. B.G. Hutchinson. Principles of Subjective Rating Scale Construction. HRB, Highway Research Record 46, 1963, pp. 60-70.
4. R.J. Weaver and R.O. Clark. Psychophysical Pavement Serviceability. Soil Mechanics Bureau, New York State Department of Transportation, Albany, May 1977.
5. J.P. Guilford. Psychometric Methods, 2nd ed. McGraw-Hill, New York, 1954.
6. W.S. Torgerson. Theory and Methods of Scaling. Wiley, New York, 1967.
7. J.B. Nick and M.S. Janoff. A Comparative Study of Psychophysical Scaling Methods for Evaluating Pavement Ride Quality. Pennsylvania Department of Transportation, Harrisburg, draft, Dec. 1981.

Notice: The contents of this paper reflect the views of Ketron, Inc., which is responsible for the facts and the accuracy of the data presented. The contents do not necessarily reflect the official views or policy of the U.S. Department of Transportation. This paper does not constitute a standard, specification, or regulation.

High-Speed Road Monitoring System

P.G. JORDAN AND J. PORTER

A high-speed road monitoring system has been developed at the Transport and Road Research Laboratory. It consists of four laser sensors mounted on a 4.5-m-long beam that is supported by a two-wheeled trailer towed behind a small van. Measurements are made by the configuration of laser sensors under the control of a computer system located in the vehicle behind which the trailer is towed. Longitudinal profile, wheel-track rutting, and surface macrotexture are measured as the system travels over the road networks in the normal traffic stream; provision is being made for the measurement of road crossfall, gradient, and horizontal curvature. The principles of system operation in the different measurement modes are described and illustrated. Use of the measurements made by the high-speed system in studies of the effects of unevenness on the road user and in detecting structural deterioration of roads is described. Its potential for use in making surveys of the road network at a relatively low cost, locating areas of distress, and guiding the deployment of other, more specialized equipment is discussed within the context of the development of a cost-effective maintenance management system.

Large sums of money are being spent throughout the world on road maintenance. Effective use of these funds demands careful allocation of resources. In the United Kingdom the Report of the Committee on Highway Maintenance (1) recommended that highway authorities should use an objective maintenance rating system as the basis of regular road inspections. In answer to this recommendation, various computerized highway maintenance systems (2) have been developed and are in widespread use in the United Kingdom. All rely heavily on visual inspection to assess the condition of the road surface by reference to, for example, wheel-track rutting and surface cracking. These visual condition surveys are increasingly being augmented by input from machines that monitor various aspects of road condition more quickly than could a team of inspectors.

A high-speed monitoring system that measures a number of different aspects of road condition has recently been developed at the Transport and Road Research Laboratory (TRRL). Measurements are made by the system as it travels over the network in the normal traffic stream. The power of the equipment lies in its ability to cover a large distance each day, gather surface profile and alignment data that describe the condition of the network, and guide the deployment of other slower, more specialized evaluation equipment.

In this paper, the road monitoring system is described and its use in research on the effects of surface unevenness and its development as a component part of a total maintenance management system are discussed.

DESCRIPTION OF THE SYSTEM AND ITS USE

The high-speed road monitoring system (3), shown in Figure 1, is a laser-based system that accurately measures road surface characteristics. Its on-board computer facilities, shown schematically in Figure 2, provide both measurement control and an on-site data-processing capability. The system operates at speeds between 5 and 80 km/h, and its performance is not affected by variations in speed. With a driver and one operator, it can cover as much as 200 km of profile, rutting, texture, and road alignment parameters each day. Data are stored on floppy disks and can be either processed on site or transferred to a central mainframe computer for permanent storage and further processing.

Figure 1. High-speed road monitoring system.