Journal of Environmental Psychology (1987), 7, 233–250

CONTEXT EFFECTS IN PERCEIVED ENVIRONMENTAL QUALITY ASSESSMENT: SCENE SELECTION AND LANDSCAPE QUALITY RATINGS

THOMAS C. BROWN* and TERRY C. DANIEL

Rocky Mountain Forest and Range Experiment Station, 3825 East Mulberry, Fort Collins, CO 80524, U.S.A. and the University of Arizona, Tucson AZ, U.S.A.

Abstract

Observer groups rated the scenic beauty of forest scenes represented by color slides presented in the context of different scene mixes. The proportion of scenes from recently harvested, low scenic beauty forests compared with those from unharvested, high scenic beauty forests had a significant effect on judgments of scenes common to both sets. The effects of different scene contexts on scenic beauty judgments can result from changes in observers' perception of the scenes and from shifts in their criteria for assigning ratings. A psychophysical scaling analysis is suggested as a means for dealing with criterion shifts. Because perceptual shifts may also occur, procedures used to assess scenic beauty should be designed to reflect accurately the context to which the assessment applies, and care should be exercised in comparing experimental results obtained in different contexts.

Introduction

Perception of environmental quality has been a major area of study for environmental psychologists, geographers, and other researchers in the environment and behavior field. A considerable battery of methods for gathering and analyzing public response to environmental change has been developed for research and increasingly for application in environmental assessment, planning and management. Methods range from traditional verbal survey instruments, to a variety of perceptual comparisons, rankings or ratings, to several economic techniques, including bidding and trade-off games. The goal of all of these methods is to obtain a reliable and valid assessment of public preferences for alternative environmental conditions. Areas to which these methods have been applied include assessments of the scenic quality of various recreation experiences, visual and aesthetic effects of air pollution in parks and in cities, perceived safety in urban parks, and perceived risk from natural or technological hazards.

Each of these methods requires that assessments be obtained in a survey or experimental situation which typically differs in several respects from the 'real world' situation to which the assessment is intended to be applied. These research situations include an array of information, some of which is presented directly (e.g., in verbal instructions) and some of which is only implied (e.g., by where and by whom the assessment is being conducted). These features of the situation establish a *context* within which the respondent makes and reports his or her perceptions, judgments or choices. Because contextual factors can exert considerable influence on the respondent (Helson, 1964; White, 1975; Fischhoff, Slovic, and Lichtenstein, 1980; Einhorn

* To whom correspondence and requests for reprints should be addressed.

0272–4944 87 030233 + 18 $03.00 0

and Hogarth, 1981), the relationship between the assessment context and the intended 'real world' context is critical. The issue is one of *external validity*; do the perceptions, judgments, and choices obtained in the assessment context generalize accurately to the real world context?

Context effects are ubiquitous in human judgment-based assessment situations. For example, Wohlwill and Kohn (1973) found that migrants from rural areas judge their city as noisier and more polluted than do migrants from urban areas. And Rowe, d'Arge, and Brookshire (1980) found that people's stated willingness to pay for environmental amenities was affected by the information they were given on what others had bid. Similarly, Lichtenstein and Slovic (1971) found that people apparently use different criteria to assess gambles depending on whether they are put in a rating or a monetary response mode. Among pairs of bets, where one had a higher probability of winning and the other offered more to win, subjects tended to choose the former, but bid more to play the latter.

In landscape quality assessment, studies have investigated several contextual effects. Four components that have been found to have little effect on study results are: the sample of participants/observers used to represent a definable population of observers (Boster and Daniel, 1972; Daniel and Boster, 1976; Shuttleworth, 1980; Kellomaki and Savolainen, 1984; Brown and Daniel, 1984); the landscape representation medium, whether slides, prints, or on-site views (Boster and Daniel 1972; Daniel and Boster, 1976; Jackson and Hudman, 1978; Shuttleworth, 1980; Kellomaki and Savolainen, 1984); the observers' response format, whether paired comparisons, rankings, or ratings, when properly scaled (Buhyoff, Leuschner, and Arndt, 1980; Buhyoff, Wellman, and Daniel 1982); and the time respondents took to view the scenes (Wade, 1982).

Two potentially significant aspects of experimental context in landscape assessment are the season when photographs are presented for judgments and the labels attached to the scenes. Buhyoff and Wellman (1979) found a significant interaction between the season in which photographs were taken and the season in which they were evaluated. For example, preference for fall foliage was greater in the late summer than spring, while preference for green foliage was greater in spring than late summer. Anderson (1981) found that labels such as 'wilderness area' consistently elevated an area's scenic quality ratings, while labels such as 'commercial timber stand' consistently reduced ratings for the same area.

Another potential influence on observer judgments is the range and relative mixture of environmental conditions presented. In landscape assessments, the set of scenes previously viewed determines, in part, the context for any subsequent scene (Russell and Lanius, 1984). If the particular set of scenes presented in an experiment significantly affects the perception and/or judgment of individual scenes, this may limit comparability of experiments, and their generalizability to 'real world' experience.

Studies of human judgment and decision making have found that early questions affect responses to later ones (e.g., Turner and Krauss, 1978; Fischoff *et al.*, 1978). However, there is little, and conflicting, evidence regarding the effects of previously viewed scenes on judgments of forest scenic beauty. Daniel and Boster (1976) reported that values for four of six assessed pine forest areas remained unchanged when slides of a strip cut and a clearcut area were replaced by a burned area and a park-like pine stand. Still, they acknowledged that there were limits (untested) to

the context stability of scenic beauty measurements. Brown and Daniel (1984) found indications that ratings of forest scenes differed depending upon the mixture of recently harvested and 'natural' scenes shown in a rating session. However, they did not test for the significance of that effect.

The experiments presented below investigated the effects of previously rated forest landscape scenes on observers' ratings of the scenic beauty of subsequent scenes. Context was manipulated in three experiments by varying the proportion and distribution of forest scenes that showed obvious evidence of recent harvesting activities. The effect of this context manipulation was assessed by observing changes in the ratings of other forest scenes common to all presentations. In the first experiment a strong context manipulation was implemented. The second and third experiments introduced progressively weaker manipulations of scene context, but ones more typical of most actual forest viewing situations.

Experiment I

The approach of this experiment was to establish a strong contextual dichotomy by preceding a common set of forest scenes by one of two very different scene sets. The *low* scenic beauty scene set, rated by one panel, contained color slides that all exhibited obvious evidence of recent harvesting activities and that had generally been rated very low in scenic beauty in previous assessment studies. Scenes in the *high* scenic beauty set, rated by a separate panel, showed no obvious evidence of management activity and had generally been rated high in scenic beauty in previous studies. Scenes of the *common* set, rated by both panels, showed no obvious evidence of management activity, and had been rated from low to high in scenic beauty in previous studies, with an average approximately midway between the high and low sets. The common scenes were presented in the same order to both panels.

Method

Two separate panels of 26 observers were shown a total of 130 photographs (color slides) of forest landscape scenes. All observers were undergraduates at the University of Arizona who had responded to an announcement and each received credit toward a class research participation requirement. The first 80 slides differed between observer panels and were used to establish different contexts for a subsequent common set of 50 scenes that were presented to both panels.

The 80 slides that established the low scenic beauty (SB) context were selected from a pool of approximately 600 slides taken in a large heterogeneous area of ponderosa pine forest in northern Arizona. Slides were taken in the summers of 1980 and 1981, immediately after a selective-cut timber harvest that removed approximately 40% of the overstory trees (see Brown and Daniel, 1984). The low SB context slides were selected to show obvious evidence of recent harvest activities, such as cut limbs, stumps, and disturbed ground, and were required to be in the lowest range of scenic beauty, as determined by previous assessments of the entire pool of slides.

The 80 slides that established the high SB context were selected from a pool of 1,500 slides taken in the same area in the summer of 1979, prior to any harvest activity. These slides were selected from the highest range of the previously established scenic beauty distribution.

The 50 common slides were taken in the same general forest area as the previously

shown 'context' slides. Common slides were selected to represent the full range of preharvest forest conditions, and their individual scenic beauty values ranged from low to high.

Instructions to each observer panel were identical and referred to the fact that public perception of scenic beauty is one of many important considerations in the management of the national forests. They were told that their participation in the experiment would help to extend our understanding of public perceptions and preferences for scenic beauty in forest landscapes. The 10-point scenic beauty rating scale was described, ranging from 'very low scenic beauty' (1) to 'very high scenic beauty' (10), and observers were instructed to assign one rating to each scene and to try to use the full range of the rating scale.

Prior to rating the scenes, a sample of 15 of the 80 context scenes appropriate to the panel was shown briefly (3 seconds each) to indicate 'the type of scenes' they would be rating. Observers were told to view the scenes and to think about how they would use the 10-point scenic beauty scale to rate them.

Immediately following the preview, the scenes were shown one at a time for 5 seconds each. Each observer independently rated each slide as it was presented. After the appropriate set of 80 experimental condition slides had been rated, the 50 common slides were presented with no interruption in the procedure. Both low and high SB context panels saw the common slides in the same order. Observers recorded their ratings of all slides on a specially designed sense-mark form which was subsequently optically scanned to enter the ratings into a computer data file.

Results

The reliability of the ratings assigned by the low and high SB context panels was assessed separately by the intra-class correlation coefficient (Ebel, 1951). This measure indicates the expected mean correlation between panels for panels of the same number of observers sampled from the same population. The reliability coefficients were 0·91 for the high SB context panel and 0·89 for the low SB context panel.

Ratings of the slides used to establish the contexts confirmed the basis for their selection for this study; the low SB context slides were rated considerably lower in scenic beauty than the high SB context slides, as shown in Table 1. On the one hand, this seems an expected result. However, if observers had followed instructions fully there would have been little difference in the mean ratings of the two groups. Both panels were told to 'use the full range of the rating scale' in responding to the set of slides they were shown, and that only 'relative differences' were of concern. The fact that there was a substantial difference in the average ratings assigned by the panels indicates that each was applying, to some extent, an extra-experimental or 'absolute' standard in rating the scenes.

The principal focus of this experiment was on the ratings of the common slide set. As shown in Figure 1, the preceding slides had rather strong effects on ratings of the common slides. The common slides were rated higher after the low scenic beauty slides than after the high scenic beauty slides. This effect proved highly significant in an ANOVA, the main effect of context (averaged over the 50 common slides) yielding $F(1, 50) = 10{\cdot}96$, $P = 0{\cdot}002$. Closer inspection of Figure 1 reveals that the context established by the first 80 slides declined steadily over the 50 slides of the common set. The difference between ratings by the low and high panels averaged 2·26 rating points over the first ten common slides and declined to 0·92 over the last

TABLE 1
Mean ratings of context and common slides

	Context		
	High SB (Pre-harvest)	Mixed SB	Low SB (Post-harvest)
Experiment I			
Context slides	6·05		4·83
Common slides *	4·85		6·20
Experiment II			
Context slides	5·80	5.64‡	5·24
Common slides A †	4·39	5·41	6·59
Common slides B *	4·55	5·20	5·99
Experiment III§			
Context slides	5·57		4·85
Common slides †	4·89		5·36

* Presented in a block at end of session
† Presented mixed in with context slides
‡ High and low SB context slides received mean ratings of 7·00 and 4·27, respectively
§ Averaged from two replications

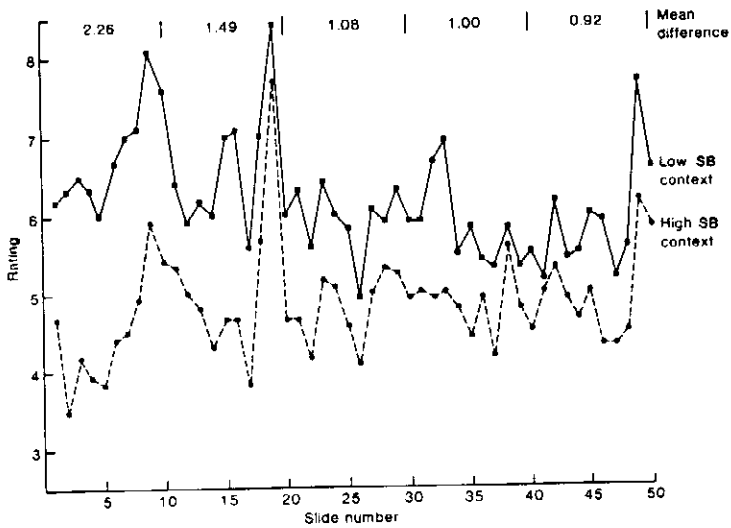

FIGURE 1. Mean ratings of common slides rated in the context of high and low scenic beauty slides.

ten slides. The context-by-slide position interaction proved significant in the ANOVA, $F(49, 2450) = 3{\cdot}16$, $P < 0{\cdot}0001$.

While the effect of the context established by the 80 context slides declined over the common slide set, it is important to note that there was still a substantial effect on the ratings of the last 10 slides. Presumably the context effect would eventually

dissipate or, more correctly, be replaced by the context established by the current slide set.

The Pearson correlation coefficient of ratings of the common slides, across the two contexts, was 0·59 (Table 2), substantially lower than the intra-class correlation coefficients of the individual panels, which were 0·89 and 0·91, as reported above. The considerable difference between the two measures suggests that the relative ratings of the individual slides differed across the two contexts.

TABLE 2

Intra-class correlation coefficients for high and low SB context panels compared with Pearson correlation coefficients across panels

	Intra-class correlation coefficients		
	High SB Context	Low SB Context	Pearson correlation coefficients*
Experiment I	0·91	0·89	0·59
Experiment II	0·89	0·85	0·67
Experiment III			
Replication 1	0·93	0·89	0·85
Replication 2	0·90	0·85	0·85

* Correlation of mean (across observers) ratings of 50 common slides from the high SB context panel with corresponding ratings from the low SB context panel

Presenting all of the context scenes first was intended to establish firmly the context before introducing the common slides. Under these conditions, effects on ratings of the common scenes were quite large and, while progressively declining over the 50 slide set, were persistent. The second experiment utilizes very similar methods to inspect further the development and maintenance of the context effect.

Experiment II

This experiment used the same slides as did Experiment I, but here the slides were arranged differently. Twenty-five of the 50 common slides were mixed in with the 80 slides used to establish the contexts. The other 25 common slides were included at the end of the presentation. This arrangement enabled assessment of the context effect on the interspersed common slides under 'continuously reinforced' conditions. It was expected that the effect of the context differences on the interspersed common slides would not decline as it did in Experiment I for the common slides presented at the end. Also, mixing the common slides in with the context scenes might be expected to 'dilute' the context difference, as compared with Experiment I.

A third observer group was shown a mixture of 40 low SB and 40 high SB context scenes along with the 25 interspersed common scenes. Again, the other 25 common scenes were presented at the end. This arrangmeent was expected to produce ratings for the common scenes that fell in between those of the other two groups.

Method

This experiment was identical to Experiment I in terms of stimuli, recruitment of observers, general slide presentation procedure, previous slides, and response scale. The difference was in the order and mixtures in which the slides were presented.

Three separate groups of 27 observers each rated a set of 130 slides of forest scenes. Each group saw 80 'context' slides with 25 of the common slides randomly mixed in, followed by the remaining 25 common slides. The context slides, used to establish the different contexts, were presented as follows: one group saw the 80 *high* scenic beauty (pre-harvest) slides; another group saw the 80 *low* scenic beauty (post-harvest) slides; and the third group saw a *mixture*, a random selection of 40 high scenic beauty and 40 low scenic beauty slides. The 25 of the 50 common slides that were mixed in with the context slides were chosen by selecting every other slide from the set of 50 used in Experiment I. They were randomly positioned among the 80 context slides, but the placement of the common slides was the same in all three presentations.

Results

Reliability was again assessed by the intra-class correlation coefficient. This measure was 0·89, 0·85, and 0·96 for the high, low, and mixed SB context panels, respectively, indicating that other panels from the same population would yield very similar ratings.

Ratings of the context slides again reflected the basis for their selection, with the slides used to establish the low SB context rated lower than those used to establish the high SB context (Table 1). Furthermore, as shown in Figure 2, the three different contexts had strong effects on ratings of the common slides. These effects were similar to those observed in Experiment I; the common slides were rated higher when shown with low scenic beauty slides than when shown with high scenic beauty

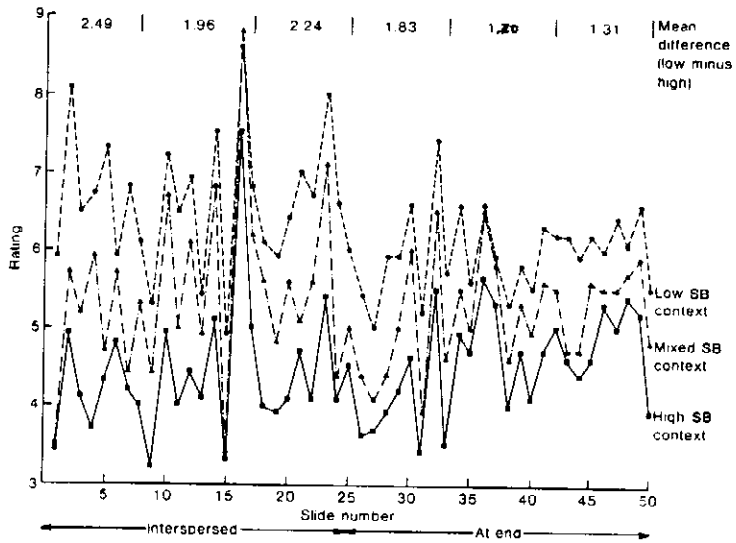

FIGURE 2. Mean ratings of common slides rated in the context of high, low, and mixed scenic beauty slides.

slides. In addition, the ratings of the common slides in the mixed high-low SB context were consistently between the two extremes. Analyses of variance showed that the differences in ratings of the common slides, across the three contexts, were significant, with $F(2, 78) = 25{\cdot}25$, $P < 0{\cdot}0001$, for the 25 slides mixed in with the experimental condition slides, and $F(2, 78) = 7{\cdot}95$, $P = 0{\cdot}0007$, for the last 25 slides.

As expected, the effect of the context difference on ratings of the interspersed common slides did not dissipate as much as for common scenes presented after the context scenes. Furthermore, the context effect was strong from the beginning of the experiments. Apparently, viewing 15 of the context scenes before beginning to rate scenes and rating the initial context scenes, coupled with the instruction to 'use the full range of the rating scale,' was sufficient to establish a significant context effect. The relative importance of these three aspects of the procedure in establishing the context effect cannot be determined from the data.*

Interspersing common slides with the context slides, at a density of one common to three experimental, appears to have diluted somewhat the context effect on the ratings of the common slides presented at the end. In Experiment I, ratings of the first 25 common slides (all presented after the context scenes) had a mean difference of 1·70 between high and low SB context groups. The average difference for ratings of the common slides that followed the context scenes in Experiment II was 1·44.

The correlations of ratings of the common slides, across pairs of panels, were 0·67 (high versus low SB context), 0·84 (high versus mixed SB context), and 0·76 (low versus mixed SB context). Two of these are considerably lower than the intra-class correlation coefficients (0·85 to 0·96), indicating that the ratings differed across those contexts (Table 2).

Experiment III

The context slides of the first two experiments were selected to establish very distinct differences between the high and low SB contexts. The differences were rather extreme compared with the differences among areas that are typically viewed by visitors to Arizona ponderosa pine forests. The approach in the third experiment was to observe the effects of a substantially more subtle difference in scene context, established by the presentation of two different mixtures of pre-harvest and post-harvest scenes. The slides included in the mixtures were selected without regard to the scenic beauty of the scenes they depicted, and much more closely approximated naturally occurring mixtures than those of the first two experiments. The study design allowed assessment of the effect of these mixtures on a common set of pre-harvest slides.

Method

The methods of this experiment were identical to those of Experiments I and II in terms of recruitment of observers, general slide presentation procedure, and response scale. The slides used in this experiment came from the same slide pools as those of Experiments I and II, but different criteria were used to select the slides.

* The instruction requesting that observers use the full range of the rating scale is commonly used where relative differences are of primary interest, and is intended to avoid compression of responses within a narrow range of the rating scale. From experience in previous experiments, we do not expect that the instruction had a major effect on observers' responses. We plan to test for the importance of the instruction and preview procedures in future work.

The original slide pools, from which the slides used in the three experiments were selected, were obtained by taking four slides at each of a large number of randomly selected points in the forest. Some points were photographed before (1979) and again shortly after (1980 and 1981) harvest. The harvest selectively removed about 40% of the trees, but not uniformly, so that some scenes remained essentially unchanged.

In this experiment, for the post-harvest condition, 30 of the sample points inventoried after harvest were randomly selected, with the restriction that at least two of the four slides showed evidence of management activity. For the pre-harvest condition, the same 30 points were used, but of course the slides were taken before harvest. No attempt was made, as it was in the first two experiments, to select scenes of low or high scenic beauty. Some slides in the post-harvest set showed scenes of high scenic beauty, with little or no harvest activity apparent, and some slides of the pre-harvest set showed scenes that were low in scenic beauty. Thus, the pre-harvest condition was represented by 120 slides (four slides at each of 30 points) showing no evidence of harvest activities and the post-harvest condition was represented by 120 slides, with almost 60% of these showing some evidence of harvest activities. Also, 15 preview slides were randomly selected from each group of 120.

In addition to the 120 slides used to establish the pre-harvest and post-harvest contexts, 25 common slides (a subset of the 50 common, pre-harvest slides in Experiments I and II) were interspersed. The same 25 common slides were shown to each panel. Slides were randomly ordered except that a common slide appeared in every fourth position. Thus, each observer viewed 15 preview slides and then viewed and rated 145 slides.

The entire experiment was repeated, with the slide order unchanged, by a different experimenter. Separate observer panels of 20 observers each were used for each condition and each replication of the experiment, requiring a total of four panels.

Results

Two measures of group-to-group reliability were examined. First, the intra-class correlation coefficients varied from 0·85 to 0·93 for the four panels. Second, Pearson product moment correlations between replications for pairs of observer groups that saw the same slides were 0·83 and 0·87 for the two pairs of panels.

Ratings of the context slides reflected the relative proportions of slides showing effects of harvest activity. On average, the pre-harvest condition was rated higher than the post-harvest condition (Table 1).

The effect of the context differences on ratings of the common slides was generally similar to the effect obtained in the first two experiments. The ratings of the common slides averaged slightly less when shown with slides of the pr-harvest condition than when shown with slides of the post-harvest condition. As shown in Figure 3, which shows average ratings across both replications, the generally more attractive, pre-harvest condition slides tended to deflate the ratings of the common slides, while the less attractive post-harvest condition slides tended to inflate the ratings of the common slides. The difference in mean ratings of the common set, however, was about three times larger in the first two experiments than in this comparison. Furthermore, analysis of variance showed that the mean difference in common slide ratings was not significant, $F(1, 76) = 3{\cdot}83$, $P > 0{\cdot}05$. Thus, for this more subtle, naturally occurring context distinction, changes in the mixture of slides from harvested and

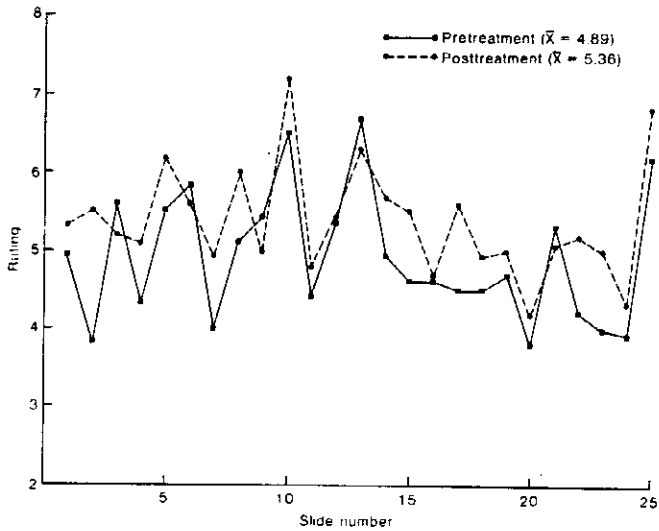

FIGURE 3. Mean ratings of common slides rated in the context of pre-treatment and post-treatment slides.

unharvested areas presented to observers did not significantly affect their mean scenic beauty ratings.

The correlation of ratings of the common slides across the two contexts was 0·85 for each replication. This is similar to the intra-class correlation coefficients (0·85 to 0·93), suggesting that ratings of the individual scenes were not significantly different across contexts (Table 2).

Discussion

Experiments I and II showed that ratings of forest scenes are influenced by the nature of previously viewed and rated scenes. In both experiments, a set of scenes that was common to all context conditions was rated lower in the high SB context than in the low SB context. The slide mixture of Experiment III, where context differences were more subtle, produced similar but non-significant effects on the common scene ratings.

These findings have important implications for comparisons among different environmental assessments, and for the external validity of such assessments. Clearly, when the same forest scenes were assessed in different contexts, scenic beauty ratings were quite different. Which assessment is to be believed? Obviously, neither outcome can be viewed as being 'correct' in an absolute sense; rather, each result has meaning only with respect to the context in which it was obtained. The critical issue is the relationship between the assessment context and the 'real world' context to which the assessment is intended to apply, i.e., the external validity of the assessment.

Context effects can result from changes in observers' *perception* of the assessed environments and/or from shifts in their *criteria* for assigning ratings to the alternative environments. This two-part model, introduced by Daniel and Boster (1976), follows the psychophysical models developed by Thurstone (see Torgerson, 1958;

Nunnally, 1967) and extended by signal detection theory (Green and Swets, 1966), and is only summarized here.

In simplified terms, the model postulates that implicit *perceptual processes* encode the features of the environmental stimulus (e.g., the color slide of a forest scene) and translate them into a subjective impression of the 'attractiveness,' 'aesthetic quality' or, in the present experiments, the 'scenic beauty' of the stimulus. This perceptual process is strongly influenced by the features of the environment in interaction with the sensory and perceptual system of the observer.

Relationships among the characteristics of the set of environments being assessed and the setting in which they are presented can affect the perceptual component by focusing observers' attention upon different features of the environments, leading observers to emphasize certain features and to overlook others. Evidence of such 'perceptual cueing' has been reported by Buhyoff and Leuschner (1978) and Buhyoff and Riesenman (1979). In these studies, information about the cause of apparent damage to trees led observers to be more sensitive to insect damage differences in their ratings of forest scenes. Brown and Daniel (1984) found evidence that the relative proportion of scenes exhibiting different forest management-related features affected the importance of those features in determining observers' ratings of the scenes. Observers may have cued more on the amount of cut tree limbs, tops and stumps when a larger proportion of the scenes being judged represented immediate post-harvest conditions.

To produce an overt response, the perception of the environment must be referenced to a *judgment criterion* scale which, in the present experiments, was a 10-category rating scale. Depending upon the judgment criterion being applied, the observer assigns an overt rating to the forest scene. For example, the perceived scenic beauty of a scene may be sufficient to meet the criterion for the seventh category, but not high enough to reach the eighth category, so the observer would assign a rating of seven to that scene. Criterion scales can vary between observers, or the same observer may change scales from one time to another. As a consequence, responses to the same environment (e.g., ratings of a particular forest scene) can differ even though the perception of the scene is the same. When differences in ratings are due only to criterion shifts, the resulting ratings will be monotonically related, and usually the differences can be resolved by a linear transformation of the rating distributions.*

Context can have substantial effects on the judgment criteria that observers use in expressing their perceptions in the assessment situation. Direct instructions, verbal labels (Hodgson and Thayer, 1980; Anderson, 1981) and more subtle 'social influences' (Simpson *et al.*, 1976) have produced large changes in ratings of scenes that appear better attributed to criterion than to perceptual differences. Differences in the range and proportions of environmental conditions represented in the assessment set could also influence the observer's judgment strategy. For example, the context manipulations in the experiments reported above might tend to require observers to shift their criteria for rating the forest scenes presented: criteria for

* Forced-choice (e.g., paired-comparison) and rank-order procedures are generally assumed to by-pass the criterion component. In these procedures the observer's response is only dependent on the relative perceived value of each alternative (Torgerson, 1958; Egan and Clark, 1966; Hull, Buhyoff and Daniel, 1984).

rating the differences among the high SB context scenes would be inappropriate for rating the low SB context scenes.

There is no direct way to observe either the perceptual process or the judgment criterion process; both are implicit psychological constructs that are only indirectly indicated by overt behaviors of observers. How, then, can changes in perception be distinguished from changes in judgment criteria? A number of 'psychophysical scaling' procedures (Torgerson, 1958; Nunnally, 1967) have been developed in an attempt to answer such questions. It is beyond the scope of this paper to discuss these models in detail, but a simplified graphic analysis is presented that captures some of the more important aspects of a psychophysical scaling analysis appropriate to the problem investigated in the experiments described above. This analysis focuses on the differences in mean ratings across panels, as reported in Table 1.

To simplify the illustration, the levels of perceived scenic beauty for the high, low, and common scene sets are treated as fixed values; in typical scaling models the perceived values would be treated as normally distributed variables (Torgerson, 1958; Hull *et al.*, 1984).* Furthermore, as indicated in Figure 4, the perceived scenic beauty values for the high and low SB context scenes are assumed equidistant above and below the values for the common scene set. The equidistant relationship for high and low scenes is largely consistent with the data of Experiments I and II; high scenes were rated an average of 1·27 above, and low scenes 1·21 below, the common scenes.

The rating assigned to a scene is determined by the relationship between the perceived scenic beauty level and the judgment criterion scale being applied by the observer. Two hypothetical criterion scale strategies are shown in Figure 4, an 'absolute' and a 'relative' scale strategy. The 'absolute' strategy assumes that observers come to the assessment situation with pre-established aesthetic criteria that are not affected by the particular assessment context. The 'relative' strategy assumes that observers develop or adjust their judgment criteria to fit the particular assessment context.

As indicated in Figure 4, application of the 'absolute' strategy results in the assignment of the same overt rating for a given level of perceived scenic beauty regardless of context. That is, assuming constant perception, the common scenes would be expected to get the same rating (denoted by 'R_c' in the figure) whether judged in the context of the high or the low scenes. The high and low scenes would be rated correspondingly above and below the common scenes regardless of the context in which they were judged.

In contrast to the 'absolute' strategy, application of the 'relative' strategy results in identical ratings for the high and low scenes in their respective contexts. Observers would shift their judgment criterion scale up or down so that the midpoint of the 10-category criterion scale (6·00) would be centered at the mean of the scenes that they were rating. As a consequence, the common scenes, which are the same in both contexts, would be given very different ratings. As shown in Figure 4, the common

* The assumption that the perceived scenic beauty of each scene set remains constant across experimental conditions panels is consistent with the approach used in signal detection theory-based studies. The approach is generally to treat perception as a relatively stable function of stimulus characteristics (the respective scenes are the same across the experiments), and attempt to attribute situational changes in response (ratings) to criterion shifts. Then, if the data shows this assumption to be untenable, a change in perception may be indicated.

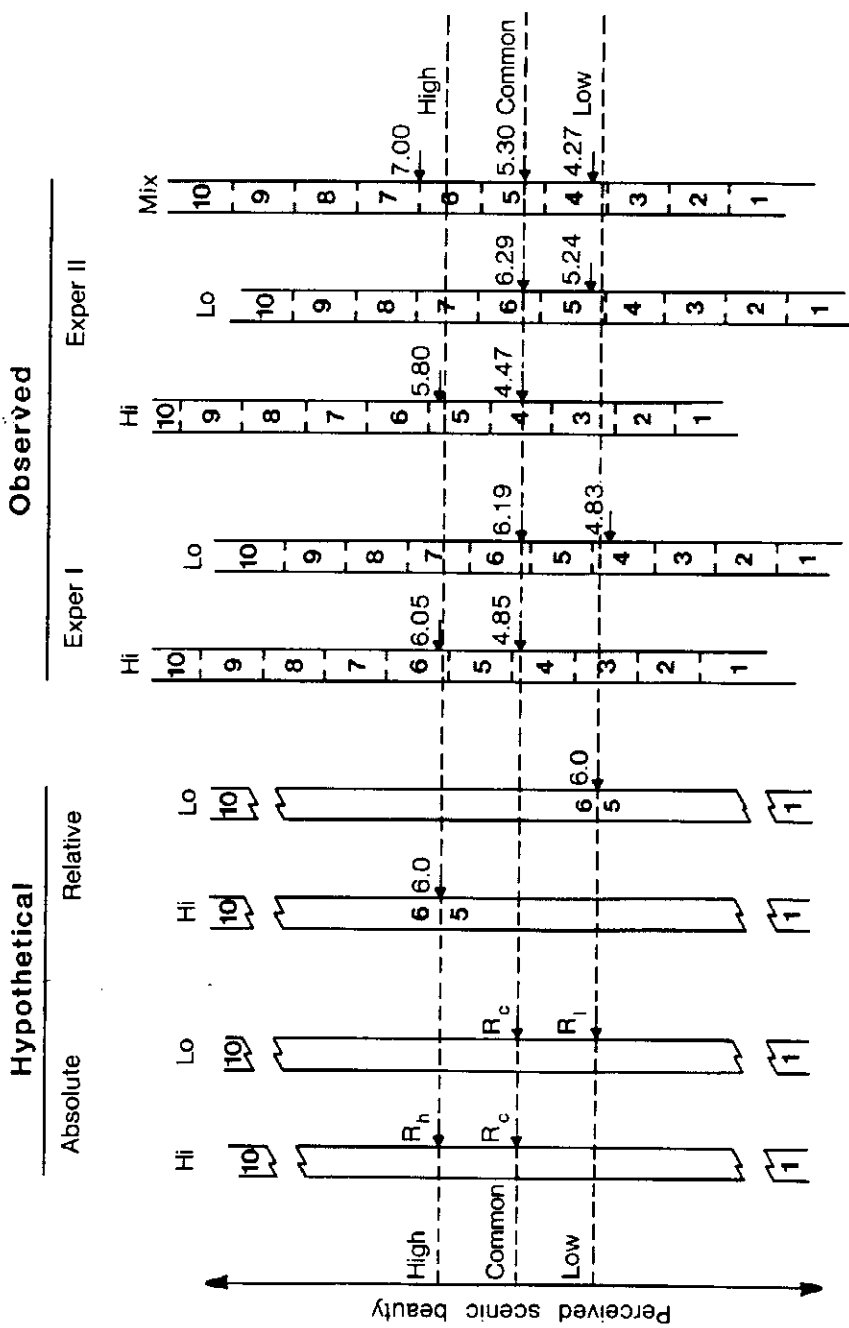

FIGURE 4. Mean ratings of high, low, and common slides according to hypothetical and observed strategies. 'Hi', 'Lo', and 'Mix' indicate the high, low, and mixed SB contexts, respectively, in which the common slides were rated. R_h, R_c, R_l refer to mean ratings of the high, common, and low scenic beauty slides, respectively.

scenes would be expected to have a mean rating less than 6·00 in the high SB context and a mean rating correspondingly more than 6·00 in the low SB context. The actual magnitude of the difference in ratings of the common and the context scenes would depend upon the distance between the scene sets on the perceived scenic beauty scale and on the size of the categories (the 'interval' size) on the judgment criterion scale.

To simplify the application of the two-component model to the actual results of Experiments I and II (Experiment III was excluded because different scene sets were used), an additional assumption was required. The judgment criterion scales of all panels were assumed identical and to each have 10 ordered categories of equal size (although the first and last categories would be 'open' at the ends of the scale). The actual interval size used in Figure 4 was based on the average of the observed differences in ratings between the common and the high SB context scenes, and the common and low SB context scenes, across the five observer panels.

The ratings shown in Figure 4 for each of the panels are the obtained mean ratings reported in Table 1. The location (up or down) of the assumed criterion scales for each panel was determined by the mean rating of the common scenes. That is, the relative position of the scale was shifted up or down as necessary to fit the data for the scenes that were rated by all observers to the assumed perceived scenic beauty of those scenes.

The obtained relationships may be compared to those predicted by the 'absolute' and 'relative' criterion strategies. The results of the two experiments indicate that neither the 'absolute' nor the 'relative' criterion was used exclusively. The very different mean ratings across contexts for the common slides suggests that a 'relative' criterion was used, but differences across contexts for the high and low slides suggests that one cannot rule out some element of an 'absolute' criterion.

Specifically, the results of Experiment I indicate that observers in the high and low SB context conditions did adopt different judgment criteria (Figure 4). For the high SB context, the mean rating of the high scenes was 6·05, very near the midpoint (6·00) value expected for the 'relative' criterion. The average rating for the common scenes was 4·85, considerably below the midpoint value and close to the expected 4·75 (given the criterion scale interval size assumed). The observers in the low SB context, on the other hand, do not appear to have adjusted their judgment criteria sufficiently to fit the low SB context scenes to the 'relative' criterion strategy. The mean rating of 4·83 for the low scenes is well below the 6·00 expected under the 'relative' strategy. The common scene rating of 6·29 is also below the 7·25 that would be expected for the 'relative' strategy.

Results of Experiment II show a very similar pattern. Responses in the high SB context again approximated those expected for the 'relative' criterion strategy. Indeed, the mean rating for the high SB context slides (5·80) suggests a slight 'overshift;' the criterion scale was shifted up about 0·20 units as compared to the hypothetical 'relative' criterion scale. In contrast, the low SB context observers made only modest adjustments in the direction of a relative criterion scale. Their scale was about 0·75 units high as compared to the hypothetical 'relative' scale. The minor differences from Experiment I may, in part, be attributed to the change in procedure; half of the common scenes were mixed in with the context scenes in Experiment II.

The same reluctance to 'compromise aesthetic standards' to accommodate the low SB context scenes is indicated for the mixed presentation group in Experiment

II. A 'balanced' scale, centered at the midpoint of the perceived scenic beauty values assumed in Figure 4, would produce a mean rating of 6·00 for the common scenes. Apparently, however, the criterion scale applied by the mixed group was more like the high relative scale than the low. The common scenes achieved a mean rating of only 5·30. Because the mixed condition data are based on only half of the high and low SB context scenes, some of this shift in ratings may be due to scene sample differences.

As the analysis in Figure 4 indicates, the context effects on mean response per panel observed in the present experiments can, for the most part, be attributed to a simple shift in the location (up or down) of the judgment criterion scale. The assumption of equal interval sizes was largely confirmed, as indicated by the close correspondence of the mean ratings of the high and low SB context scenes to their respective assumed perceived scenic beauty levels. Thus, a relatively simple adjustment, such as expressing ratings in each context condition as a difference from the common scene mean, would eliminate most of the differences between contexts; it does not seem necessary to conclude that the observers on average *perceived* the common scenes differently in the two contexts.

More sophisticated methods for dealing with criterion scale problems include an array of 'standard score' computations, such as 'z scores,' and any of several methods developed by Thurstone (see Torgerson, 1958; Nunnally, 1967). These methods are designed to account for changes in criterion scale interval sizes, as well as the simple 'origin' shifts shown in Figure 4.

The 'Scenic Beauty Estimation' (SBE) method (Daniel and Boster, 1976) is an extension of traditional psychophysical scaling methods to landscape quality assessment (Hull *et al.*, 1984).* The results of applying the SBE scaling to the ratings obtained in Experiments I and II are shown in Figure 5.† The mean ratings differed considerably across the contexts; the common scene averages ranged from a low of 4·47 in the high SB context to a high of 6·29 in the low SB context, with corresponding variations in the ratings for the context scenes. In contrast, the corresponding mean SBE values, which are intended to represent only differences between the scene sets on the perceived scenic beauty scale, are very similar across contexts in both experiments.

One goal of psychophysical scaling methods is to achieve an index that reflects consistent changes in perception regardless of temporary situational (context) changes in judgment criteria. On the basis of the panel means for the above reported experiments, the SBE method would seem to have achieved this goal; the observed SBE values for the scene sets remained quite stable, in spite of substantial shifts in contexts (and ratings). However, the fact that mean SBEs are stable across contexts

* As in Figure 4, responses were related to the assumed perceived scenic beauty of the three conditions by computing a common interval size for the response scales. The interval size for the SBE scales of Figure 5 was based on the observed differences between the common and high SB context scenes, and the common and low SB context scenes, across all observer panels.

† The 'by slide' SBE procedure was used to scale the rating responses. Using this procedure, (1) ratings of each observer group were converted to a set of standardized scores, one per slide, based on the frequency distribution of the observers' ratings for a given slide, (2) the mean of the standardized scores for the common slides for each observer group was then subtracted from the standardized scores for each slide the group rated to yield a standardized difference score for each slide, and (3) the difference score was multiplied by 100 to yield SBEs for each slide. Mean SBEs were then computed for the common and context conditions.

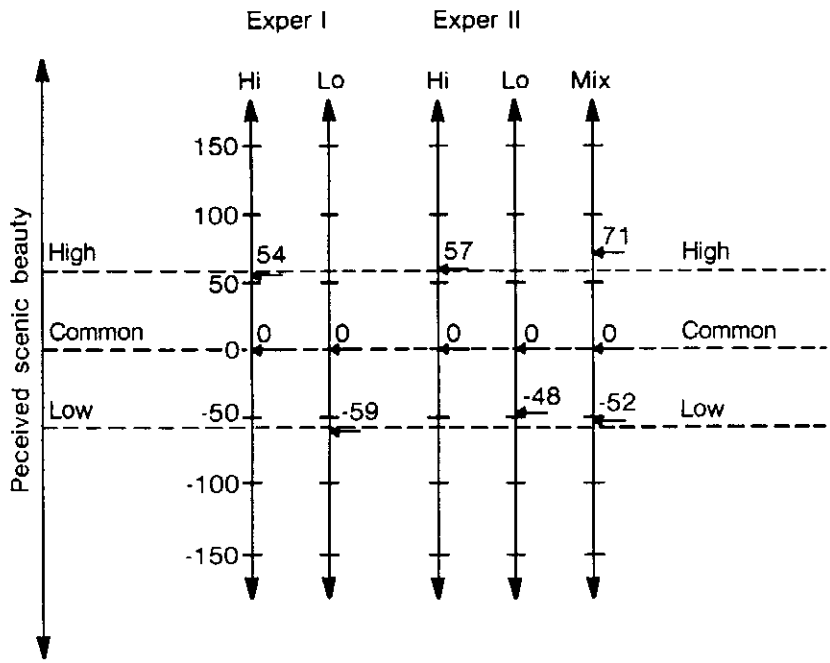

FIGURE 5. SBEs of high, low, and common slides. 'Hi', 'Lo', and 'Mix' indicate the high, low, and mixed SB contexts, respectively, in which the common slides were rated.

does not insure that individual slide values will be comparable. Correlations of common slide SBEs from the high and low SB context conditions were 0·59 and 0·67 in Experiments I and II, respectively, indicating that substantive differences between values for individual common slides remain after the SBE scaling. This pattern of results could indicate context-induced changes in perceptions.

Closer inspection of the data presented in Figures 1 and 2 suggests, however, that a more complex criterion effect may be involved, one that is not accounted for by the SBE or standard z-score adjustments of the ratings. In particular, there is an obvious order effect in Figures 1 and 2; the difference between high and low SB context ratings progressively declines over the 50 common slides. When serial order is included as a covariate in a multiple regression between high and low SB context SBEs for the common slides, multiple correlations improve to 0·77 and 0·82 for Experiments I and II, respectively. These values are still short of the internal reliability coefficients in Table 2, but indicate that most of the difference between contexts can be attributed to shifts in judgment criteria.

It seems clear that observers in the above experiments used different judgment (rating) criteria in the contexts established by the high and low slide sets. This conclusion is supported by the fact that differences in ratings of the common slides can largely be accounted for by relatively simple, monotonic scale adjustments. However, context-induced changes in perception cannot be dismissed entirely; even with the modifying effect of serial order taken into account, there were still differences between common slide ratings in the two contexts. Moreover, the attribution of the observed context effects to the judgment criterion component of the

two part model is substantially directed by *a priori* theoretical considerations. Perceived environmental quality is postulated to be a relatively consistent process that is strongly related to the features of the environmental stimuli. Judgment criteria, on the other hand, are assumed to be less stable and more strongly affected by personal, social and situational (context) factors. Of course, these data might also be fitted by a model which assumes that judgmental standards are stable, and that perception is the more volatile of the two components. A choice between these two models cannot be made on empirical grounds.

Conclusion

The context effects demonstrated above have important implications for methods of assessing perceived environmental quality. The comparability of different environmental assessments and the external validity of assessment results (i.e., the extent to which the results may be generalized to the 'real world') may potentially both be compromised by changes in context. Sometimes the context change affects only judgment criteria. Methods that account for criterion effects and provide standardized indices can help to make assessment results more generalizable. However, assessment contexts may have effects that are not addressed by scale manipulations—such as context-specific cueing on different aspects of a multidimensional stimulus. In these cases, the choice of contexts must be based on the goals of the assessment. Because the true nature of a context effect cannot necessarily be determined, the assessment context should, to the extent possible, be made to match the 'real world' context to which the results are to be applied. In any case, further research is needed to determine the impact of the assessment context on judgments and to better understand how and to what extent perceptual cueing occurs.

Acknowledgements

We thank the reviewers, whose thoughtful comments have helped make this a better paper, and Joanne Vining and Sara Kocher for their assistance to this research.

References

Anderson, L. M. (1981). Land use designations affect perception of scenic beauty in forest landscapes. *Forest Science*, **27**, 392–400.

Boster, R. S. and Daniel, T. C. (1972) Measuring public responses to vegetative management. *Proceedings of the Sixteenth Annual Arizona Watershed Symposium*. Arizona Water Commission. Phoenix, Az., pp. 38–43.

Brown, T. C. and Daniel, T. C. (1984). *Modeling forest scenic beauty: concepts and application to ponderosa pine*. (USDA Forest Service Research Paper RM-256), Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.

Buhyoff, G. J. and Leuschner, W. A. (1978). Estimating psychological disutility from damaged forest stands. *Forest Science*, **24**, 424–432.

Buhyoff, G. J. and Riesenman, M. F. (1979). Manipulation of dimensionality in landscape preference judgments: a quantitative validation. *Leisure Science*, **2**, 221–238.

Buhyoff, G. J. and Wellman, J. D. (1979). Seasonality bias in landscape preference research. *Leisure Sciences*, **2**, 181–190.

Buhyoff, G. J., Leuschner, W. A. and Arndt, L. K. (1980). Replication of a scenic preference function. *Forest Science*, **26**, 227–230.

Buhyoff, G. J., Wellman, J. D. and Daniel, T. C. (1982). Predicting scenic quality for mountain pine beetle and western spruce budworm damaged forest vistas. *Forest Science*, **28**, 827–838.

Daniel, T. C. and Boster, R. S. (1976). *Measuring landscape aesthetics: the scenic beauty method.* (USDA Forest Service Research Paper RM-167), Rocky Mountain Forest and Range Experiment Station, Fort Collins, CO.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, **16**, 407–424.

Egan, J. P. and Clark, F. R. (1966). Psychophysics and signal detection. In J. Sidowski (ed), *Experimental Methods and Instrumentation in Psychology*. New York: McGraw-Hill.

Einhorn, H. J. and Hogarth, R. M. (1981). Behavioral decision theory: processes of judgment and choice. *Annual Review of Psychology*, **32**, 53–88.

Fischhoff, B., Slovic, P. and Lichtenstein, S. (1980). Knowing what you want: measuring labile values. In T. S. Wallsten (ed), *Cognitive Processes in Choice Decision Behavior*, Hillsdale, NJ: Erlbaum Associates.

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S. and Combs, B. (1978). How safe is enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, **8**, 127–152.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Helson, H. (1964). *Adaptation Level Theory*. New York: Harper and Row.

Hodgson, R. W. and Thayer, R. L. (1980). Implied human influence reduces landscape beauty. *Landscape Planning*, **7**, 171–179.

Hull, R. B., IV, Buhyoff, B. J. and Daniel, T. C. (1984). Measurement of scenic beauty: the law of comparative judgment and scenic beauty estimation procedures. *Forest Science*, **30**, 1084–1096.

Jackson, R. H. and Hudman, L. E. (1978). Assessment of the environmental impact of high voltage power transmission lines. *Journal of Environmental Management*, **6**, 153–170.

Kellomaki, S. and Savolainen, R. (1984). The scenic value of the forest landscape as assessed in the field and the laboratory. *Landscape Planning*, **11**, 97–107.

Klukas, R. W. and Duncan, D. P. (1967). Vegetational preferences among Itasca Park visitors. *Journal of Forestry*, **65**, 18–21.

Lichtenstein, S. and Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, **89**, 46–55.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

O'Brien, R. M. (1979). The use of Pearson's R with ordinal data. *American Sociological Review*, **44**, 851–857.

Rowe, R. D., d'Arge, R. C. and Brookshire, D. S. (1980). An experiment on the economic value of visibility. *Journal of Environmental Economics and Management*, **7**, 1–19.

Russell, J. A. and Lanius, U. F. (1984). Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology*, **4**, 119–135.

Shuttleworth, S. (1980). The use of photographs as an environmental presentation medium in landscape studies. *Journal of Environmental Management*, **11**, 61–76.

Simpson, C. J., Rosenthal, T. L., Daniel, T. C. and White, G. M. (1976). Social-influence variations in evaluating managed and unmanaged forest areas. *Journal of Applied Psychology*, **61**, 759–763.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.

Turner, C. F. and Krauss, E. (1978). Fallible indicators of the subjective state of the nation. *American Psychologist*, **33**, 456–470.

Wade, G. (1982). The relationship between landscape preference and looking time: a methodological investigation. *Journal of Leisure Research*, **14**, 217–222.

White, G. M. (1975). Contextual determinants of opinion judgments: field experimental probes of judgmental relativity boundary conditions. *Journal of Personality and Social Psychology*, **32**, 1047–1054.

Wohlwill, J. F. and Kohn, I. (1973). The environment as experienced by the migrant: an adaptation level approach. *Representative Research in Social Psychology*, **4**, 135–164.

Revised Manuscript Received 5 *October* 1987