ELSEVIER

Landscape and Urban Planning 54 (2001) 149–161

LANDSCAPE AND URBAN PLANNING

www.elsevier.com/locate/landurbplan

Rating reliability and representation validity in scenic landscape assessments

James F. Palmer*, Robin E. Hoffman

SUNY College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, USA

Received 25 September 1999; received in revised form 20 March 2000; accepted 17 May 2000

Abstract

The US Supreme Court recently determined that experts from all fields of knowledge must demonstrate the reliability and validity of their testimony. While the broader implications of their finding have yet to manifest itself, it clearly has the potential to challenge all manner of professional practices. This paper explores the reliability of visual quality ratings of landscapes, and the validity of photographic representations used when making these ratings. A review of the literature finds that relatively few studies report reliability or validity coefficients. Those including such reports give reason for some concern. Data from several studies are re-analyzed to demonstrate how professionals should evaluate the reliability and validity of their visual landscape assessments.

Keywords: Visual quality; Landscape aesthetics

1. Introduction

Visual landscape assessments involve the inventory and evaluation of diverse visible attributes of the landscape for purposes of planning, design and management. Such assessments may involve expert appraisals or public judgements, but they are always conducted by landscape professionals. As currently practiced, visual assessments are firmly grounded in a tradition of knowing that requires the collection of empirical (often quantitative) data for analysis through systematic means. That is, we believe the landscape has a physical reality independent of people that can be characterized through various measurements. The landscape also has a reality that depends on our individual perceptions. These perceptions can be characterized or measured by various means. While beauty (and other perceptions) may be ''in the eye of the beholder.'' we believe that large groups of people share similar landscape perceptions, either because of biological heritage or common cultural and personal experiences. Attendant with this perspective of research or practice is the normative belief that to be good our work must be both reliable and valid.

This paper reviews two issues of particular importance to scenic landscape assessments: (1) the degree of similarity among evaluators, which is a test of reliability, and (2) the equivalence of judgements made from photographs and in the field, which is a test of validity. The rest of the introduction describes why reliability and validity are a concern for visual landscape assessment. In the following sections, the literature supporting the common practice in landscape assessments is reviewed for each of these concerns. Data are presented that cast doubt on common practices, and a more responsible approach is proposed.

* Corresponding author. Tel.: +1-315-470-6548.
E-mail address: zooey@mailbox.syr.edu (J.F. Palmer).

0169-2046/01/$20.00
PII: S0169-2046(01)00133-5

1.1. Reliability and validity as professional responsibilities

What is it that distinguishes a visual landscape assessment prepared by an expert compared to an ordinary citizen? Landscape architects form the primary professional group conducting visual landscape assessments in the United States. The American Society of Landscape Architects' Code and Guidelines for Professional Conduct requires that its members "not mislead ... clients," and "undertake to perform professional services only when they ... are qualified" (ASLA, 1998: 6). However, this says little about what standards to expect from visual landscape assessment experts. Recent rulings from the US Supreme Court are much less ambiguous about what it means to be an expert.

Since 1923 the standard under which US Courts permitted expert testimony was whether he or she was "generally accepted" by the relevant scientific community (Marshall, 1993). In Daubert (1993), the US Supreme Court asserted a new standard for determining what qualifies an expert to provide factual testimony. Before accepting testimony on facts or data, the trial judge must ensure that the "expert's testimony both rests on a reliable foundation and is relevant to the task at hand." The judge must further determine that the testimony will be relevant "by demanding a valid scientific connection to the pertinent inquiry as a precondition to admissibility." The focus in *Daubert* was on scientific findings. The court offered four possible considerations for determining the admissibility of scientific testimony: (1) whether the theory or technique is falsifiable and has been tested, (2) whether it has been subjected to peer review and published, (3) what is its known or potential error rate and what are the standards to control it, and (4) widespread acceptance by the relevant scientific community. In Kumho (1999) the Court reaffirmed their ruling and extended it "not only to 'scientific' testimony, but to all expert testimony." The facts of this case concerned whether a tire failed from a manufacturing defect or other cause. Of particular importance to us is that "the specific issue before the court was not the reasonableness *in general* of a tire expert's use of a visual and tactile inspection" to obtain data, but the "particular method of analyzing the data thereby obtained, to draw a conclusion regarding *the particular matter to which the expert testimony was directly relevant*." They further clarify that "the relevant issue was whether the expert could reliably determine the cause of *this* tire's separation." In *Kumho*, the expert "failed to satisfy either *Daubert's* factors or *any other* set of reasonable reliability criteria."

A central feature of environmental decision-making in the United States, as well as elsewhere, is its adversarial nature. These US Supreme Court rulings shed significant light on what all experts must do to establish the reliability and relevance of both their data and methods of analysis. It is the experts' responsibility to establish the reliability and validity of their methods generally, and as applied to specific studies. We believe that few involved with visual landscape assessment are prepared to do this. The following review of the literature establishes that there is cause for concern. Procedures to evaluate the reliability of landscape ratings and the validity of landscape representations are presented here in the hope that they will be widely employed by those preparing visual landscape assessments.

1.2. Visual landscape assessments in practice

We believe that few of those conducting visual landscape assessments are prepared to meet the Supreme Court's criteria for expertise. Most assessments rely on one of the established procedures (Countryside Commission, 1993; Smardon et al., 1988; USDA, 1995, 1974; USDI, 1980; USDOT, 1981). However, they are frequently adapted for local circumstances, or in other ways "improved." The assessment begins with a desk study to become oriented to the area, determine its boundaries, and prepare the fieldwork. Typically only one professional conducts the field evaluation at any particular site. This may include completing systematic evaluation sheets or simply taking field notes. Photographs are normally taken to document the area, but often without any explicit approach to selection beyond a desire to be "representative" (Hull and Revell, 1989). Upon returning to the office, the field data are organized. If the study is associated with an impact assessment, then visual simulations may be prepared. If the project is politically charged, a group of evaluators may rate slides or photographs of the site with and without the proposed project. However, evaluations are always the

result of an individual's experience, not a group's experience. The professional analyzes these data and prepares a report that includes only mean ratings.

Issues concerning the reliability and validity of the photographic representations and the evaluations are not considered. If a reviewer raises these issues, the response will be that research by Daniel and Boster (1976), Shafer and Richards (1974) and Zube and colleagues (1974), has shown that photographs can be used with confidence in visual assessments. As for the procedures used, the reply will be that they are ''widely accepted'' and the evaluations were made by a qualified professional. It is unlikely that they are prepared to demonstrate how the work of Daniel, Shafer or Zube relates to their specific project. Nor could they refer to work demonstrating the reliability of the evaluation methods they use, particularly as applied to their specific project.

1.3. Units of analysis

We found reason to be concern after conducting the literature reviews summarized below. When reporting the reliability of landscape ratings, authors almost always report the reliability of the group's mean rating and not the reliability of individual ratings. Similarly, when evaluating the validity of photographic landscape representations, authors almost always report the correlation of mean ratings for a group of representations compared to ratings of the actual settings they represent. The validity of individual representations are rarely considered. We believe this practice is in error because it fails to represent the way landscapes are experienced by the public or evaluated by landscape professionals. This error is known by social scientists as the ''ecological fallacy.''

Robinson (1950) first demonstrated the error of using a group measurement to substitute for individual measurements. He defines an analysis of individuals as being when ''the statistical object or thing described is indivisible.'' An ecological analysis occurs when the ''statistical object is a group'' of individuals. Robinson provides dramatic examples of where grouped data are erroneously used to describe individual attributes. Since landscape ratings are an individual activity in practice, then an analysis of the reliability of ratings must begin with individual rather than group data. Similarly, we normally evaluate individual sites or views. Therefore, the validity of representations must also be evaluated at the site level, and not for an aggregation of sites.

However, we acknowledge that there may be times when it is appropriate to consider the reliability of a group's mean rating or the validity of a group of representations. There is a recent exchange in the literature about the appropriate unit of analysis for environmental assessments involving human behavior or perceptions (Levine, 1994, 1996; Richards, 1990, 1996; Richards et al., 1991). While the unit of analysis may not be an open and closed decision, we believe that most landscape assessments are conducted by an individual of a specific site or sites. Evaluations of landscape assessment reliability and validity must reflect this practice. As described above, the Federal courts also expect experts to demonstrate the reliability and validity of their specific judgements, as well as the reliability and validity of their general procedures and practices.

2. Evaluating the reliability of ratings

Reliability refers to the dependability or consistency of something that is done repeatedly. Measurements, such as those made with rating scales, are reliable when they are similar, even though made by different people, or at different times. The average correlation among separate measurements is the normally accepted test of their reliability. Among psychometricians, reliability coefficients of 0.70 or 0.80 are normally expected from sound research. In applied settings, where the measurements are the basis of important decisions, reliabilities of 0.90 and above are expected (Nunnally, 1978).

Since professional or public landscape ratings may have important consequences, it is surprising that relatively little attention is paid by most researchers and practitioners to the reliability of landscape assessment methods — and there is a lot on which to focus. For instance, how many photos are needed to reliably represent different landscapes (Daniel et al., 1977; Hoffman, 1997)? When a long series of landscape scenes are being evaluated, are the same standards being reliably applied throughout the sequence (Palmer, 1998)? Are landscape evaluations stable after a year or two (Hull and Buhyoff, 1984); how about after

10 years (Palmer, 1997)? While these and other questions of reliability are all important, this paper focuses on the reliability of raters.

2.1. Calculating reliability coefficients

There are diverse ways to estimate the reliability or agreement of a rater's judgements (Ebel, 1951; Jones et al., 1983; Tinsley and Weiss, 1975). The two most common approaches to describe reliability are the average interrater correlation and the intraclass correlation. The average interrater correlation is the mean of the pairwise product moment (e.g. Pearson) correlations between all the members of a group. Estimates of the reliability for any size group are made by adjusting the average interrater correlation using the Spearman–Brown formula (Nunnally, 1978). Its advantage is simplicity of calculation. Most statistical software programs require only one instruction to calculate all of these correlations and a spreadsheet can be used to find their average. Its disadvantage is that ratings with parallel profiles have a high correlation, even though their systematic differences may be significant. The loss of between-rater variation gives a more optimistic estimate of reliability, which also may be significant.

The intraclass correlation is calculated from specific mean square components of an analysis of variance (ANOVA). There are many possible ways to calculate the intraclass correlation. Specifying the appropriate ANOVA model and properly calculating the intraclass correlation requires a level of statistical sophistication not often found among landscape assessors (Shrout and Fleiss, 1979). However, there are situations where the intraclass correlation can be calculated, and the interrater correlation cannot. In addition, the intraclass correlation can easily account for between-rater variation, which may result in lower reliability. This approach is advised by most authors (Ebel, 1951; Jones et al., 1983; Tinsley and Weiss, 1975).

2.2. Literature review

The general practice for reporting reliability of scenic landscape assessments is to focus on the correlation between the mean ratings of two or more groups, or the reliability of a group's mean rating (Palmer, 2000). Table 1 summarizes the reliability of scenic preference ratings as reported in 13 studies involving a total of more than 1000 people. A cursory review of these values gives one an overall feeling of confidence in landscape ratings.

2.3. Reliability of individual ratings

All the rating reliabilities in Table 1 are for group means, or represent the mean correlation among groups. "However, if the raters ordinarily work individually, . . . then the reliability of individual ratings is the appropriate measure" (Ebel, 1951). In most cases, landscape ratings are actually made by individuals, and evaluating only mean group ratings is an example of the "ecological fallacy."

The few studies that report the reliability of individuals are instructive. For instance, Patsfall et al. (1984) calculated an individual reliability of 0.23 for preference ratings, down from a reliability of 0.92 for the group of 41 raters. Feimer et al. (1979) called for more studies of individual reliability of landscape ratings. Their group of 22 students had an individual reliability of 0.02 for scenic beauty judgements, while 22 landscape professionals with the Bureau of Land Management (BLM) had an individual reliability of only 0.14. Results such as these leave one with much less confidence in individual ratings.

The difference between individual and group reliability coefficients is further illustrated for three studies in Table 2. In 1986, Palmer and Smardon (1989) surveyed a random sample of residents and attendees at a public workshop to study the human-use values of wetlands in Juneau, Alaska. The survey included 16 photographs representing the range of local wetland types and conditions. The second study in Table 2 began as part of a community effort to develop a comprehensive plan for Dennis, Massachusetts. In 1976, a random sample of registered voters evaluated 56 photographs representing the town (Palmer, 1983). These ratings are compared to those from employees of the US Army Corps of Engineers which were gathered in preparation for a training course in landscape aesthetics (Palmer, 1985). The third study evaluated simulations of different harvesting intensities, patterns, and patch sizes of clearcuts in the White Mountain National Forest, in New Hampshire (Palmer, 1998). Respondents included a random sample of regional residents, opinion leaders in the management

Table 1
Studies reporting the reliability of mean group scenic preference ratings[a]

Ref.	Correlation	Number of raters	Region	Settings			Reliability coefficient			Sampling method
				Forested	Open countryside	Urban/ suburban	Intrclass correlation	Pearson correlation	Other	
Brown and Daniel, 1987	0.85–0.93	20–27	SW	•			•			S
Daniel et al., 1989	0.92–0.98	38–101	SW	•			•			U
Gobster and Chenoweth, 1989	0.91–0.96	44	NE	•	•		•			S
Herzog and Bosley, 1992	0.95	104		•	•				•	S
Herzog, 1985	0.95	85		•	•				•	S
Herzog, 1987	0.97	274	SW	•	•				•	S
Herzog, 1989	0.98	76	NE			•			•	S
Hetherington et al., 1993	0.95	16	SW				•			S
Parsons and Daniel, 1988	0.91	27–36	SW				•			S
Patsfall et al., 1984	0.92	41	SE	•			•			S
Ribe, 1994	>0.90	16–40	NE	•				•		C
Rudis et al., 1988	0.85–0.92	35	SW	•			•			S
Schroeder, 1986	0.66–0.91	13–28	NE			•		•		C

[a] Notes: sampling method for respondents: available students (S), civic groups (C), random (R), or users (U). Region of study: US northeast (NE), northwest (NW), southeast (SE), southwest (SW), Australia (Au), or Europe (E).

Table 2
Reliability coefficients for scenic preference ratings from three studies[a]

Location	Respondents	*n*	Interrater	Intraclass(1)	Intraclass(*n*)
Juneau, AK	Residents	406	0.307	0.309	0.995
	Public meeting attendees	41	0.355	0.243	0.928
Dennis, MA	Registered voter	68	0.603	0.607	0.991
	Environ. professionals	118	0.672	0.633	0.995
White Mountains, NH	Residents	73	0.512	0.247	0.960
	Opinion leaders	97	0.532	0.345	0.979
	USFS employees	205	0.619	0.447	0.994

[a] Notes: the number of photographs evaluated in Juneau was 16, in Dennis it was 56, and in the White Mountains it was 64.

of the area's forests, and environmental professionals stationed on National Forests in the northeastern quarter of the US.

In several instances, the interrater and intraclass reliability coefficients for individuals give essential the same results. In those cases where the intraclass(1) correlations are significantly lower, it reflects the substantial variation among the raters within the group. In all cases, the intraclass(*n*) correlation measuring the reliability of the group's mean rating is much higher than the reliability coefficients for individuals. While these group mean ratings can be used with confidence, the use of only one or two individual's ratings should be viewed with suspicion.

3. Evaluating the validity of representations

Validity refers to the degree that something is as it purports to be. The validity of measurement scales and indices has long concerned social scientists. Almost 40 years ago, Ebel (1961) questioned whether this obsession is well placed. ''So long as what a test is suppose to measure is conceived to be an ideal quantity, unmeasurable directly and hence undefinable operationally, it is small wonder that we have trouble validating our tests'' (Ebel, 1961: 643). He suggests that we settle for reliable and useful measurement.

However, the validity of using photographs in visual landscape assessments is testable by comparing them to the settings they are intended to represent. Sheppard (1982: 14–15) identifies two basic types of validity: accuracy and realism. Accuracy refers to ''replicating the physical and visual qualities,'' such as color, position, scale, shape, and texture. All of these characteristics have a physical reality that can serve as the evaluation criterion. Verifying accuracy is particularly critical for the simulations of proposed conditions used in visual impact assessment (Sheppard, 1989).

Realism refers to an observer's ''response equivalence'' when viewing the real setting or a photographic representation. This may be a difficult determination because it relies on indirect measures of the observer's experience of the type questioned above by Ebel (1961). Even with the difficulties associated with its measurement, the response equivalence is a fundamental requirement for many landscape assessments and is the focus of this section.

3.1. Calculating validity coefficients

The first issue in testing the response equivalence between photographic representations and the actual setting they represent is how to capture meaningful responses. The normal practice is a pairwise comparison of ratings completed on site and while viewing the representation. It is assumed in this paper that rating scales are a valid way to measure visual experience, though this is a topic that warrants further investigation.

There are two approaches to calculating validity coefficients. The most common is to demonstrate a similar response pattern by calculating a correlation coefficient, normally the Pearson's product-moment correlation coefficient. However, correlation only measures a similar pattern of response. A high correlation is possible even though the actual values may systematically differ by a significant amount.

Student's *t*-test or analysis of variance is used to identify whether a significant difference exists between particular field and photograph ratings. We were not able to find any statement of expected validity standards, but believe that Nunnaly's (1978) requirement of a minimum correlation of 0.70 and a preferred correlation of 0.90 are desirable targets. There should not be a significant difference between the field and photograph ratings. However, as a practical matter, significant *t*-tests may occur simply by increasing the number of respondents. It may be appropriate to consider a measure of agreement, rather than difference (Tinsley and Weiss, 1975).

3.2. Literature review

Stamps (1990) identified over 1300 citations that appeared to use photographs to evaluate environmental preference. His purpose was to use the statistical principles of meta-analysis to calculate the combined size-effect over all the studies (Rosenthal, 1991). However, he found only 44 studies that could possibly be used in a meta-analysis, and only 11 published articles that met the meta-analysis requirements of reporting the correlation between preferences based on photographs and the actual scenes from at least 4 sites. In other words, just a few percent of the studies Stamps identified attempted to validate the use of photographic representations.

Stamps' meta-analysis has been extended here to include a total of 19 studies that explicitly tested the validity of single photographs or slides to represent the landscape for visual preference evaluation. The findings of these studies, which are briefly characterized in Table 3, are reported in more detail by Hoffman and Palmer (1994). Collectively, these 470 sites produce an overall size effect (i.e. the weighted average correlation) of 0.80, which appears to support the reasonableness of using photographic representations for scenic landscape evaluation.

However, there are problems with this approach. The first is known as the "file drawer problem," an effect with which most researchers are familiar (Iyengar and Greenhouse, 1988). Publishable research generally requires positive results, while negative results are simply stored away in the file drawer. However, there are a few cases that indicate photographic representations may not always be valid. For instance, after evaluating the validity at eight sites in their study of the Connecticut River Valley, Zube et al. (1974: 49) conclude that: "All of the analyses suggest that panoramic and wide-angle color photography *may be* valid landscape simulation media" (emphasis added). Yet two pages later they devote a full page to discussing the significant differences found at two of the eight sites between the field and photograph ratings using 18 landscape description scales and the presence of 51 landscape features.

Brown and coworkers (1988) evaluated the validity of photographic representations at 11 campgrounds. While they found there was a 0.76 correlation between the mean direct field and mean photograph ratings, these means were also significantly different from each other ($t = 6.8$, $p < 0.0001$). They conclude by stating that, "our results suggest that the campers, who had demonstrated in photo-based evaluations that they were discriminating judges of forest conditions, still generally liked where they were more than any photographically represented alternative. Which evaluation, then, is to be believed? Should society disregard the preferences of public panels that provided their reasoned discrimination among a set of options if, in the act of participation, campers report a strong preference for things as they are?"

In perhaps the most disquieting study of all, Danford and Willems (1975) used 16 semantic differential scales to evaluate the setting of the Bates College of Law at the University of Houston. A group of students were taken on a tour of the College with 62 "standard stops." Another group viewed slides of the 62 standard stops. The results indicated a high agreement between the ratings given during both experiences of the site. However, another group without any previous experience with the site were asked "to complete the response instrument based solely upon personal expectations concerning (a) what one would *expect* a law school to *look* like and (b) how one would *expect* a law school to make one feel." This group's evaluation of the site was indistinguishable from the other two groups! The authors conclude by warning that "using comparisons of subjective ratings to show convergent validity between simulations and real, intact environments is not enough." Clearly more attention needs to be given to questions of representational validity.

Table 3
Studies testing the validity of photographic media compared to field evaluations of scenic preference[a]

Ref.	Correlation	Number of sites	Region	Settings			Media				Experimental design		
				Forested	Open countryside	Urban/ suburban	Black and white	Color	Prints	Slides	Different field/ photo subjects	Same field/ photo subjects	Sampling method
Brown et al., 1988	0.76	11	SW	•				•	•			•	U
Brush, 1979	0.67	10	NE	•				•	•		•		S/U
Clamp, 1975	0.80	170	E	•	•	•		•	•			•	R
Coughlin and Goldstein, 1970	0.64	92	NE	•	•	•		•		•	•		S
Daniel and Boster, 1976	0.97	6	SW	•				•		•	•		S/C
Daniel and Boster, 1976	0.98	6	SW	•				•		•	•		S/C
Dearinger, 1979	0.79	34	SE	•	•			•		•	•		S/E
Dunn, 1974	0.64	6	E						•		•		–
Hull and Stewart, 1992	0.91	12	SE	•	•			•	•			•	U
Kane, 1981	0.96	10	Au	•	•	•		•		•	•		–
Kellomaki and Savolainen, 1984	0.83	34	E	•			•		•		•	•	S/R
Kroh and Gimblett, 1992	0.38	16	NE	•	•			•		•		•	S
Rabinowitz and Coughlin, 1971	0.89	14	NE	•	•	•		•		•	•	•	R
Seaton and Collins, 1972	0.93	4	NW			•		•	•		•		U
Seaton and Collins, 1972	−0.27	4	NW			•	•		•		•		U
Shafer and Richards, 1974	0.71	8	NE	•	•	•		•	•	•	•		S
Shelby and Harris, 1985	0.76	20	NW	•				•	•		•		U
Stewart et al., 1984	0.67	5	SW		•	•		•		•		•	U
Zube et al., 1974	0.76	8	NE		•	•		•	•			•	S/R

[a] Notes: sampling method for respondents: available students (S), experts (E), civic groups (C), random (R), or users (U). Region of study: US northeast (NE), northwest (NW), southeast (SE), southwest (SW), Australia (Au), or Europe (E).

3.3. Validity of individual photographic representations

All of the studies in Table 3 validating photographic representations are for groups of scenes. Therefore, these studies all fall prey to the ''ecological fallacy'' by failing to recognize that each photo represents a specific view at a specific site, not the landscape generally. The remainder of this section will present results that illustrate this issue.

The materials and data to investigate photo validity are drawn from three studies summarised in Table 4 that were conducted in northeastern hardwood forests to better understand major forest management issues. The visual affect of harvesting alternatives were simulated in this study from two vista sites in the White Mountain National Forest: Sugarloaf and Welsh Ledge (Palmer et al., 1995). In all there were 30 simulations representing different sized clearcuts, patterns and intensities. Also included among these simulations were the photograph of the original view taken the previous year and a completely revegetated simulation. A total of 26 different hikers at each site, used a 10-point bi-polar scale to rate the scenic value of all 32 scenes. The interviews were conducted in the field at the view point. The hikers also rated the actual view and were asked to locate the scene that most resembled the view from among the 32 photographs.

Cuyler Hill State Forest in Central New York is a demonstration site for uneven-age single-tree selection silviculture (Hoffman and Palmer, 1994). The area was marked according to research guidelines and harvested by a commercial logger in late-August 1993. The last previous harvest was in mid-1970. Two existing data collection points with similar appearance were selected in the treatment area and a third point from a neighboring unharvested stand with similar forest structure was selected as a control site. All three sites were photographed in mid-July 1993. The two treatment sites were rephotographed a week after the harvest. The slash was lopped to the ground at one of these sites, and rephotographed. This resulted in the sites representing three different field conditions: uncut, slash, and lopped. Two non-overlapping views were photographed using color 35 mm film at all three sites in their uncut condition, at two sites in their slash condition, and one with the slash lopped. All three sites were visited by 12 landscape architecture students during September 1993. A total of 24 bi-polar scales were used to describe various visible forest characteristics, including overall scenic

Table 4
The validity of photographs to represent the scenic preference of views in three studies[a]

White Mountains, NF: view point[b]	Mean rating		Statistics	
	Field	Photo	t	r
Sugarloaf	6.23	3.54	4.53^{***}	0.26
Welsh Ledge	7.08	5.71	2.70^{*}	0.24
Cuyler Hill SF: treatment[c]				
Uncut	3.25	3.17	0.24	0.25
Cut w/slash	3.83	4.88	−1.71	0.05
Cut and lopped	3.67	3.21	2.03	0.81
Allegany NF: density, treatment[c]				
10 deer/m^2, uncut	2.13	1.96	1.16	0.78
10 deer/m^2, thinned	2.72	2.56	0.46	0.10
10 deer/m^2, clearcut	3.96	3.58	0.89	0.20
64 deer/m^2, uncut	2.65	2.00	3.16^{**}	0.50
64 deer/m^2, thinned	2.66	2.11	1.97	0.36
64 deer/m^2, clearcut	4.59	4.59	0.00	0.38
Uncontrolled, Sfailed clearcut	2.95	3.05	−0.37	0.79

[a] Significance levels are $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.
[b] Rating scale of 1: low scenic value to 10: high scenic value.
[c] Rating scale of 1: beautiful to 7: ugly.

beauty. All participants rated the 3 field sites and 12 photographic slides.

Regeneration failure after harvesting in the Allegheny National Forest of western Pennsylvania has become a serious problem due to excessive browsing by the deer population (Hoffman and Palmer, 1994). Fenced plots at Fool's Creek have been established to study the effects of deer density on forest vegetation. There were two controlled densities of 10 and 64 deer per square mile. Within each density, there is an area that has been uncut, thinned to 60% stocking, and clearcut in 1968. A third unfenced site was clearcut and remained open to the natural density of >100 deer per square mile. During the second week of September 1994, a color 35 mm slide was taken at each site to represent each management condition, for a total of seven slides. On 16 October 1994, a group of 27 students and their professors on a biology field trip evaluated all the slides and sites. They used 24 bi-polar scales to describe the forest's visible characteristics, including an overall rating of scenic beauty.

The results from these three studies give reason for concern about the uncritical use of photographic representations, at least for studies of the scenic quality in the northeastern hardwood forest. There are significant differences and very low correlations between the field and photograph ratings in the White Mountain National Forest study. The respondents were interviewed at the sites where the photos were taken, a viewpoint with a breathtaking panoramic view. These hikers understood the scene they were looking at. Over 80% at each site recognized the clearcut patches in the on-site view, and over one third were able to pick the photograph of the site's actual condition from among the 32 very similar photographs. However, their comments made it clear that they were not able to objectively ignore the dramatic context of the panorama when they were directed to evaluate that portion of the on-site view represented in the photographs. In this case, the simulations were accepted as useful representations for the limited purpose of comparing generic alternative harvesting scenarios. However, they would not be appropriate for assessing the visual impacts of harvesting alternatives from these particular sites.

At the Cuyler Hill State Forest, the mean rating of the two slides was compared to the scenic quality rating completed while in the forest stand. There are no significant differences between the field and photo ratings. However, two of the three sites have unacceptably low correlations between the field and photo ratings.

The results from the Allegany National Forest deer study show the scenic ratings for one of the seven photos is significantly different from the corresponding field ratings. While the correlations between the field and photo scenic ratings are acceptable at two sites, and marginal at a third, they are unacceptable at the remaining four sites. Once again, the results suggest caution in the use of photographic representations.

3.4. Single representations

Another potential problem with using photographic representations is that they record a limited field of view. While an observer may experience the visual condition within an angle of 120° or more through slight movements of the eyes or head, the standard 35 mm wide angle lens covers approximately 60°. All of the results reported in Table 3 are for single frame photographic representations. It is very unusual to find a study that uses more than one photograph to represent the visual conditions from a particular viewpoint.

The results in Table 5 compare scenic beauty ratings by 12 landscape architecture students of two photographs taken at each of six conditions in Cuyler Hill State Forest. All the images are taken within what foresters consider homogenous conditions, leading one to expect no visual differences between them.

Table 5
A comparison of scenic preference ratings for two photographs taken from the same viewpoint at Cuyler Hill[a]

Site	Treatment	Mean rating		Statistics	
		Slide A	Slide B	t	r
X	Uncut	2.90	2.60	1.10	0.69
Y	Uncut	2.35	3.30	0.21	0.54
Y	Cut with slash	4.45	3.25	2.76*	0.52
Z	Uncut	2.95	2.65	1.28	0.41
Z	Lopped slash	3.05	3.50	−0.78	0.67
Z	Cut with slash	4.25	2.30	2.30*	0.21
Overall		3.33	2.93	3.11**	0.55

[a] Significance levels are *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

The correlations among the six matched pairs range from poor to mediocre. The ratings are significantly different for the two viewpoints with significant amounts of slash and tree tops left from the recent single tree selection harvest, as well as for the group as a whole. The mean rating for the two photographs at each site is used in the field-photo validity test.

4. Conclusions

Over the past 30 years, researchers have developed confidence in the reliability of rating scales to evaluate landscape qualities, and the validity of photographs to represent those qualities. While most studies have been limited to the assessment of scenic preferences, this confidence has been extended to all visible landscape qualities. It seems that there is cause for concern if those professionals conducting landscape assessments embrace this confidence.

The wholesale extension of high reliability for mean group scenic preference ratings to the judgements of one or perhaps a few professional landscape assessors is demonstrated in this paper to be unfounded. This is particularly true for visible landscape qualities other than scenic preference, which have not been as extensively studied.

Similarly, past demonstrations of the validity of photographic representations have been averaged over a group of views, rather than for individual views. This paper calls attention to several instances where specific photographs poorly represented the scenic quality of the actual view. In some instances, this may be because no single photograph can represent the diversity readily seen from a particular viewpoint. Further, the extension of results based on scenic preference to other visible qualities is also unfounded.

We would offer some recommendations to professionals preparing landscape assessments, in the light of the US Supreme Court's recent findings about the need of an expert to demonstrate the reliability and validity of their methods and data both generally and specifically.

1. *Establish the reliability of professional ratings*. This may be accomplished by having several professionals evaluate each view and then calculating the reliability coefficient. It is also desirable to establish the relationship of professional ratings to a criterion group, such as a random sample of the public.
2. *Establish the validity of each landscape representation*. This may be accomplished by comparing the ratings of the representations and actual field conditions from several individuals. In situations where the visual condition may be quite diverse, use panoramic images or more than one photograph from each viewpoint.
3. *Establish a record of preparing valid visual simulations*. It is, of course, not possible to establish the validity of a simulation before it is built, but it is possible to validate the existing condition's representation. Professionals who frequently create visual simulations should establish a record of their validity by comparing ratings of them to ratings of the project as built.

If professionals conducting landscape assessments follow this advice, then they should be able to meet any challenges to the reliability or validity of their work.

Acknowledgements

Portions of this research were funded by the USDA Forest Service's North Central Forest Research Station, Chicago, IL, and the New York Center for Forestry Research and Development, Syracuse, NY. The data from Fool's Creek were collected by Susan Stout, USDA Forest Service, Northeastern Research Station, Warren, PA.

References

American Society of Landscape Architects, 1998. The 1999 Members Handbook. Washington, DC.

Brown, T.C., Daniel, T.C., 1987. Context effects in perceived environmental quality assessment: scene selection and landscape ratings. J. Environ. Psychol. 7 (3), 233–250.

Brown, T.C., Richards, M.T., Daniel, T.C., King, D.A., 1988. Recreation participation and the validity of photo-based preference judgments. J. Leisure Res. 20 (4), 40–60.

Brush, R.O., 1979. The attractiveness of woodlands: perceptions of forest landowners in Massachusetts. For. Sci. 25 (3), 495–506.

Clamp, P., 1975. A study in the evaluation of landscape and the impact of roads. Landscape Res. News 1 (11), 6–7.

Coughlin, R.E., Goldstein, K.A., 1970. The extent of agreement among observers on environmental attractiveness. Regional Science Research Institute, Discussion Paper Series: No. 37. Philadelphia, PA.

Countryside Commission, 1993. Landscape Assessment Guidance. CCP3 423. Cheltenham, UK.

Danford, S., Willems, E.P., 1975. Subjective responses to architectural displays: a question of validity. Environ. Behavior 7 (4), 486–516.

Daniel, T.C., Boster, R.S., 1976. Measuring landscape esthetics: the scenic beauty estimation method. Research Paper RM-167. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station. Fort Collins, CO, 66 pp.

Daniel, T.C., Anderson, L.M., Schroeder, H.W., Wheeler III, L., 1977. Mapping the scenic beauty of forest landscapes. Leisure Sci. 1 (1), 335–352.

Daniel, T.C., Brown, T.C., King, D.A., Richards, M.T., Stewart, W.P., 1989. Perceived scenic beauty and contingent valuation of forest campgrounds. For. Sci. 35 (1), 76–90.

Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993. 509 US Supreme Court 579.

Dearinger, J.A., 1979. Measuring preferences for natural landscapes. In: Proceedings of the American Society of Civil Engineers. Journal of the Urban Planning and Development Division. Vol. 150(UP1), pp. 63–80.

Dunn, M.C., 1974. Landscape evaluation: a further perspective. J. R. Town Plann. Inst. 60 (10), 935–936.

Ebel, R.L., 1951. Estimation of the reliability of ratings. Psychometrika 16 (4), 407–424.

Ebel, R.L., 1961. Must all tests be valid? Am. Psychol. 16, 640–647.

Feimer, N.R., Craik, K.H., Smardon, R.C., Sheppard, S.R.J., 1979. Appraising the reliability of visual impact assessment methods. In: Elsner, Smardon (Technical Coordinators), Our National Landscape. General Technical Report. PSW-35. USDA Forest Service, Pacific Southwest Forest and Range Esp. Stn., Elsner, G.H. Smardon, R.C. (Technical Editors) Berkeley, CA, pp. 286–295.

Gobster, P.H., Chenoweth, R.E., 1989. The dimensions of aesthetic preference: a quantitative analysis. J. Environ. Manage. 29 (1), 47–72.

Herzog, T.R., 1985. A cognitive analysis of preference for waterscapes. J. Environ. Psychol. 5 (3), 225–241.

Herzog, T.R., 1987. A cognitive analysis of preference for natural environments: mountains, canyons, and deserts. Landscape J. 6 (2), 140–152.

Herzog, T.R., 1989. A cognitive analysis of preference for urban nature. J. Environ. Psychol. 9 (1), 27–43.

Herzog, T.R., Bosley, P.J., 1992. Tranquility and preference as affective qualities of natural environments. J. Environ. Psychol. 12 (2), 115–127.

Hetherington, J., Daniel, T.C., Brown, T.C., 1993. Is motion more important than it sounds?: the medium of presentation in environmental perception research. J. Environ. Psychol. 13 (4), 283–291.

Hoffman, R.E., 1997. Testing the validity and reliability of slides as representations of northern hardwood forest conditions. Doctoral dissertation. SUNY College of Environmental Science and Forestry, Syracuse, NY, 296 pp.

Hoffman, R.E., Palmer, J.F., 1994. Validity of using photographs to represent visible qualities of forest environments. In: Clark, J.D. (Ed.), Proceedings of the Council of Educators in Landscape Architecture 94 Conference on History and Culture. Landscape Architecture Foundation/Council of Educators in Landscape Architecture. Washington, DC, pp. 160–169.

Hull, R.B., Buhyoff, G.J., 1984. Individual and group reliability of landscape assessments. Landscape Plann. 11 (1), 67–71.

Hull, R.B., Revell, G.R.B., 1989. Issues in sampling landscape for visual quality assessments. Landscape Urban Plann. 174, 323–330.

Hull, R.B., Stewart, W.P., 1992. Validity of photo-based scenic beauty judgments. J. Environ. Psychol. 122, 101–114.

Iyengar, S., Greenhouse, J.B., 1988. Selection models and the file drawer problem. Statist. Sci. 3 (1), 109–135.

Jones, A.P., Johnson, L.A., Butler, M.C., Mai, D.S., 1983. Apples and oranges: an empirical comparison of commonly used indices of interrater agreement. Acad. Manage. J. 26 (3), 507–519.

Kane, P.S., 1981. Assessing landscape attractiveness: a comparative test of two new methods. Appl. Geogr. 1 (2), 77–96.

Kellomaki, S., Savolainen, R., 1984. The scenic value of the forest landscape as assessed in the field and the laboratory. Landscape Plann. 11 (2), 97–108.

Kroh, D.P., Gimblett, R.H., 1992. Comparing live experience with pictures in articulating landscape preference. Landscape Res. 17 (2), 58–69.

Kumho Tire Co., Ltd., et al., v. Carmichael et al. (1999) 526 US Supreme Court 137.

Levine, D.W., 1994. True scores, error, reliability, and unit of analysis in environment and behavior research. Environ. Behavior 26 (2), 261–293.

Levine, D.W., 1996. Why choose one level of analysis? And other issues in multilevel research. Environ. Behavior 28 (2), 237–255.

Marshall, E., 1993. Supreme Court to weigh science. Science 259, 588–590.

Nunnally, J.C., 1978. Psychometric Theory. Second Edition. McGraw-Hill, New York, 701 pp.

Palmer, J.F., 1983. Assessment of coastal wetlands in Dennis, Massachusetts. In: Smardon, R.C. (Ed.), The Future of Wetlands: Assessing Visual–Cultural Values of Wetlands. Allanheld, Osmun Co., Montclair, NJ.

Palmer, J.F., 1985. The perception of landscape visual quality by environmental professionals and local citizens. Faculty of Landscape Architecture, SUNY ESF. Syracuse, NY.

Palmer, J.F., 1997. Stability of landscape perceptions in the face of landscape change, landscape. Landscape Urban Plann. 37 (1/2), 109–113.

Palmer, J.F., 1998. Clearcutting in the White Mountains: Perceptions of Citizens, Opinion Leaders and US Forest Service Employees. New York Center for Forestry Research and Development, SUNY ESF. [NYCFRD 98–01] Syracuse, NY.

Palmer, J.F., 2000. Reliability of rating visible landscape qualities. Landscape Journal, 19 (1/2), 166–178.

Palmer, J.F., Shannon, S., Harrilchak, M.A., Gobster, P., Kokx, T., 1995. Esthetics of clearcutting alternatives in the White Mountain National Forest. J. For. 93 (5), 37–42.

Palmer, J.F., Smardon, R.C., 1989. Measuring human values associated with wetlands. In: Kriesberg, L., Northrop, T.A., Thorson, S.J. (Eds.), Intractable Conflicts and their Transformation. Syracuse University Press, Syracuse, NY.

Parsons, R., Daniel, T.C., 1988. Assessing visibility impairment in class I parks and wilderness areas: a comparison of policy-relevant methods. Soc. Nat. Res. 1 (3), 227–240.

Patsfall, M.R., Feimer, N.R., Buhyoff, G.J., Wellman, J.D., 1984. The prediction of scenic beauty from landscape content and composition. J. Environ. Psychol. 4 (1), 7–26.

Rabinowitz, C. B., Coughlin, R.E., 1971. Some experiments in quantitative measurement of landscape quality. Regional Science Research Institute, Discussion Paper Series: No. 43. Philadelphia, PA.

Ribe, R.G., 1994. Scenic beauty perceptions along the ROS. J. Environ. Manage. 42 (3), 199–221.

Richards Jr., J.M., 1990. Units of analysis and the individual differences fallacy in environmental assessment. Environ. Behavior 22 (3), 307–319.

Richards Jr., J.M., 1996. Units of analysis, measurement theory, and environmental assessment: a response and clarification. Environ. Behavior 28 (3), 220–236.

Richards Jr., J.M., Gottfredson, D.C., Gottfredson, G.D., 1991. Units of analysis and the psychometrics of environmental assessment scales. Environ. Behavior 23 (4), 423–437.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15 (3), 351–357.

Rosenthal, R., 1991. Meta-Analytic Procedures for Social Research. Applied Social Research Methods Series, Vol. 6. Revised Edition. Sage, Newburry Park, CA.

Rudis, V.A., Gramann, J.H., Ruddell, E.J., Westphal, J.M., 1988. Forest inventory and management-based visual preference models of southern pine stands. For. Sci. 34 (4), 846–863.

Schroeder, H.W., 1986. Estimating park tree densities to maximize landscape aesthetics. J. Environ. Manage. 23 (4), 325–333.

Seaton, R.W., Collins, J.B., 1972. Validity and reliability of ratings of simulated buildings. In: Mitchell, W.S. (Ed.), Environmental Design: Research and Practice. University of California Press, Los Angeles, CA, pp. 6.101–6.10.12.

Shafer, Jr., E.L., Richards, T.A., 1974. A comparison of viewer reactions to outdoor scenes and photographs of those scenes, USDA Forest Service Research Paper. NE-302.

Shelby, B., Harris, R., 1985. Comparing methods for determining visitor evaluations of ecological impacts: site visits, photographs, and written descriptions. J. Leisure Res. 17 (1), 57–67.

Sheppard, S.R.J., 1982. Landscape portrayals: their use, accuracy and validity in simulating proposed landscape changes. (Doctoral dissertation, University of California, Berkeley, 1982) UMI International, Ann Arbor, Mich.

Sheppard, S.R.J., 1989. Visual Simulation: A User's Guide for Architects, Engineers, and Planners. Van Nostrand Reinhold, New York.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428.

Smardon, R.C., Palmer, J.F., Knopf, A., Grinde, K., Henderson, J.E., Peyman–Dove, L.D., 1988. Visual Resources Assessment Procedure for US Army Corps of Engineers. (Instruction Report EL–88–1) US Army Engineer Waterways Experiment Station. Vicksburg, Mississippi, 71 pp. plus appendices.

Stamps, A.E., 1990. Use of photographs to simulate environments: a *meta*-analysis. Perceptual Motor Skills 713, 907–913.

Stewart, T.R., Middleton, P., Downton, M., Ely, D., 1984. Judgements of photographs versus field observations in studies of perception and judgement of the visual environment. J. Environ. Psychol. 4 (4), 283–302.

Tinsley, H.E.A., Weiss, D.J., 1975. Interrater reliability and agreement of subjective judgements. J. Counsel. Psychol. 22 (3), 358–376.

US Department of Agriculture, Forest Service, 1974. National Forest Management. Vol. 2, Chapter 1: The Visual Management System. (Agriculture Handbook No. 462) US Government Printing Office, Washington, DC.

US Department of Agriculture, Forest Service, 1995. Landscape Aesthetics: A Handbook for Scenery Management. (Agriculture Handbook No. 701) USDA Forest Service, Washington, DC.

US Department of Interior, Bureau of Land Management, 1980. Visual Resource Management Program. US Government Printing Office, Washington, DC, 39 pp.

US Department of Transportation, 1981. Visual Impact Assessment for Highway Projects. Federal Highway Administration, Washington, DC.

Zube, E.H., Pitt, D.G., Anderson, T.W., 1974. Perception and Measurement of Scenic Resources in the Southern Connecticut River Valley. (Pub. No. R-74-1) Institute for Man and His Environment, University of Massachusetts, Amherst, MA, 191 pp.