Three Ways to Assess Reliability in Professional Visual Impact Assessment

James F. Palmer[1]

[1]T. J. Boyle Associates, Burlington, VT/USA · palmer.jf@gmail.com

Abstract: The public expects that the services provided by professionals, such as physicians or accountants are reliable. As the public becomes more concerned about visual impacts, it is to be expected that questions will be raised about the reliability of visual impact assessment methods. This paper presents a case study investigating three types of reliability: rater reliability, test-retest reliability, and firm reliability. Reliability is generally found to be good but may not reach the highest professional standard. The comparison of two firms suggests there may be a subtle client bias.

Keywords: Visual impact ratings, rater reliability, test-retest reliability, firm reliability

1 Introduction

Environmental impact assessments are widely required worldwide as a condition for permitting projects (SADLER 1969). Visual impacts are among the public's top concerns when large development projects are proposed that have the potential to change the landscape and its appearance significantly. Examples include wind energy development (BISHOP 2011), high voltage transmission lines (IETPP 1996), and forest management (RIBE 1989). While there are decades of research about the public's perception, it may be surprising that there has been little or no investigation of the reliability of the professional judgements made in visual impact assessments (VIAs).

"Reliability" is being used in the scientific sense of whether a method produces consistent results. For instance, if a group of professionals all apply the same method to evaluate the potential visual impacts at a key observation point (KOP), do they all arrive at the same conclusion? If they are all exactly the same, then it is really only necessary that one professional conduct the evaluation. But if there is variation among their evaluations, then it is necessary to average the findings across several professionals to obtain a reliable evaluation. A related concern is whether the evaluation of a professional is stable over time. For instance, if the evaluations are conducted six months or a year apart will the results be the same? Finally, a third concern is whether professionals representing clients with different interests arrive at the same conclusion or does their evaluation tend to tilt toward their client's interests (BAZERMAN, LOWENSTEIN & MOORE 2002).

This paper investigates each of these three ways of looking at reliability of professional visual impact evaluations: (1) the reliability within a group of professionals on the same team making their assessments at the same time, (2) the test-retest reliability of the same professionals making judgements separated by a substantial time interval, and (3) the comparison of results from two groups of professionals using similar methods but representing different clients.

The judgements made in VIAs can have real-world consequences. Therefore the standards for establishing an acceptable reliability coefficient should be higher than for basic research. PALMER & HOFFMAN (2001) recognize that while reliabilities of 0.7 are fair and 0.8 are very good for basic research, reliabilities of 0.9 should be expected from professional assessments.

Journal of Digital Landscape Architecture, 8-2023, pp. 236-244. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-740-3, ISSN 2367-4253, e-ISSN 2511-624X, doi:10.14627/537740025.

2 Methods

This investigation uses data from the VIAs prepared for the approximately 187-mile (306 km) Northern Pass Transmission Project, which would deliver 1,200 MW of hydropower electricity from the Canadian border south through the state of New Hampshire to the Boston, Massachusetts metropolitan area. For most of its length, the aboveground portion is collocated in the right-of-way of an existing 115 kV transmission line. The analysis primarily relies on data from the VIA prepared for the US Department of Energy by T. J. Boyle Associates (2017) as a requirement to obtain a federal permit (hereafter referred to as DOE). A second VIA was prepared by T. J. Dewan & Associates (2015) for Eversource Energy for submission to the Site Evaluation Committee as part of the New Hampshire permitting process (here after referred to as SEC).

Both VIAs used a formalistic analysis of simulated views from a number of KOPs selected to represent the range of conditions encountered along the proposed project route. The procedure used in the DOE VIA was based on SHEPPARD & NEWMAN (1979), which rated the project's color (0-9), form (0-6), line (0-3), texture (0-3) and scale (0-6) contrast with the surrounding landscape as well as its scale (0-12) and spatial (0-6) dominance. The sum of these values determines the visual impact as severe (36-45), strong (27-35), moderate (18-26), weak (9-17) or negligible (0-8). The SEC VIA adapted a rating system from SMARDON & HUNTER (1983) that was based in part on SHEPPARD & NEWMAN (1979). However, it applied a slightly different approach to weighting the components. For this analysis, the SEC ratings were adjusted as shown in Table 1 to be equivalent to those used for the DOE VIA.

Table 1: DOE and SEC factors and possible ratings with SEC adjustment

DOE Factors	Possible Rating	SEC Factors	Possible Rating	Adjustment to SEC Ratings
Color contrast	0-9	Color	0-3	x 3
Form contrast	0-6	Form	0-3	x 2
Line contrast	0-3	Line	0-3	x 1
Texture contrast	0-3	Texture	0-3	x 1
Scale contrast	0-6	Scale	0-12	÷ 2
Spatial dominance	0-6	Horizontal Field of View	0-3	Sum
		Interfere with Existing View	0-3	
Scale dominance	0-12	Perceived Dominance	0-3	Sum x 2
		Distance Zone	0-3	

In December 2014 six landscape architects who were involved in the NPTP's field inventory for DOE were trained to conduct the VIA ratings. They evaluated the no change (alternative 1) and proposed project (alternative 2) photorealistic simulations from 15 KOPs for the DOE VIA. The simulations were 11"x17" high-resolution color prints with only a minimum of text to identify their location. They were considered in a randomly assigned order; each evaluator's judgements were made independently without any discussion.

After submission of the VIA report, Eversource Energy proposed a new preferred route (alternative 7) that buried a substantial portion of the route around a scenic National Forest.

Seven new KOPs were added to provide better representation of project impacts. This resulted in 22 KOPs for alternatives 1 and 2, but alternative 7 had only 14 KOPs with above ground views of the project. The team evaluated all three alternatives in November 2016.

The Pearson *r* is used to measure the inter-rater correlation among the 6 evaluators. Fisher's z transformation is used to calculate the mean Pearson inter-rater correlation (COREY, DUNLAP & BURKE 1998). In addition, the intraclass correlation coefficient (ICC) is calculated (PALMER & HOFFMAN 2001). The Type 2 ICC is used because all six evaluators evaluated all of the simulations. It incorporates the variation among both raters and KOPs, and reflects the absolute agreement among raters. ICC(2,1) is the expected reliability for one evaluator; ICC(2,k) is the reliability for the group of six evaluators. In addition, these data also provide an opportunity for a test-retest reliability using the Pearson *r* to compare the ratings from 2014 and 2016 for alternatives 1 and 2 at 15 KOPs.

For the SEC VIA three landscape architects evaluated alternative 7. The SEC VIA included six KOPs that were at the same location as KOPs used for the DOE VIA. This permitted a comparison between the visual impact ratings of the two firms.

3 Results

3.1 Rater Reliability

The ICC and inter-rater (i. e., mean Pearson correlation) reliabilities for the six landscape architects evaluating the three alternatives in 2016 are given in Table 2. The ANOVA analyses to compute the ICC values are all significant at the .001 level.

Table 3 presents the Pearson correlations between the six individual evaluators. Since alternatives share the same base photo, the alternatives are not independent from each other. Therefore, their results are presented separately. In general, the Pearson correlations are significant at the 0.001 level.

Table 2: Mean Pearson and ICC correlations for Alternatives 1, 2 and 7

Alternative	N sites	x̄ Pearson r	ICC(2,1)	ICC(2,k)
1	22	0.835	0.665	0.924
2	22	0.818	0.570	0.888
7	14	0.732	0.585	0.894

Table 3: Pearson correlations between raters for Alternatives 1, 2 and 7

Alternative 1	Raters				
Raters	G	H	I	J	K
H	0.848***				
I	0.839***	0.759***			
J	0.771***	0.725***	0.928***		
K	0.766***	0.693***	0.867***	0.896***	
L	0.788***	0.718***	0.941***	0.855***	0.879***

Alternative 2	**Raters**				
Raters	G	H	I	J	K
H	0.768***				
I	0.813***	0.887***			
J	0.693***	0.844***	0.812***		
K	0.802***	0.839***	0.872***	0.801***	
L	0.658***	0.905***	0.834***	0.758***	0.794***
Alternative 7	**Raters**				
H	0.648**				
I	0.840***	0.665**			
J	0.792***	0.844***	0.860***		
K	0.615*	0.275$^{n.s.}$	0.716**	0.455$^{n.s.}$	
L	0.790***	0.582*	0.932***	0.807***	0.597*

Significance: n.s. > .05, * ≤ .05, ** ≤.01, *** ≤ .001

3.2 Test-Retest Reliability

There were 15 KOPs with views of aboveground structures that were rated in both 2014 and 2016. Pearson correlation is used to determine the test-retest reliability for the six evaluators and the mean correlation for the group. The test-retest Pearson correlations in Table 4 are generally significant at the 0.001 level. The mean correlation for the group is calculated using Fisher's z transformation. These results indicate that even after a year the ratings are very consistent for all raters.

Table 4: Test-retest reliability for Alts. 1 and 2

Rater	**Alternative 1**	**Alternative 2**
G	0.884***	0.889***
H	0.719**	0.882***
I	0.949***	0.921***
J	0.857***	0.956***
K	0.850***	0.807***
L	0.828***	0.916***
Group mean	*0.848*	*0.868*

Sig.: n.s. > .05, * ≤ .05, ** ≤.01, *** ≤ .001

3.3 Firm Reliability

The mean visual impact ratings for the six KOPs that are common between the DOE and SEC VIAs are shown in Table 5. The Pearson correlation between the mean DOE and SEC VIA ratings for the six common KOPs was 0.842 (p = 0.158). While this correlation is high, the very small sample size means that the p-value is higher than is normally acceptable. In addition, the Pearson correlation measures consistency rather than absolute agreement, which may not be the most useful way to compare two firms.

Table 5: Mean visual impact for six KOPs rated by two firms

KOP	**DOE**	**SEC**
CO-4	30.17	23.33
DE-2	23.00	22.83
LA-2	17.50	12.33
NH-3	40.83	33.00
SE-3	20.67	25.33
ST-3	20.50	17.33

These ratings are plotted in Fig. 1. The general trend of the lines rises from left to right, which is why the correlation is high. However, it also appears that There is also a trend for the SEC values to be lower than the DOE values. This suggests the possibility of a client bias.

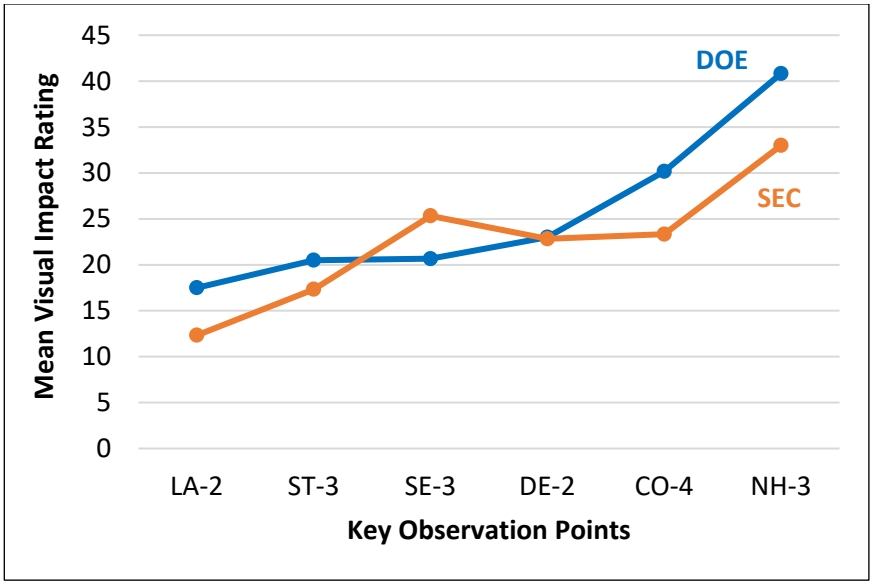

Fig. 1: Plot of the mean visual impact for six KOPs rated in common

The procedure used for the DOE VIA interprets the ratings into five levels of severity: Severe (36-45), Strong (24-35), Moderate (18-26), Weak (9-17), Negligible (0-8). Another way to compare the firms is to look at the probability that the individuals from each firm judged the visual impact to be severe or strong (i. e., a greater impact), or at lower level. Fisher's exact test is used to determine that the evaluators for the DOE VIA were significantly more likely to rate the visual impact as greater than the evaluators for the SEC VIA ($p = 0.02$).

Table 6: Greater and lower visual impacts determined by raters in the DOE and SEC VIAs

	Count			Percentage of a VIA's Count		
VIA	**Greater***	**Lower**	**Total**	**Greater ***	**Lower**	**Total**
DOE	21	15	36	58.3	41.7	100
SEC	4	14	18	22.2	77.8	100
Total	25	29	54	46.3	53.7	100

* Greater ratings are Severe or Strong. Lower ratings are Moderate, Weak or Negligible.

4 Discussion

The difficulty with determining the reliability of professionally conducted VIAs is that normal practice is to have only one evaluator; highly contested projects may use three or four raters, as was done for the VIA submitted to the SEC. The DOE VIA used six independent

raters, which is very unusual – this author is not aware of another VIA using this many professionals. While more raters would lead to a more robust VIA analysis, six evaluators and 22 KOPs is adequate to obtain the statistically significant ICC, inter-rater and test-retest results reported here. The group's results are high, sometimes meeting the 0.90-standard for professional services, but often falling just short of it. Correlations at this level are statistically significant with only six raters.

The group's rater reliability for the contrast ratings is higher than previously reports (PALMER 2000). One reason for this may be that the VIA professionals are both experienced in using contrast ratings and have field experience throughout the project area. Nonetheless, it is worth noting that the single-rater reliabilities (i. e., ICC(2,1)) were not up to professional standards, which supports BLM's (1986) direction to use multiple raters.

The question of how many raters are necessary to obtain the group reliability expected of professionals is difficult to answer. In practice, evaluating a large number of representative KOPs is probably more important than a large number of trained raters. SMARDON evaluated the reliability of the more common ratings used in VIAs and recommended a team of ten raters (SMARDON et al. 1983, 93). Writing in a medical journal, KOO & LI (2016, 157) suggest "as a rule of thumb, researchers should try to obtain at least 30 heterogeneous samples [e. g., KOPs] and involve at least 3 raters whenever possible." BUJANG & BAHARUM (2017) review how to determine the minimum number of raters for a given number of observations (KOPs) under different assumptions and provide tables to guide that determination. The number of raters is generally determined based on the assumption that there is no agreement among them (i. e., reliability is 0). If the rating for 20 KOPs indicate that the ICC is 0.8 or 0.9, then three raters are sufficient for statistical significe (alpha = .05) under that assumption. This changes if it is assumed that the typical reliability is 0.7 for a VIA and a firm wants to demonstrate that their calculated reliability of 0.9 is statistically significant; then 9 to 12 raters are required.

The test-retest reliability presented here is a found opportunity made possible because the developer changed the project design, which is not an unusual event. It would be a benefit to the profession if the practice reported here is followed by others, and an evaluation of the original simulations is repeated as well as evaluating the new design. The results indicate high reliability in applying the contrast ratings on the same scenes, even after a year has passed, though sometimes falling short of the 0.90-standard for professional services.

Interpreting the comparison of firms is more difficult. There is a strong correlation between how the firms evaluated six sites, meaning that their ordering of sites for impact severity was very similar. However, there is also a tendency of the SEC ratings to be assessed as less severe than the DOE ratings. This could be because of client bias, but it might also be a result of converting the SEC ratings to be equivalent to the DOE rating scale. Client bias is potentially a significant problem, since the developer is normally responsible for preparing the technical reports supporting an agency's environmental impact assessment – the DOE report was an exception. This is an area that deserves further research.

BAZERMAN et al. (2002, 3-4) investigated client bias among accountants doing audits. Like VIAs, accounting audits may have the appearance of deterministic objectivity, but actually require a substantial amount of professional judgement and interpretation resulting in unintended distortions. They identify three opportunities for bias that also apply to VIAs.

- **Ambiguity**. Bias thrives wherever there is the possibility of interpreting information in different ways.
- **Attachment**. Auditors have strong business reasons to remain in clients' good graces and are thus highly motivated to approve their clients' accounts. ... it is well known that client companies fire accounting firms that deliver unfavorable audits.
- **Approval**. Research shows that self-serving biases become even stronger when people are endorsing others' biased judgements – provided those judgements align with their own biases – than when the are making original judgement themselves.

They suggest that there is a need to provide for auditor independence and removal of the threat of being fired for unfavorable findings. Perhaps VIAs could benefit from similar provisions.

This study has additional limitation that others interested in this work should consider.

- The contrast ratings made for both the DOE and SEC reports were made in the office, not in the field. The BLM (1986) has long stipulated that contrast ratings need to be made at the KOPs in the field.
- The two firms used slightly different photosimulations, though the viewpoints are very near to each other. The selection of a simulation's viewpoint in itself could be a form of client bias (SULLIVAN et al. 2021).
- The simulations used in this study are based on summer-like photography; the DOE evaluations were done in the winter and the date of the SEC ratings is unknown. There is evidence suggesting that the field evaluations should be conducted in the season represented in the simulations. PALMER (1990) found that "when people evaluate scenic quality, they do so within their present seasonal context."

5 Conclusion

The public is justified in expecting that VIA professionals produce reliable reports. Three ways to evaluate the reliability of visual impact judgements for KOPs are demonstrated. In this case study, the ICC(2,k) reliability of the six evaluators is 0.924, 0.888 and 0.894 respectively for the three alternatives. This result is very high, as it should be for professional services. In contrast, the ICC(2,1) reliability for a single rater is 0.655, 0.570 and 0.585, which is unacceptable for professional services. The implication is that multiple trained professionals must be used to evaluate the visual impact at each KOP.

A second approach to reliability is to determine the stability of the evaluations over time. This case study compared the same 15 KOPs evaluated for two alternatives by the same individuals using the same procedures at different times, nearly two years apart. The test-retest mean Pearson correlations are 0.847 and 0.868, which indicates substantial stability.

Finally, the results from two VIAs prepared by different firms for different clients are compared for six KOPs that were evaluated in both VIAs. The Pearson correlation between these firms' evaluations is 0.842, which shows high consistency comparable to the test-retest reliability. However, when the interpretation thresholds for impact severity are applied to the individuals' ratings, the firm whose client was the developer was much more likely to assign a lower-level impact rating than the firm whose client was the government permitting agency.

Reliability is an important attribute of professional services and should be required by regulatory agencies and demonstrated as part of VIAs. This requires that a panel of trained evaluators independently rate the same KOPs using the same methods. A rigorous demonstration of reliability would include test-retest evaluation, and a comparison of VIAs prepared by separate firms for different clients.

References

BAZERMAN, M. H., LOWENSTEIN, G. & MOORE, D. A. (2002), Why good accountants do bad audits. Harvard Business Review, 80 (11), 96-102, 134.

BISHOP, I. D. (2011), What do we really know? A meta-analysis of studies into public responses to wind energy. World Renewable Energy Congress-Sweden, 8-13 May 2011, Linköping, Sweden, 57 (15), 4161-4168. Linköping University Electronic Press.

BUJANG, M. A. & BAHARUM, N. (2017), A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. Archives of Orofacial Sciences, 12 (1), 1-11.

COREY, D. M., DUNLAP, W. P. & BURKE, M. J. (1998), Averaging correlations: Expected values and bias in combining Pearson rs and Fisher's z transformations. Journal of General Psychology, 125 (3), 245-261.

INTERNATIONAL ELECTRIC TRANSMISSION PERCEPTION PROJECT (IETPP) (1996), Perception of Transmission Lines: Summary of Surveys and Framework for Further Research. Edison Electric Institute, Washington, DC.

KOO, T. K. & LI, M. Y. (2016), A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15, 155-163.

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

PALMER, J.F. (1990), Aesthetics of the northeastern hardwood forest: the influence of season and time since harvest. In: Proceedings of the 1990 Northeastern Recreation Researchers Symposium, MORE, T., DONNELLY, M., GRAEFE, A. & VASKE, J. (Eds.). Gen. Tech. Rep. NE-145. Northeastern Forest Experiment Station, Radnor, PA., 185-190.

PALMER, J. F. & HOFFMAN, R. E. (2001), Rating reliability and representation validity in scenic landscape assessments. Landscape and Urban Planning, 54 (1-4), 149-161.

RIBE, R. G. (1989), The aesthetics of forestry: What has empirical preference research taught us? Environmental Management, 13 (1), 55-74.

SADLER, B. (1969), International Study of the Effectiveness of Environmental Assessment. Canadian Environmental Assessment Agency and the International Association for Impact Assessment: Ottowa, ON.

SHEPPARD, S. R. J. & NEWMAN, S. (1979), Prototype visual impact assessment manual. SUNY College of Environmental Science and Forestry, Syracuse, NY.

SMARDON, R. C., FEIMER, N. R., CRAIK, K. H. & SHEPPARD, S. R. J. (1983), Assessing the reliability, validity and generalizability of observer-based visual impact assessment methods for the Western United States. In: Managing Air Quality and Scenic Resources at National Parks and Wilderness Areas, ROWE, R. D. & CHESTNUT, L. G. (Eds). Westview Press, Boulder, CO.

SMARDON, R. C. & HUNTER, M. (1983), Procedures and methods for wetland and coastal area visual impact assessment (VIA). In: The Future of Wetlands: Assessing Visual-Cultural Values, SMARDON, R. C. (Ed.). Allanheld Osmun, Totowa, NJ.

SULLIVAN, R. G., MEYER, M. E. & PALMER, J. F. (2021), Evaluating photosimulations for visual impact assessment. Natural Resource Stewardship and Science, Air resources Division, National Park Service, Lakewood, CO.

T. J. BOYLE ASSOCIATES, (2017), Visual Impact Assessment: A Technical Report for the Northern Pass Transmission Line Project Final Environmental Impact Statement. US Dept. of Energy, Washington, DC.

TERRENCE J. DEWAN & ASSOCIATES (2015), Northern Pass Transmission Line Visual Impact Assessment. Terrence J. DeWan & Associates, Yarmouth, ME.

U.S. DEPARTMENT OF INTERIOR, BUREAU OF LAND MANAGEMENT (1986), Visual Resource Contrast Rating; BLM Manual H-8431-1; USDI, BLM, Washington, DC.