VALIDITY AND RELIABILITY OF RATINGS OF SIMULATED BUILDINGS *

R. W. Seaton and J. B. Collins

School of Architecture
University of British Columbia, Vancouver

Office of Academic Planning
University of British Columbia, Vancouver

How does one represent a designed environment before it is built? What are the right questions to ask about such a representation? Real or proposed physical spaces are notoriously difficult to model or manipulate experimentally, not only because they are expensive and time-consuming to construct but also because they are highly complex, their effect may be revealed only over extended time, and their connotations will vary with different kinds of self-selected users in variously-defined groups.

In a test of scales and simulations, the exteriors of 4 recently-constructed campus buildings were evaluated on five 7-point labelled scales by 2 control groups (N=38 each) of subjects randomly selected from a pool of 304 naive adults. Other equal-size groups from the same pool evaluated the same 4 buildings simulated by either 3-dimensional models, or color photographs, or black and white photographs. Some groups viewed each building from one position only, while others viewed each building from two positions.

In general, the 4 buildings rated high on some scales and low on others, and the different simulations did not much affect average ratings pooled across buildings. What the simulations did significantly affect (on all 5 criteria) were the relative mean values between different buildings.

This finding, if valid, bears importantly both on user preconstruction judgments between design alternatives and (often) on the post-construction selection of buildings meriting architectural awards. The architectural simulation is apparently not typically a psychological surrogate for the real facade.

Introduction and Overview

Research on the behavioral and aesthetic implications of architecture has been impeded by very considerable methodological difficulties.(1) Firstly, the unit demarking the independent variable (typically, an alternative design of a building or environment) is extraordinarily complex, so that it is most difficult to relate behavioral differences observed between alternatives with particular structural differences between them. Secondly, the effects of structure (of environments or buildings) on inhabitants may become manifest only over a very long time frame ... perhaps years. Thirdly, effects on inhabitant groups may be quite different from effects on individual visitors or clients. Fourthly, even if two structures are very similar in all but a few significant features, an investigator would be hard-pressed to establish that behavioral differences between their inhabitants are attributable to differences between the structures rather than physiological or psychological differences inherent between inhabitant populations who usually are self-selected ... people in social groups are "confounded," in an experimental design sense, with places. Finally, buildings, and other designed environments which characteristically surround their inhabitants while the latter move about within them, are most difficult to simulate in research laboratory settings.

Given these difficulties, most researches on behavioral results of architectural design are restricted to case studies. As in the early stages of scientific investigation of any field of natural phenomena, case studies provide insight into the range of aspects of human behavior which can be affected by environmental (and especially architectural) forms. However, except in isolated instances, such studies do not allow one to make assertations about which features of the environment are associated in a causative way with which features of observed behavior.

Given the difficulties of assessing the affects of complex structures, two paths are open. Firstly, a sufficiently large population of case studies, each applying at least roughly equivalent methods of evaluation to standard features of behavior, can be accumulated so that multivariate or cross-classification analytic techniques can be applied.(2)

Secondly, one can attempt to experiment with structures, either in full scale or real time or in terms of simulated settings abbreviated in time and space. Full-scale real time simulations are relatively rare. There are two kinds: the "natural experiment" and the full-scale mockup. Natural experiments are naturally infrequent, while still retaining many of the disadvantages of the case study. Full-scale mockups are expensive to create realistically and difficult to investigate because often the process of investigation itself reduces their realism.(3) Accordingly, behavioral scientists have turned to small-scale

simulations which allow them to make practical application of the established experimental laboratory methodology of the behavioral and life sciences.(4)

Architectural Simulation

Small-scale representations, simulations and mockups of built places are the very "stuff" of architectural practice; in a sense, any time a designer sketches alternative forms and judges one is better than the others, he conducts a simulation experiment. The great difficulty is that one never knows whether responses to a simulated architectural environment (c.f., e.g., Ritter & Hibb, 1969) or conversely respondents' expression of environmental preferences through the manipulation of small modeling (e.g., on sandtable ... see Michelson, 1966) are the same as when architectural forms are expressed in full scale.

For example, when one views a model on a table, his perspective and the angle the model subtends at his eye are different from when he views the image at eye level. Sky-lighting, shadowing and the surroundings extending from the model or mockup may also lack verisimilitude. Because small-scale simulations typically can more readily be seen in their entirety from one viewing position, building designs may impart a greater sense of integration than would be inferred by a viewer at ground level. Perspective drawings or photographs may avoid some of these difficulties, but two-dimensional approaches to simulations of buildings and designed places, have other faults in addition to their inherent two-dimensionality.

Experimentation with architectural simulations become even more difficult when one is concerned with the insides of structures surrounding their occupants, rather than merely external facades (Langdon, 1970; Lau, 1970; Winkel & Sasanoff, 1970). And the whole question of the merits of simulations becomes extraordinarily complex when we begin to discover ... as we will ... that some simulations give good verisimilitude for some purposes but not others; thus in experimentation the simulation technique one uses will vary with the attitudinal, social or behavioral dependent variables measures, as well as the kinds of forms and spaces to be represented.

Yet simulation is vital to the development of a body of knowledge in architecture and planning which includes generalizations about how features of form shape human responses. Valid simulation also has great practicality, in that it permits one to have confidence in his or others' interpretations of the effectiveness of proposed designs (represented by small-scale modeling in two or three dimensions) upon human functioning. There are as yet no grounds for such confidence. To put it bluntly, all h'mmming and nodding of heads that occurs today when hundreds of architects present thousands of pretty sketches and handsome models to scores of thousands of clients' building committees is just as significant, and no more, as the ooh'ing and ah'ing of the same people at a noontime fashion show in a men's bar. The sole purpose of the exercise is sociological rather than critical, for the viewers are unskilled in reading architect's plans and the intentions expressed in his plans. Even where it is otherwise, predictions of what people will do in planned buildings (rather than what they are supposed to do) are rarely if ever corroborated.

Architectural Scales

A consideration of "architectural meaning" involves an attempt to define what we mean by "mean". Invoking Osgood's language of the "sign" and "significant", architectural meaning (like semantic meaning) involves the mediational links between the significate (the architectural display) and the signs used to represent it (the word descriptors to which we respond). However, the incautious assumption that a semantic differential applied to a stimulus object (building) yields equivalent information as a differential applied to a stimulus concept (word - itself a sign) may have led us to conclude yet unproven conjectures about the nature of architectural meaning. Osgood stands on firm ground in using a sign (word) to investigate meaning of another sign (word). As yet, no one in the environmental-behavioral disciplines has demonstrated a technique to study the meaning of a significate (building) via the medium of a "significant" differential. Pending such discovery, architectural "meaning" can only be investigated in terms of threads of unity found among multiple samples of respondents, investigators and procedures, through the emergence of similarly ranked loadings on common sets of dimensions gathered in a multiplicity of settings.

Despite inherent weaknesses of prior studies, a closer look reveals an emerging pattern of what constitutes architectural meaning. The most frequently referenced users of the semantic differential approach are Vielhauer (1965), Canter (1968), Hershberger (1969), Craik (1969), Collins (1970) and Janiskee (1971) (whose work is not yet in print). A compilation of the findings of these six investigators affords a look at their factor structures (arranged by rotated factor emergence).

There is noteworthy agreement among the six on the first factor. All find a factor strongly indicating aesthetic evaluation; but for none is the loading pattern identical to Osgood's

semantic evaluation. Note that good-bad appears well down the ranked loadings, if at all. Also for four of the six, there is clear evidence of a confounding of aesthetic evaluation with activity (loadings on dynamic, exciting, revolutionary, lively, active).

The second factor emerges almost as cleanly for the six. All report high loadings on neat, orderly, tidy, clear, stable, simple, calm, peaceful, etc. Closer inspection of lower ranked loadings suggests that there may be some confounding of orderly, tidy, neat into simple, rational, straightforward.

Four of the six studies show a third factor related to size, but usually broken into separate components of physical size or phenomenological size. For Hershberger, spaciousness and open were confounded into Factor I; his third factor then reflected strength, boldness, etc. Janiskee did not include spaciousness but instead had a third factor relating to distance and accessibility.

TABLE 1

Factor Structures for Six Researchers of Environmental Descriptors*

	Factors I	II	III
Vielhauer	pleasant appealing inviting gay cheerful	neat orderly tidy organized clean	roomy free space uncrowded comfortable
Canter	pleasant interesting lively active	tidy clean coherent clear stable	spacious constant flexible
Hershberger	cheerful welcoming beautiful active exciting interesting (spacious)	ordered clear strtfrwrd. rational simple	strong bold profound rugged good
Craik	dismal distressing expressionless glamorous gay	calm cozy civilized	big huge elongated broad
Collins	interesting inviting dynamic exciting animated	peaceful quiet orderly neat secure	spacious roomy uncluttered open
Janiskee	interesting beautiful pleasant impressive	constant stable calm inflexible	near-far central accessible

*Collins, J.B. Scales for Evaluating the Architectural Environment. Presented at American Psychological Association Convention, Washington, D.C., September 3-7, 1971.

Experimental Design

The major classes of factors which affect judgments about architectural space have been well summarized by Brunswick et al. (1943) and Craik (1968). Judgments depend on what (kind of) spaces are being judged, who is doing the judging, what kinds of judgments are being asked of the judges, how the different building spaces (stimuli) are represented (simulated), and under what conditions judges view the representations. The experiment described below attempts investigation of three of these factors.

The experiment focused on judges' responses to the overall exterior form of four different buildings on the University of British Columbia campus which were constructed within the past decade. Thus the study gave explicit consideration to the first factor discussed by Craik: the kind of space (facade) to be judged varies at four levels.

When building facades are being judged, it would be better if judgments were not contaminated by previous knowledge about the particular buildings. Fortunately, the University in 1970 hosted a tri-annual "open house" for the regional community, with many persons (e.g., parents) visiting the campus for the first time. These persons made good subjects for experimental purposes, both because they could give judgments based only on what is shown to them and because in most cases they came to the campus expressly to look at buildings as well as other things.

Thus the second factor of Craik (who is doing the judging?) was not varied. Judges were visitors to, rather than users of, the campus. They presumably were unfamiliar with the test buildings. The buildings tested were not differentiated to subjects with respect to purpose but instead were appraised under the rubric "campus buildings."

An advantage of full-scale or small-scale exterior models is that they can be viewed from various perspectives. However, small-scale mockups and full-scale representations are very expensive. Much cheaper representation can be obtained with plans, perspectives, elevations and sections. These in turn may be ineffective in inducing valid judgments; this failure may derive either from the fixed viewing perspective which graphic representations entail, or from their lack of three-dimensional depth.

The four test buildings were each visually represented to judges in four different ways: in

full scale, in scale models, in color photographs and in black-and-white photographs. These various representations or simulations constitute a test of Craik's third factor: the way that buildings are represented to judges.

To provide a limited test of the fourth factor discussed by Craik ...the conditions of judgment with respect to simulation and viewing angle ...judges were asked to view real buildings or their three-dimensional models from one or alternatively two perspectives, which were also specifically those used in the two-dimensional photographs of the same buildings.

A final consideration raised by Craik is the kind of judgments asked, with respect both to quality and scale of measurement (e.g., rankings vs. ratings). Previous research suggests that people tend to conceive of their environment in terms of (among other things) peacefulness, strength, orderliness, and potential interest-excitement. A general dimension is pleasingness. Each subject was asked to judge each of the four buildings in terms of verbal scales reflecting these conceptual dimensions.

Quite likely, some ways of simulating space are valid in terms of some evaluative dimensions and not in others; however, in this experiment only four fixed simulations and five fixed conceptual dimensions were considered with respect to four selected buildings. Otherwise the scope of this study, already highly complex, would have had to be expanded to include consideration of a wide range of dimensions in terms of which subjects can judge space.

Details of the Experiment

The experiment consisted of the evaluation of the facades of four different buildings (see Figures 2-5) on the University of British Columbia campus under four different simulations, each viewed from two positional alternatives, with each test subject using five rating scales. Generally, then, the experiment had a 4 x 4 x 2 x 5 "mixed" design (partly factorial and partly hierarchical) with buildings, simulations positional alternatives and rating scale criteria forming the four variables. Test subjects, all aged 16 or older, were recruited from passers-by attending the triennial University of British Columbia Open House. Each test subject was initially handed an instruction sheet briefly describing the project. Next, he had a "dry run" on the five different criterion scales (Figure 1), using as stimulus an architect's drawing of a proposed campus building (Figure 6). After the subject read the instructions and completed the "dry run," his responses were scanned for assurance that he understood use of the scales and was aware that any of seven scale levels could be used in responding. He then became one

ARCHITECTURAL EVALUATION PROJECT

Researchers in Architecture and Environmental Psychology are concerned with determining how the visitor to the University of British Columbia campus experiences and evaluates the various buildings on the campus complex.

A few minutes of your time will greatly assist us in planning more effectively for persons who visit the campus less frequently than faculty or students.

To familiarize you with our information collecting procedure, we would like you to examine an Architect's drawing of the Pharmacy Building Addition and to report your impressions of it on the following set of five scales.

Study the building -- then consider each scale and mark an "X" at that point which to you represents the best assessment of the building.

PEACEFUL / QUIET — PEACEFUL / QUIET
none at all slightly somewhat moderately considerably very extremely

STRONG / BOLD — STRONG / BOLD
none at all slightly somewhat moderately considerably very extremely

DYNAMIC / EXCITING — DYNAMIC / EXCITING
none at all slightly somewhat moderately considerably very extremely

ORDERLY / TIDY — ORDERLY / TIDY
none at all slightly somewhat moderately considerably very extremely

PLEASING / APPEALING — PLEASING / APPEALING
none at all slightly somewhat moderately considerably very extremely

When you have completed this sample evaluation, the project director will provide you with an Evaluation Booklet and will direct you to a viewing station at which you will assess actual campus buildings.

Thank you for your interest and cooperation!

Figure 1. The test form, showing the scales

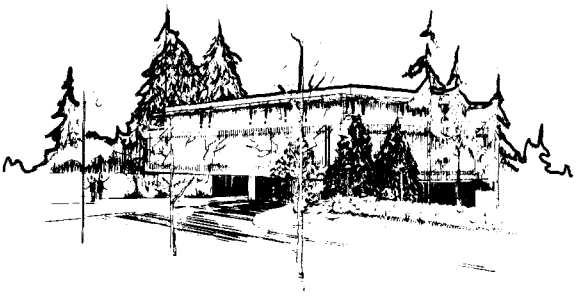

Figure 2. Graduate Student Centre

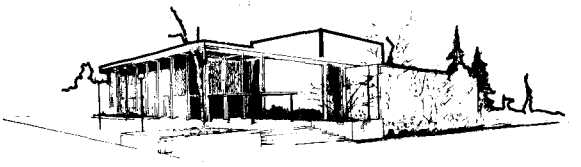

Figure 3. Frederic Wood Theatre

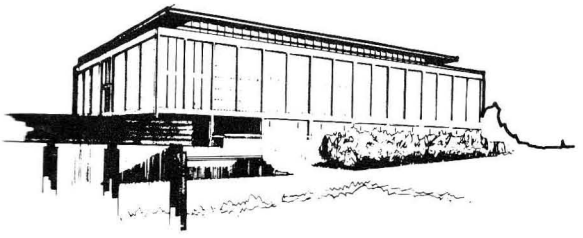

Figure 4. Lasserre Building (architecture, art and planning)

Figure 5. Music Building

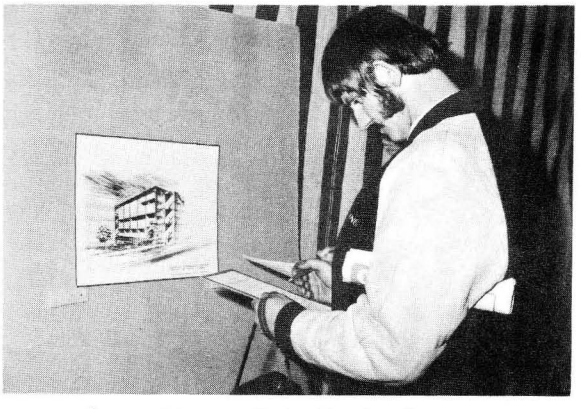

Figure 6. A "dry run" in the headquarters tent

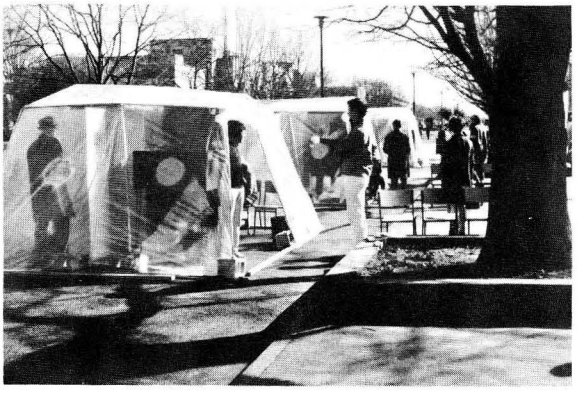

Figure 7. Tents housing models and photographs

Figure 8. Model viewing tent. Note eye-level viewing tunnel on box at right

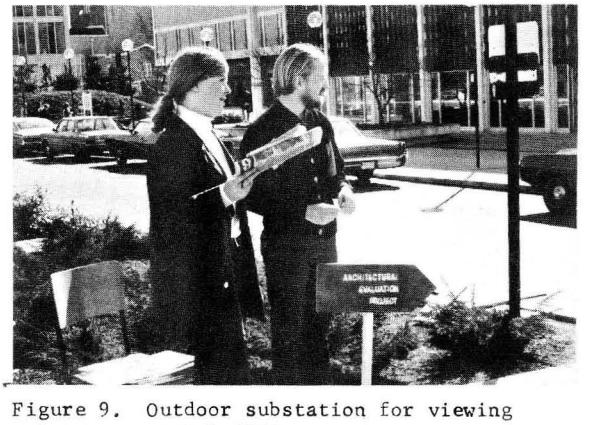

Figure 9. Outdoor substation for viewing real buildings

of 38 subjects assigned to each of eight test conditions (one vs. two viewing positions, x 4 modes of representation). He received a pack of five 5-inch by 8-inch cards, including a "cover" card and four answer cards (Figure 1). Each answer card applied to a particular building facade, identified to the test subject only by number; the sequence with which buildings were viewed by test subjects varied systematically between subjects, in eight different orders. The four answer cards (one for each building) each included five rating scales, in a fixed order throughout a subject's pack; however, different test subjects received packs in which the order of scales varied among five alternative sequences. At the bottom of each answer card was space for the subject to write in words describing each stimulus building.

After receiving his pack of cards, the subject was directed to one of four test stations at which one of the four experimental simulation modes was represented. Three stations were enclosed in translucent tents along the Main Mall of the campus (Figures 7 and 8); the fourth ("real building") simulation station was actually a pair of substations (Figure 9) from which subjects were directed to marked positions from which test buildings were to be viewed. Subjects moved from building to building (or simulation thereof) in the order shown in their pack of cards, and after completion of their task they submitted their pack of cards to a station attendant.

All this sounds very neat and orderly, but of course it wasn't at all. Outdoor experiments are full of devilish resistances and obstacles which rise to confound the experimenter. The tents housing the simulations, for example, had to be made of a reasonably light polyethylene stretched on a metallic frame. The structures were not very strong, which became manifest on the first day of Open House, characterized by drenching rains and high winds. The models and photos were saved at the last minute, before the tents collapsed. This was as well, for the models each cost several hundred dollars, being hand-made to exactly the same scale for all four buildings, using exactly the same qualities of materials and finish for each, to architectural standard. The same is true of the photographs taken by a professional architectural photographer at comparable fees. Then too, getting the subjects ordered into test groups dizzied our lives for a few days, for we had eight different arrangements of the stimuli times five arrangements of the scales times four different simulation color codes times two viewing position codes for each simulation, making a total of 320 different decks of cards to order, each replicated twice and thereafter systematically interleaved so as to achieve practical balance of sequencing and treatments. Then there were the usual personnel problems ... our dozen test monitors quit for lunch just when the big crowds came. It was a frenzied day.

A total of some 600 cases were run, but many had to be discarded due to being under age, or being students or failing to complete the full set of response cards. The latter difficulty particularly occurred in the viewing of the four real buildings, which had to be visually separated from each other and therefore had long distances between them. Time delays entailed in viewing the models, which were to be squinted at eye level through narrow slots also lead to the loss of some subjects. After screening the responses for all sources of error and omission, the least populous of the eight test subject groups held 38 cases, so other cells were pruned down randomly to this size (with some loss of reliability) so that all cells could have equal N's for reasons of computational convenience.

Results

Under the experimental design used, each test subject was put in a group (N=38) which used five scales to rate the four stimulus buildings. It would be nice if a person's twenty ratings were independent of each other, so that means and variances calculated over persons could be regarded as tantamount to independent replications of test effects. To this end, the order of looking at buildings and the ordering of dimensions judged for each building were varied systematically. However, independence of scales was not fully achieved.

Others' factor analyses of semantic studies indicated common loadings (see Table 1) on some of the scales selected for test: thus "pleasing, appealing" appears to have a factor structure relation to "exciting, dynamic", and "orderly, tidy" may share some common structure with "quiet, peaceful".

Intercorrelations among the scales are all positive and mostly of reliable order. Those of the pleasing, appealing scale with others are the highest. Noteworthy are the intercorrelations for pleasing, appealing vs. dynamic, exciting and for dynamic, exciting vs. strong, bold. Clearly there was a strong "halo" affecting raters' evaluations of buildings on supposedly independent dimensions.

On the other hand, the four test buildings did differ in their ordinal positions on different scales (see right-hand column, Table 2). Generally, the Graduate Center and the Music Building received higher marks than the other two buildings, but not on all of the scales. Also, in all but one case, the four buildings were rated lower on the "dynamic, exciting" dimension than on the other scales; this suggests a blandness or austerity of frame-and-

TABLE 3

Mean Ratings (N=38) of Buildings Varying in Representation and Position

Scale	Building	Mode of Building Representation: Real Pos 1	Real Pos 2	Model Pos 1	Model Pos 2	Color Photo Pos 1	Color Photo Pos 2	B&W Photo Pos 1	B&W Photo Pos 2	Mean (N=304)
Pleasing, appealing	GC	4.63	4.50	4.03	4.66	4.71	4.39	4.37	4.10	4.42
	FW	4.61	4.29	3.71	4.34	4.16	4.82	3.51	3.79	4.12
	LA	3.71	3.47	4.05	3.66	3.34	3.61	4.05	4.13	3.75
	MB	4.71	4.50	4.16	4.71	4.47	4.47	3.68	3.95	4.33
	Mean	4.41	4.19	3.99	4.34	4.17	4.32	3.91	3.99	4.17
Dynamic, exciting	GC	3.39	3.71	3.76	3.50	3.40	3.45	3.24	3.32	3.47
	FW	3.97	3.63	3.45	3.42	3.50	3.74	2.97	3.29	3.50
	LA	3.05	3.26	3.24	3.13	3.03	2.97	3.58	3.34	3.20
	MB	4.84	4.26	3.63	3.84	4.08	4.42	3.82	3.87	4.10
	Mean	3.82	3.72	3.52	3.47	3.50	3.64	3.40	3.45	3.57
Orderly, tidy	GC	4.66	4.53	4.08	4.50	4.66	4.76	4.37	4.16	4.46
	FW	4.45	4.39	4.21	5.03	4.47	4.79	3.89	3.50	4.34
	LA	4.24	3.53	5.21	5.03	4.00	3.82	4.92	4.74	4.43
	MB	4.89	4.58	4.82	5.11	5.53	5.26	4.63	4.58	4.92
	Mean	4.56	4.26	4.58	4.91	4.66	4.66	4.45	4.24	4.54
Strong, bold	GC	3.45	4.13	4.76	4.00	3.84	3.84	3.89	3.68	3.83
	FW	4.34	3.71	3.97	3.58	3.92	3.89	3.71	3.39	3.82
	LA	3.82	3.66	4.18	3.58	3.68	3.32	4.63	4.05	3.86
	MB	5.37	5.24	4.37	4.16	5.05	5.13	4.55	5.08	4.87
	Mean	4.24	4.18	4.07	3.83	4.13	4.05	4.20	4.05	4.09
Peaceful, quiet	GC	5.05	4.79	4.13	4.03	4.47	4.79	4.08	4.13	4.42
	FW	3.58	3.95	3.82	4.03	3.79	3.74	3.18	3.16	3.65
	LA	3.03	3.84	4.16	3.79	3.61	3.58	3.63	3.76	3.67
	MB	3.74	3.55	3.76	3.84	4.47	3.76	3.37	3.39	3.72
	Mean	3.85	4.03	3.97	3.92	4.09	3.97	3.57	3.61	3.87

sheath architecture on the campus studied.

Validity. A key issue in this study is whether one can simulate the outside of buildings by models or photographs in order to create viewer impressions veridical with those generated by real buildings when viewed from the outside. By way of an example of our results, Table 2 gives mean ratings on the pleasing, appealing scale of the four stimulus buildings for each mode. On the seven-point scale (see Figure 1), an overall average of 4.17 as shown in the lower right hand corner of the pleasing, appealing data in Table 2 indicates that our four test buildings impressed visitors as no more than moderately appealing. We were glad to see that the real buildings generally rated as high (in the "mean" row) as did their doublegangers...but not much higher. We also see (along the extreme right column) that the scale discriminates clearly between the test buildings, with lower ratings allocated to FW and LA than to the other two buildings. Note, however, that this overall pattern is not replicated under different simulations; in the color photo mode, FW is a winner on the pleasing, appealing scale, and in the black-and-white photo mode LA rates higher than MB.

Table 2 confirms the reliability of observed variations in ratings. Because the experimental design is mixed, being partly factorial and partly hierarchical, different error terms are used to estimate the significance of different effects. The analysis confirms that the average ratings of the different buildings do in fact differ significantly, but that the relative pleasantness or appealingness of a building much depends on how it is represented to the public ... that is, there is a significant Building-Simulation interaction.

Note that the position variable ... with levels depending on whether a person appraised a building or model from just one viewpoint angle (or saw just one photograph), or from two ... has no main effect on the data nor any significant interaction with other variables; this was found generally to be the case across all scales.

TABLE 3

Analysis of Variance of Pleasing, Appealing Ratings

Source	df	MS	Error	F
Position of view	1	2.58	A	<1
Simulation mode	3	7.25	A	1.79
Pos. x Sim.	3	4.37	A	1.08
Subjects (error A)	296	4.05	B	1.76***
Buildings	3	26.84	B	11.66***
Building x Pos.	3	2.29	B	<1
Building x Sim.	9	7.20	B	3.13***
B x P x S	9	2.33	B	1.01
Subj. x Bldg. (error B)	888	2.30		
Total	1,215			

***p<.01

On all of the scales for which data are reported in Table 2, the main effect of buildings tends to emerge more strongly among the real buildings than among the simulations; on two of the five scales, the top-rated real building is scored more than a full scale point higher than its competitors. While the same pattern of differentiation also appears in the simulations, it does so much less markedly. Of the three simulations, that which seems to gain results more closely resembling those elicited by the real buildings is the color photograph mode, while ratings obtained via the models and black-and-white photos tended to blur the contrasts noted just above.

TABLE 4

Summary of Three ANOVA Tables

Scale	Average value of ratings	MS Error B	Probability of F ratios: Bldgs.	Bldg. x Sim.	Other
Dynamic, exciting	3.57	2.17	<.001	<.02	n.s.
Orderly, tidy	4.54	1.59	<.001	<.001	<.05*
Strong, bold	4.09	1.93	<.001	<.001	<.02**
Peaceful, quiet	3.87	1.79	<.001	<.001	n.s.

*Simulation effect
**Building x Position interaction

Data in Table 4 summarize the ANOVA tables for the four scales dynamic, exciting; orderly, tidy; strong, bold; and peaceful, quiet. On all scales both the Building main effect and

Building-Simulation interaction were highly significant, telling us both that the buildings differed reliably in terms of the scales, and that the nature of these differences varied with simulation.

To determine more exactly how the simulation modes represented reality, the means shown in Table 2 for each of the five scales in the "real" mode were used as standards against which to correlate means obtained from simulations. The two viewing positions were treated as replicates. Thus, for a given scale (e.g., pleasing, appealing), mean scores for both positions in the "real" mode were first ranked and then correlated by means of Goodman and Kruskal's gamma(5).

By this means, four correlations were generated between the two positions for "reality" and the two positions for any given simulation; over five scales, twenty gamma values resulted. The gamma values for a given pair of simulation positions were averaged (arithmetically) over all five scales. Resulting values are shown in Table 5.

TABLE 5

Mean (N=5 Scales) Gamma Values between Eight Subject Groups (N=38) Viewing Test Buildings under Alternative Representation Modes and Positions

Positions correlated	Representation: Real	Model	Color Photo	B&W Photo
1 vs 2	.60	.32	.73	.86
Real position 1 vs. simulation position 1	-	.07	.80	.00
Real position 1 vs. simulation position 2	-	.47	.80	.00
Real position 2 vs. simulation position 1	-	.17	.71	.17
Real position 2 vs. simulation position 2	-	.76	.51	.20

The data in Table 5, showing gamma correlations averaged over scales, indicates the rating data from black-and-white photographs tends to correlate highly (first row) with itself (i.e., is reliable) but has a very low positive relationship with reality. The best average performance over all scales, is provided by the color photographs, which correlated highly both with themselves and also with reality. The models seemingly do well at predicting reality only when both models and reality entail viewing a building from two distinct positions.(6)

By a somewhat analogous procedure, the 16 mean gammas in Table 5 were decomposed and re-averaged over scales, with results as shown in Table 6. The better average gamma performances, over all representations and positions, are those deriving from the dynamic, exciting and strong, bold scales.

Thus, taking Tables 5 and 6 in combination, it would appear that greatest degree of verisimilitude using simulations obtains when one uses color photographs to appraise boldness, excitingness or like concepts.

TABLE 6

Mean (N=16) Gamma Values between Reality and Simulation Ranks of Buildings on Five Scales

Pleasing, appealing	.30
Dynamic, exciting	.63
Orderly, tidy	.36
Strong, bold	.53
Peaceful, quiet	.42

Reliability. Eight subject groups (N=38 each) under varying viewing conditions used five scales to evaluate the test buildings. From these data, forty separate analyses of variance were generated, one for each scale in each condition. A typical example of one ANOVA is shown in Table 7. Note that the mean square error is somewhat high in this example (relative to 7-point scales generally) and inter-person error is very low (relative to most rating groups).

TABLE 7

Analysis of Variance of Pleasing, Appealing Ratings of Real Buildings Viewed from Position 1

Source	df	Sum of Squares	Mean Square	F
Buildings	3	25.33	8.45	3.32*
Persons	37	93.14	2.52	.99
Error	111	282.41	2.54	

*$p < .05$

The error term in the example (Table 7) is an estimate of the reliability of obtained scale values of buildings represented under a specified set of circumstances. To estimate scale reliability under varying circumstances, the forty error mean square values were extracted from the forty ANOVA and arrayed in a table showing error values for five scales, four simulations and two positions. Marginal median values for the five scales, and for the eight test groups, are shown in Table 8.

TABLE 8

Median Values of Mean Square Error Terms

Scale		Simulation and Position	
Pleasing, appealing	2.44	Real, position 1	2.20
		Real, position 2	2.37
Dynamic, exciting	2.26	Model, position 1	1.63
		Model, position 2	2.24
Orderly, tidy	1.56	Color Photo, pos. 1	1.89
		Color Photo, pos. 2	1.80
Strong, bold	1.91	B&W Photo, position 1	1.77
		B&W Photo, position 2	1.79
Peaceful, quiet	1.69		

To test the reliability of observed differences in average values, the forty mean square errors were transformed into natural logarithms; means were calculated on transformed values and related back to values in Table 8 with only trivial differences emerging. Then the transformed data were subject to analysis of variance with 39 degrees of freedom. Results are shown in Table 9. The results show a clear-cut main effect for scales. Reference to Table 8 shows

TABLE 9

Analysis of Variance of Forty Transformed Error Variances

Source	df	Sum of Squares	Mean Square	F
Scales	4	666.2	171.6	10.3***
Representations	3	163.6	54.5	3.3
S x R	12	205.2	17.1	1.0
Positions	1	1.6	1.6	0.1
S x P	4	26.7	6.7	0.4
R x P	3	205.3	68.4	4.1*
Error	12	199.1	16.6	

*p= <.05
***p= <.001

lowest subject-building "error" occurring when the orderly, tidy scale is used, while relatively poor consistency obtains when the pleasing, appealing and dynamic, exciting scales are used. Seemingly our test subjects were less clear in concensus about what is pleasing or dynamic than they were about orderliness or peacefulness.

The analysis of variance of error variances (Table 9) also suggests (at about the 5% level) some reliability in observed differences in error terms between representations and positions. The data (Table 8) indicate lowest concensus among persons judging the real things rather than the simulations. (It were ever thus!) Better reliability seems to obtain for judgments of photographs, while judgments of models varied in unreliability, depending on viewing position.

A technical difficulty in building evaluations is moving judges around the country judging different structures. Things would be easier (and cheaper) if from a population of judges one group is randomly selected and sent to look at this building, another sent to look at that building, and still others randomly selected to appraise a third, fourth, etc. In such a case, comparisons between judge groups would depend for reliability estimates on variation between judges within groups.

The forty analyses of variance discussed in connection with Table 7 were first arrayed (as in Table 8) with respect to inter-subject variances, then transformed into natural logarithms and subject to analysis of variance. None of the values, which were similar relationally but not absolutely to those in Table 8 were found reliable except that (shown in Table 10) relating

TABLE 10

Inter-Judge Variance in Eight Groups Varying in Mode of Representation and Viewing Position

Mode of Representation	Viewing Position 1	2
Real	4.3	4.5
Model	4.7	5.2
Color photo	4.1	3.8
B&W photo	6.2	3.4

to an interaction between mode of building representation and viewing position. As can be seen, the interaction derives from greater inter-judge reliability obtaining from viewing two photographs while slightly lower inter-judge reliability happens when a real building or model is viewed from several angles. This interaction, significant at the .05 level, may however be artifactual.

Power. Reliability of scales and viewing conditions is important, but the ultimate concern is for power in measurement. One measure of power, applying to type I error, is the F-ratio relating stimulus (building) variance to error variance.

To test power obtaining under the various measurement conditions appraised, F-ratios for buildings were extracted from the forty analyses of variance illustrated in Table 7. Marginal median values obtained are shown in Table 11. As we might have expected from earlier analysis

TABLE 11

Median Values of F-ratios for Buildings

Scale		Simulation and Position	
Pleasing, appealing	3.4	Real, position 1	10.1
		Real, position 2	5.0
Dynamic, exciting	2.7	Model, position 1	1.3
		Model, position 2	1.6
Orderly, tidy	5.7	Color Photo, pos. 1	5.5
		Color Photo, pos. 2	7.3
Strong, bold	8.3	B&W Photo, position 1	3.7
		B&W Photo, position 2	4.1
Peaceful, quiet	4.1		

of error variances, the orderly, tidy and strong, bold scales have highest power and the pleasing, appealing and dynamic, exciting scales show least discriminatory power.

To test the reliability of observed trends, the forty F-ratios were normalized by transformation to Fisher's z(1957, p.2 and Table V), to which unity was added so that ratios lower than one would be expressed positively. The transformed values retained in large part the direction and extent of relationships shown in Table 11. These values were then subject to analysis of variance according to the procedure outlined by Anderson (1961). Results show the scale effect on power to be suggestive but not significant at low level of probability. However, the mode of

representation effects appeared as quite reliable ($p < .01$). Seemingly, when viewing the models, the judges could not nearly as well distinguish between buildings in terms of the dimensions requested as when they viewed photographs or the real buildings. No other effects or interactions were found to relate reliably to power.

Summary and Discussion

The results summarized above suggest that the qualities that buildings impart to viewers are generally similar over simulations. All buildings scored relatively high on tidiness and low on excitingness, irrespective of the simulations used. On the other hand, the relative pleasingness of the Lasserre (LA) Building clearly improved in its black-and-white photographic rendering, and the same kind of observation can be made about other buildings' rankings on other dimensions in other renderings.

We were glad to observe that the Thea Koerner Graduate Students Centre (GC) rated generally quite well on the scales. This is good, for in 1962 this building won a Massey Gold Medal as a leading architectural design in Canada; had it scored poorly, the validity of our scales (or the validity of the award) would have become suspect. The Massey Medal is awarded solely on the basis of simulations ... judges do not travel about the country to view candidate structures in vivo. The results of the present experiment do not invalidate this procedure, they only make it suspect by revealing significant Building-Simulation interaction on every dimension tested.

The results of the present study do not definitively set the rules for valid use of surrogates. The seven-point scales used were somewhat arbitrary in format and content, the selection of buildings used was fixed and relatively homogeneous, the five dimensions studied were a fixed selection, and ratings on the five dimensions were performed concommitantly and therefore interdependently.

Nonetheless, the results seem to make sense. The error variances are reasonably good for scales of this length, the real buildings serve as more contrasty stimuli than their simulations, and there is a measure of agreement that some buildings are more outstanding than others on this or that dimension. The chief merit of the study is in the size of its sample of people, so that the reliability of trends can be firmly established. Also important is the consistency of procedures: all models were made by one artisan; all photographs were taken by a professional on a given day at a specified time; all representations were viewed from closely specified points; all subjects were run in one day, a day similar to that used for photography; all subjects were adults and strangers to the campus; all subjects were systematically spread among different treatment groups, and all test sequences were balanced between groups.

Validity, reliability and power data obtaining from the study seem at least consistent. Color photographs appear to give good representation of reality, relative to models and black-and-white photographs. The color photographs also provided good reliability and power. Viewing position (single vs. double angles) had little effect on the data. Among the scales, the strong-bold scale seemed to give good reliability and validity.

This work is, of course, only beginning. Appraisals of method are planned to extend further, to architectural renderings and photographs of models. Other students are exploring videotaping, modelscope photography and full-scale mockups. More work should be focussed on isolating the scales most pertinent to building facades. The present authors are now considering study of scaling and simulation of the interiors of proposed buildings; a set of new problems arises in these considerations.

Other studies (Peterson, Woodman, and Eaton, 1968; Lau, 1970; Holmberg et al, 1967; Galvin, 1970) have suggested that generally simulations appear to give results similar to those of reality; but of course such results can not be identical with those from reality, and the deviations between real and simulation results has not heretofore been subject to direct statistical test. The statistical tests herein all confirm that results from simulations are not congruent with those from reality, no matter what the scalar dimension, although they may be similar. These results argue for scepticism about the merit of evaluative judgment from simulations and models. We should sympathize with Leonard Fein when he remarks,

> I am dazzled, as is any layman, by the splendid models, complete to the last detail, of tomorrow's buildings. I am dazzled, but unpersuaded, for the models are the architect's, not mine. They are not mine even when he has finished explaining all their virtues and conveniences. They are not mine because their virtues and conveniences are the children of the architect's conception. They are, at best, a grafting of my groping hopes and the architectural wisdom. They are, as they are shown to me, an elaborate and compelling diagnosis, and they are, as well, an imposition. (1968, p. 198)

As Lowe (1969) remarks, today's architectural prize winners are tomorrow's fiascos and ruins, and between the conception and the creation, between the idea and the reality, remains the shadow of doubt.

Notes

*An earlier version of this paper was read at the Western Psychological Association Convention, San Francisco, April 1971.

1Prior to 1960, except in the housing field (see Beyer, 1965, for review of that literature), little architectural behavior research was performed (Evans, 1966) although there were many researches in the behavioral science fields...anthropology, psychology, sociology,...which had implications for environmental design (see SER, 1965, v.i.). The "state of the art" was similar to that faced by researchers two decades ago in bureaucracy and complex organization (see Selznick, 1949; Blau, 1955).

2Applications of this approach in social anthropology, for example, include Cohen's 1955 study of food-sharing practices or Murdock's 1949 analysis of kinship structures, both of which drew upon accumulated case studies of societies in the Yale University Human Relations Area File; for a review of comparable research in the area of formal organization, see Bass, 1965, or Blau and Scott, 1962.

3Examples of natural experiments appear in civil defense and disaster research literature...see Tyhurst, 1957; other well-known natural experiments are reported by Festinger, 1956 and Kerckhoff & Back, 1968. Experiments involving full-scale mockups are fewer, but can be illustrated in social psychology by the work of Sherif, 1961, and other followers of the psychologist K. Lewin; in bureaucracy by the work of the Non-Linear Systems Corporation...see Bass, 1965, pp.278-9; and in architecture in the report by Sanoff, 1965.

4In social psychology such experimental simulations of complex settings began after WWI and have since become almost without number (McGrath & Altman, 1966). Small scale simulations of bureaucratic structures began a decade later, for example in the work reported by the Gullahorns (1965) and the development of elaborate business games at Carnegie Institute of Technology (Cohen et al, 1964) and elsewhere; these have the merit of having been validated by correspondence with the outcomes of detailed case studies and natural experiments in the real world.

5Since each correlation entailed only four paired observations, use of gamma mitigated the effects of extreme values on the results.

6The impact of these validity data is modified by the consideration that the scales were not wholly independent, so that a high average correlation between reality and a simulation might be an artifact of the composition of the subjects in the two groups being correlated.

We are grateful to Mr. John McMaster, Office of Academic Planning, University of British Columbia, for programming the mixed-model ANOVA reported in this paper.

References

Anderson, N.H. Scales and statistics: parametric and nonparametric. Psychological Bulletin, 1961, 58, 305-316.

Bass, Bernard M. Organizational psychology. Boston: Allyn & Bacon, 1965.

Beyer, Glenn H. Housing and society. New York: MacMillan, 1965.

Blau, Peter M. The dynamics of bureaucracy. Chicago: University of Chicago Press, 1955.

Blau, Peter M. & Scott, W. Richard. Formal organizations. San Francisco: Chandler, 1962.

Brunswick, E., Hull, C.L. & Lewin, K. Symposium on psychology and scientific method. Psychological Review, 1943, 50, 255-310.

Canter, David. An intergroup comparison of connotative dimensions in architecture. Environment and Behavior, 1969, 1, 37-48.

Cohen, Kalman J., Dill, W.R., Kuehn, A.A., & Winters, Peter R. The Carnegie Tech management game. Homewood, Illinois: Richard D. Irwin, 1964.

Cohen, Kalman J. Food and its vicissitudes. Unpublished ms. 1955.

Collins, J.B. Perceptual dimensions of architectural space validated against behavioral criteria, Salt Lake City, Utah: University of Utah Department of Psychology unpublished Ph.D. thesis, 1969.

Collins, J.B. & Cockram. Lighting of hospital wards. Building Research Station News, 1970 (Autumn), 14, 12-15.

Collins, J.B. & Seaton, R.W. Semantic dimensions as architectural discriminators, paper read at Western Psychological Association Annual Convention, San Francisco, 24 April, 1970.

Craik, K.H. The comprehension of the everyday physical environment. Journal of the American Institute of Planners, 1968 (January) 34 (1), 29-37.

Craik, K.H. Personal communication, September 1967.

Evans, B. Architecture and research. Journal of the American Institute of Architects, 1966 (June), 58-59.

Fein, Leonard J. Ideology and architecture: dilemmas of pluralism in planning, in S. Andersen (Ed.), Planning for diversity and choice, Cambridge, Mass: The MIT Press, 1968.

Festinger, L., Riecken, H.W. & Schacter, S. When prophecy fails. Minneapolis, University of Minnesota Press, 1956.

Fisher, R.A. & Yates, F. Statistical tables for biological, equcultural and medical research. New York: Hafner, 1957.

Galvin, Franklin J. The design of an experimental simulation instrument to be used as a tool for correlating architectural space with psychological effectuation. Urbana, Ill.: University of Illinois Department of Architecture bachelor's thesis, 1970.

Gullahorn, J.T. & Gullahorn, J.E. Some computer applications in social science, American Sociological Review, 1965 (June), 36, 353-365.

Hershberger, R.G. A study of meaning and architecture, Philadelphia, Penn.: University of Pennsylvania Department of Architecture unpublished Ph.D. dissertation, 1968.

Holmberg, L., Almgres, S., Soderpalm, A.C. & Duller, R. The perception of volume content of rectangular rooms: comparison between models and full-scale experiments. Psychological Res. Bulletin, 1967, 7 (9), Lund University.

Janiskee, Robert L. Unpublished paper, Department of Geography, University of Illinois, Urbana, 1971.

Kerckhoff, A. & Back, K. June bug: a study of hysterical contagion. New York: Appleton-century - crafts, 1968.

Langdon, F.J. Human factors in environmental design. Building Research Station News, 1970 (Autumn), 14, Cover, inside cover and 7-8.

Lauf, J.J.H. Differences between full-size and scale-model rooms in the assessment of lighting quality. In D.V. Canter (Ed.) Architectural Psychology, London: RIBA Publications 1970, 43-48.

Lowe, J.B. The appraisal of design, RIBA Journal, 1959, 76, 379-380.

Lowenthal, D. An analysis of environmental perception. New York: American Geographical Society, 1967.

McGrath, J.E. & Altman, I. Small group research. New York: Holt, Rinehart & Winston, 1966.

Michelson, W. An empirical analysis of urban environmental references. Journal of the American Institute of Planners, 1966, 32, 358-360.

Murdock, George P. Social Structure. New York: MacMillan, 1949.

Peterson, J.M., Woodman, D., & Eaton, R. Critical judgments based on direct vs indirect experience: photos vs reality, DMS Newsletter, 1968 (April), 2 (4), 5 (abstract).

Ritter, P. & Hibb, Ralph. A method of color cinematography of design models through a modelscope in architecture, planning and other fields, Architectural Science Review, 1969 (March), 12, 78-84.

Sanoff, Henry. Low income housing demonstration. Berkeley, California: University of California Department of Architecture Research Office, 1965.

Seaton, R.W. Architectural simulation -- a mini-bib. Council of Planning Librarians, Exchange Bibliography #200, 1971.

Selznick, P. TVA and the grass roots. Berkeley, California: University of California Press, 1949.

SER (School Environments Research Project), SER 1: Environmental abstracts. SER 2: Environmental evaluations. SER 3: Environmental analysis. Ann Arbor, Michigan Architectural Research Laboratory, 1965.

Scheffe, Henry. The analysis of variance. New York: Wiley, 1959.

Sherif, Mustafer, Harvey, O.H. et al. Intergroup conflict and cooperation: The Robber's Cave experiment. Norman, Oklahoma: University of Oklahoma Press, 1961.

Tyhurst, J.S. Psychological and social aspects of civilian disaster. Canadian Medical Association Journal, 1957, 76, 385ff.

Vielhauer, Joyce. Development of a semantic scale for the description of the physical environment. Unpublished doctoral dissertation, Louisiana State University (Baton Rouge), 1965.

Winkel, G.H. & Sasanoff, R. An approach to an objective analysis of behavior in architectural space. In Proshansky, H.M., Ittelson, W.H., & Rivlin, L.G., (Eds.) Environmental Psychology, New York: Holt, Rinehart & Winston, 1970, 619-631.