Calibration of subjective scaling responses
Rea, M. S.

Publisher's version / Version de l'éditeur:

Lighting Research and Technology, 14, 3, pp. 121-229, 1982

National Research Council Canada
Conseil national de recherches Canada

CALIBRATION OF SUBJECTIVE SCALING RESPONSES

by M.S. Rea

Reprinted from
Lighting Research & Technology
Vol. 14, No. 3, 1982
p. 121 - 129

DBR Paper No. 1076
Division of Building Research

Price $1.00 OTTAWA NRCC 20869

Canada

RÉSUMÉ

On utilise de plus en plus l'échelle d'évaluation subjective pour déterminer les caractéristiques de l'éclairage des intérieurs, bien que peu d'efforts aient été faits pour valider les réponses des sujets. On soutient que les chercheurs qui utilisent l'échelle d'évaluation subjective devraient étalonner d'une façon ou d'une autre leurs sujets, les idiosyncrasies pouvant modifier les résultats moyens.

Dans la présente expérience, la tâche est soumise à un contraste variable, et la performance visuelle (vitesse et précision) est évaluée en même temps que le sujet répond subjectivement suivant une échelle sémantique en sept points. Les performances des sujets soumis à l'expérience sont stéréotypées ; leurs réponses dénotent des idiosyncrasies, même si certaines évaluations subjectives des paramètres expérimentaux varient dans le même sens que les performances. On avance que des mesures de performances visuelles peuvent être utilisées pour étalonner des évaluations subjectives.

Summary Subjective scaling is being used more frequently to assess the lighting characteristics of interiors, although there has been little attempt to validate the responses that subjects provide. It is argued that investigators using subjective scaling should calibrate their subjects in some way because idiosyncratic responses can alter averaged scaling estimates. In this experiment task contrast was varied and estimates of visual performance (speed and accuracy) were collected simultaneously with subjective responses on seven-point semantic scales. Subjects exhibited sterotypical behaviour on the performance measures; their scaling responses were quite idiosyncratic, although some subjective responses to the experimental parameters varied like the performance responses. It is argued that visual performance measures could be used to calibrate subjective scaling responses.

Calibration of subjective scaling responses

M. S. REA

1 Introduction

Subjective scaling has become a widely used technique in measuring the properties of objects or environments, for example such attributes as the brightness of tasks[1] or the spaciousness of rooms[2]. In most other areas where measurement is important a great deal of attention is placed upon standardising the units and tools for assessing object attributes, but with subjective scaling there has been little, if any, discussion about calibrating the responses of observers. It would appear that in most contemporary studies face validity alone dictates the acceptance of subjective scaling data in 'quantifying' the properties of objects and environments.

It is fairly obvious that an observer's responses can be influenced by a variety of parameters, some of them superfluous to the task at hand. For example, having a fight with one's spouse might influence the behaviour of a subject in a scaling experiment, yet it is unlikely that this factor is of interest to the investigator trying to define a perceived brightness function. Without suitable checks (i.e. calibration) an investigator cannot validate the subject's response. If data from a subject who cannot or will not attend to brightness are included with those of other subjects, it can lead to erroneous interpretation of how people perceive brightness.

The calibration of a subject in a scaling experiment can be much like the calibration of a photocell, where a systematic relation must be found between response under unique situations and response under a standard set of conditions. Thus, when subsequent measurements are made with the cell, its response can be placed in some orthodox context. As with a photocell, a person's responses in subjective scaling could be calibrated if they could be related to responses under a standard set of conditions.

Visual performance measures based upon speed and accuracy have been used to indicate the response of the human perceptual system to complex, realistic stimuli.[3-7] As such, these response functions may be regarded as a basis for gauging visual processing by human subjects (for example, work by Weston[3] has been used to recommend illumination levels in Great Britain). Subjective scaling responses to visual parameters could be calibrated if they could be correlated with visual performance data under comparable conditions. It is not mandatory that the responses in subjective scaling be linked only to those from visual performance experiments, but without a calibration of this type to some kind of measure it is impossible to decide unequivocally whether the subjective scaling responses are produced from the variable important to the investigator or from extraneous parameters.

This paper deals with the calibration of observers' responses in a subjective scaling experiment. Responses of the same observers under the same stimulus conditions in visual performance and scaling experiments are compared. These comparisons illustrate some of the problems associated with using subjective scaling when observers' responses are not calibrated.

2 Procedures

Subjective scaling data were collected in conjunction with a visual performance experiment. Briefly, the subject's task in the performance experiment[7] was to compare two number lists as

Dr. Rea is with the Energy and Services Section, Division of Building Research, National Research Council of Canada.

The paper was first received on 3 November 1981, and in revised form on 1 March 1982.

POST BLOCK RATING FORM

EXPERIMENT ________________ NAME ________________

SUBJECT ________________ DATE ________________

SESSION ________________

BLOCK ________________

FEELINGS: INDICATE HOW YOU FELT DURING THE PRECEDING BLOCK OF TRIALS

F1	BAD	:___:___:___:___:___:___:___:	GOOD
F2	TENSE	:___:___:___:___:___:___:___:	RELAXED
F3	SLEEPY	:___:___:___:___:___:___:___:	ALERT
F4	TIRED	:___:___:___:___:___:___:___:	RESTED
F5	UNCOMFORTABLE	:___:___:___:___:___:___:___:	COMFORTABLE
F6	DISCOURAGED	:___:___:___:___:___:___:___:	SATISFIED

TASK EVALUATION: INDICATE YOUR OVER-ALL EVALUATION OF THE TASK AND ENVIRONMENT FOR THE PRECEDING BLOCK OF TRIALS

TE1	BAD	:___:___:___:___:___:___:___:	GOOD
TE2	DIFFICULT	:___:___:___:___:___:___:___:	EASY
TE3	UNPLEASANT	:___:___:___:___:___:___:___:	PLEASANT
TE4	SUBDUING	:___:___:___:___:___:___:___:	STIMULATING
TE5	TIRING	:___:___:___:___:___:___:___:	RELAXING
TE6	DIM	:___:___:___:___:___:___:___:	BRIGHT
TE7	HAZY	:___:___:___:___:___:___:___:	CLEAR
TE8	FRUSTRATING	:___:___:___:___:___:___:___:	SATISFYING

COMMENTS:

Fig. 1. Post block rating form.

quickly and accurately as possible. The contrast of the stimulus materials was changed before each trial by realistic techniques (see below). Illumination was provided by a single luminaire (aperture 95×95 cm) and kept constant at the centre of the subject's task for the entire experiment. A performance score was calculated from both speed and accuracy measures.

Two sets of 7-point semantic scales similar to those employed by Flynn *et al*[8], Hawkes *et al*[9] and McKennan *et al*[10] were used in the scaling experiment. One set of six scales dealt with how the subject felt during the preceding "half-block" of performance trials; the other set of eight scales dealt with the subject's evaluation of the task and environment for the same half-block (Fig. 1). The two scaling sets were identical to those employed by Smith and Rea[1] in a similar visual performance experiment. Dividing the scales into two distinct sets was designed to help subjects more clearly differentiate their evaluation of the task from assessment of personal feelings. All scales were intended to have negative valence terms on the left and positive valence terms on the right; the better the perceived value of the attribute the higher the scale value.

Each half-block change corresponded to a change in the illumination angle or degree of polarisation of illumination. Two illumination angles were produced by rotation of the subject's work desk about the centre of the task (0 degree veiling reflection angle and 90 degree illumination from the subject's left). Three degrees of vertically polarised illumination were produced by different luminaire panels (Plexiglas and mylar, PM, approximately 4 per cent; multilayer polariser, MP, approximately 20 per cent; linear dichroic polariser, LP, approximately 100 per cent). Two illumination angles and three degrees of polarisation produced six unique combinations; each was assigned to a half-block. Four randomly distributed examples of each type of reference sheet (black matte, black gloss, grey matte and grey gloss) were presented in a half-block. Each of the six subjects gave scale evaluations (Fig. 1) immediately following a half-block and before changing the illumination angle and degree of polarisation for the next half-block.

Subjects did the task and the scalings during twenty-four half-blocks; six half-blocks were completed in each of four sessions. One morning and one afternoon session were conducted on two consecutive days. All experimental conditions were presented in a counterbalanced, randomised design. Further details regarding stimuli, observers and experimental protocol may be obtained in Ref. 7.

3 Results

The basic data are given in Tables 1 and 2; entries are means. Those in Table 1 are based upon responses for the six half-blocks, and those in Table

Table 1. Subjective scaling results; half-blocks

		FEELINGS							TASK EVALUATION									
	SUBJECT	F1	F2	F3	F4	F5	F6	MEAN	TE1	TE2	TE3	TE4	TE5	TE6	TE7	TE8	MEAN	GRAND MEAN
0 DEGREE PLEXIGLASS & MYLAR	S1	4.75	5.00	4.75	4.50	5.00	5.00	4.83	4.75	6.50	5.00	5.00	5.00	5.25	5.75	5.00	5.28	5.09
	S2	4.00	4.00	4.00	3.75	4.00	4.00	3.96	4.00	4.00	4.00	4.00	3.75	4.00	4.00	4.00	3.97	3.96
	S3	6.75	6.25	6.75	6.50	6.25	6.00	6.42	7.00	6.00	6.00	6.00	6.00	7.00	5.25	5.50	6.09	6.23
	S4	5.50	6.00	4.50	3.25	4.50	4.50	4.71	5.75	5.25	4.50	3.00	3.50	5.25	3.25	4.75	4.41	4.54
	S5	5.00	5.50	4.50	5.00	5.00	5.25	5.04	4.75	4.75	5.50	4.25	4.75	3.75	3.50	4.75	4.50	4.73
	S6	5.00	5.00	4.75	4.75	4.75	4.50	4.79	4.50	3.75	4.00	4.00	4.00	5.00	4.75	4.00	4.25	4.48
MEAN		5.17	5.29	4.87	4.62	4.92	4.87	4.96	5.12	5.04	4.83	4.37	4.50	5.04	4.42	4.67	4.75	4.84
0 DEGREE MULTILAYER POLAROID	S1	4.75	4.75	4.25	4.75	5.00	5.00	4.75	4.75	6.00	5.00	4.50	5.00	5.00	5.50	4.75	5.06	4.93
	S2	4.00	4.00	4.00	3.75	4.00	4.00	3.96	4.00	4.00	4.00	4.00	3.75	4.00	4.00	4.00	3.97	3.96
	S3	7.00	6.50	6.50	6.50	6.00	6.00	6.42	6.25	6.00	6.00	6.00	6.25	7.00	5.25	5.50	6.03	6.20
	S4	3.75	4.25	2.50	2.50	2.75	5.00	3.46	5.00	5.00	2.75	2.50	2.50	6.00	3.75	4.75	4.03	3.79
	S5	5.50	5.50	5.75	6.00	5.25	6.00	5.67	5.00	4.75	5.50	5.25	4.75	5.25	3.75	5.75	5.00	5.29
	S6	5.25	5.00	4.75	4.75	5.00	4.75	4.92	4.50	3.50	4.00	4.00	4.00	4.75	4.75	4.00	4.19	4.50
MEAN		5.04	5.00	4.62	4.71	4.67	5.12	4.86	4.92	4.87	4.54	4.37	4.37	5.33	4.50	4.79	4.71	4.78
0 DEGREE LINEAR POLAROID	S1	4.75	4.75	4.25	4.25	5.00	4.75	4.62	5.00	6.00	5.00	4.75	5.00	5.00	5.50	5.00	5.16	4.93
	S2	4.00	4.00	4.00	3.75	4.00	4.00	3.96	4.00	4.00	4.00	4.00	3.75	4.00	4.00	4.00	3.97	3.96
	S3	6.50	6.25	6.00	6.50	6.50	6.00	6.29	6.75	6.25	6.50	6.00	6.50	6.25	5.75	6.00	6.25	6.27
	S4	3.75	4.00	2.00	1.75	4.25	4.50	3.37	4.75	5.00	3.25	2.25	2.00	6.00	2.75	4.25	3.78	3.61
	S5	6.00	6.00	6.50	7.00	6.50	6.25	6.37	6.00	6.25	6.25	6.00	5.75	5.25	6.00	6.00	5.94	6.12
	S6	5.00	5.00	4.75	4.75	5.00	4.50	4.83	4.50	4.00	4.00	4.00	4.00	5.00	4.75	4.00	4.28	4.52
MEAN		5.00	5.00	4.58	4.67	5.21	5.00	4.91	5.17	5.25	4.83	4.50	4.50	5.25	4.79	4.87	4.90	4.90
90 DEGREE PLEXIGLASS & MYLAR	S1	5.25	5.25	5.00	4.75	5.25	5.00	5.08	5.25	6.75	5.00	5.00	5.00	5.25	5.75	5.00	5.37	5.25
	S2	4.00	4.00	4.00	3.50	4.00	4.00	3.92	4.00	4.00	4.00	4.00	3.50	4.00	4.00	4.00	3.94	3.93
	S3	7.00	6.75	6.50	6.50	6.25	7.00	6.67	7.00	7.00	7.00	6.00	6.50	6.00	7.00	7.00	6.69	6.68
	S4	4.25	5.75	3.50	2.00	3.75	5.75	4.17	5.25	5.75	3.75	3.75	2.00	6.00	3.00	5.00	4.31	4.25
	S5	5.75	6.25	5.25	5.50	5.75	6.25	5.79	6.25	6.25	6.25	5.00	5.50	6.25	6.50	6.00	6.00	5.91
	S6	5.00	5.00	4.75	4.75	5.00	4.25	4.79	5.00	4.75	4.00	4.00	4.00	4.50	4.75	4.00	4.38	4.55
MEAN		5.21	5.50	4.83	4.50	5.00	5.37	5.07	5.46	5.75	5.00	4.62	4.42	5.33	5.17	5.17	5.11	5.10
90 DEGREE MULTILAYER POLAROID	S1	5.25	5.00	4.75	4.75	5.00	5.00	4.96	5.25	7.00	5.00	5.00	5.00	5.50	6.25	4.75	5.47	5.25
	S2	4.00	4.00	4.00	3.00	4.00	4.00	3.83	4.00	4.00	4.00	4.00	3.00	4.00	4.00	4.00	3.88	3.86
	S3	7.00	6.75	7.00	6.25	6.50	7.00	6.75	7.00	7.00	6.50	6.00	6.50	6.00	7.00	7.00	6.63	6.68
	S4	2.75	4.00	3.50	2.25	2.75	5.50	3.46	5.25	4.75	3.00	2.25	2.00	6.00	2.50	5.25	3.87	3.70
	S5	6.00	6.50	5.75	6.25	6.00	6.00	6.08	6.00	6.50	6.00	5.25	5.75	6.00	6.25	6.00	5.97	6.02
	S6	5.00	5.00	5.00	4.75	5.00	5.00	4.96	5.00	5.00	4.25	4.00	4.00	4.25	4.75	4.00	4.41	4.64
MEAN		5.00	5.21	5.00	4.54	4.87	5.42	5.01	5.42	5.71	4.79	4.42	4.37	5.29	5.12	5.17	5.04	5.02
90 DEGREE LINEAR POLAROID	S1	5.25	5.25	4.75	4.75	5.00	5.00	5.00	5.25	7.00	5.00	4.75	5.00	5.50	6.00	5.00	5.44	5.25
	S2	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
	S3	7.00	6.50	6.50	6.50	6.50	7.00	6.67	7.00	7.00	6.75	6.00	6.50	6.00	7.00	7.00	6.66	6.66
	S4	3.25	4.00	1.75	1.75	3.75	5.00	3.25	4.75	4.50	3.25	1.75	1.50	6.25	4.00	4.75	3.84	3.59
	S5	6.25	6.25	6.50	6.50	5.75	6.50	6.29	6.00	6.50	6.25	5.25	5.50	5.00	5.50	5.50	5.69	5.95
	S6	5.00	5.00	4.75	4.50	5.00	5.00	4.87	5.00	4.75	4.00	4.00	4.00	4.25	4.75	4.00	4.34	4.57
MEAN		5.12	5.17	4.71	4.67	5.00	5.42	5.01	5.33	5.62	4.87	4.29	4.42	5.17	5.21	5.04	4.99	5.00
OVERALL MEAN		5.09	5.19	4.77	4.62	4.94	5.20	4.97	5.24	5.37	4.81	4.43	4.43	5.24	4.87	4.95	4.92	4.94

2 are based upon responses for the four experimental sessions. F1 to F6 and TE1 to TE8 correspond to each 'feeling' scale and each 'task evaluation' scale, respectively, numbered from first to last in the two sets (Fig. 1).

Typically (and in this experiment as well), numerical values are given to stimuli (e.g. by a tick mark on a pre-drawn scale) by several subjects, using a variety of scales. One usually tries to simplify matters by averaging across subjects and scales under the assumption that these variables are replicates[11]. To test the replication assumptions the data averages were broken down by subject, by scale, and by subject and scale. The averaged data and each breakdown are presented in turn.

3.1 Averaged data

The averaged visual performance data are shown in Fig. 2, with averaged subjective scaling data for the six half-blocks (two orientations by three types of luminaire panels). Scaling responses were averaged across all subjects but separately across the scales for task evaluations and feelings. There was significant correlation ($r = 0.955$, $p < 0.01$)* between average visual performance scores and average task evaluation values, but there was not a significant correlation between performance scores and feeling values. One might conclude that, on average, subjects' task evaluation responses were based upon the same parameters as those influencing performance but that their feeling responses were not.

Fig. 3 shows a similar plot of the visual performance data with scaling data for the four experimental sessions. In this case there was not a significant correlation between performance scores and either task evaluation or feeling values.

One might conclude that, again on average, subjects' average task evaluation and average feeling responses were not influenced by the same factors that influenced average performance in the sessions. One can see from Fig. 3, however, that there is a clear correlation between feeling and task evaluation values ($r = 0.953$, $p < 0.01$), indicating that session effects influence subjects, on the average, in the same way for feeling and task evaluation responses. This perhaps indicates a lack of differentiation between feelings and task evaluations by subjects across sessions.

3.2 Breakdown by subject

Subjects' performance data obtained during the six half-blocks are more homogeneous than their scaling data under the same conditions (Fig. 4).

*Correlations were obtained between paired performance score and scaling means across half-blocks or across sessions. Significance tests were based upon the students' 't' distribution and N-2 degrees of freedom (4 df for half-blocks, and 2 df for sessions).

Table 2. Subjective scaling results; sessions

	SUBJECT	FEELINGS F1	F2	F3	F4	F5	F6	MEAN	TASK EVALUATION TE1	TE2	TE3	TE4	TE5	TE6	TE7	TE8	MEAN	GRAND MEAN
SESSION 1	S1	5.00	5.00	4.00	4.17	5.00	4.83	4.67	5.00	7.00	5.00	4.50	5.00	5.00	5.83	5.00	5.29	5.02
	S2	4.00	4.00	4.00	3.67	4.00	4.00	3.94	4.00	4.00	4.00	4.00	3.67	4.00	4.00	4.00	3.96	3.95
	S3	6.83	6.17	6.17	6.00	5.83	6.67	6.28	6.67	6.67	6.17	6.00	6.17	6.33	6.33	6.50	6.35	6.32
	S4	4.17	5.00	3.83	2.17	4.17	5.17	4.08	5.33	5.33	4.00	3.00	2.17	5.67	2.33	4.83	4.08	4.08
	S5	6.00	6.33	6.00	6.33	5.67	6.17	6.08	5.67	6.33	6.00	5.33	5.67	5.67	5.17	5.83	5.71	5.87
	S6	5.00	5.00	4.17	4.00	5.00	4.67	4.64	5.00	4.50	4.17	4.00	4.00	4.33	4.00	4.00	4.25	4.42
MEAN		5.17	5.25	4.69	4.39	4.94	5.25	4.95	5.28	5.64	4.89	4.47	4.44	5.17	4.61	5.03	4.94	4.94
SESSION 2	S1	4.50	4.67	4.50	4.33	4.83	5.00	4.64	4.67	6.17	5.00	5.00	5.00	5.00	6.00	4.67	5.19	4.95
	S2	4.00	4.00	4.00	3.67	4.00	4.00	3.94	4.00	4.00	4.00	4.00	3.67	4.00	4.00	4.00	3.96	3.95
	S3	7.00	6.17	6.50	6.00	6.00	6.33	6.33	7.00	6.33	6.67	6.00	6.00	6.33	6.17	6.17	6.33	6.33
	S4	4.17	4.50	3.83	3.50	3.83	5.00	4.14	5.00	5.00	3.50	3.67	2.83	6.17	4.67	4.83	4.46	4.32
	S5	5.00	5.00	5.17	5.00	4.67	5.50	5.06	4.67	4.83	5.33	4.33	4.00	3.67	4.50	5.00	4.54	4.76
	S6	5.17	5.00	5.00	4.83	4.83	4.67	4.92	4.67	4.17	4.00	4.00	4.00	4.83	5.00	4.00	4.33	4.58
MEAN		4.97	4.89	4.83	4.56	4.69	5.08	4.84	5.00	5.08	4.75	4.50	4.25	5.00	5.06	4.78	4.80	4.82
SESSION 3	S1	5.00	5.00	4.67	4.83	5.17	5.00	4.94	5.00	6.50	5.00	5.00	5.00	5.50	5.67	5.00	5.33	5.17
	S2	4.00	4.00	4.00	3.83	4.00	4.00	3.97	4.00	4.00	4.00	4.00	3.83	4.00	4.00	4.00	3.98	3.98
	S3	7.00	7.00	6.67	6.83	6.67	6.67	6.81	6.83	6.67	6.67	6.00	6.67	6.33	6.17	6.67	6.50	6.63
	S4	2.50	5.83	1.17	1.33	3.17	4.83	3.14	4.67	4.83	3.17	1.67	1.67	6.00	2.50	4.33	3.60	3.40
	S5	6.00	6.17	5.83	6.67	6.67	6.00	6.22	6.00	6.17	6.17	5.50	5.83	5.50	5.83	6.00	5.87	6.02
	S6	5.00	5.00	5.00	5.00	5.00	4.67	4.94	4.50	4.17	4.00	4.00	4.00	4.50	5.00	4.00	4.27	4.56
MEAN		4.92	5.50	4.56	4.75	5.11	5.19	5.00	5.17	5.39	4.83	4.36	4.50	5.31	4.86	5.00	4.93	4.96
SESSION 4	S1	5.50	5.33	5.33	5.17	5.17	5.00	5.25	5.50	6.50	5.00	4.83	5.00	5.50	5.67	5.00	5.37	5.32
	S2	4.00	4.00	4.00	3.33	4.00	4.00	3.89	4.00	4.00	4.00	4.00	3.33	4.00	4.00	4.00	3.92	3.90
	S3	6.67	6.67	6.83	7.00	6.83	6.33	6.72	6.83	6.50	6.33	6.00	6.67	6.50	6.17	6.00	6.37	6.52
	S4	4.67	3.33	3.00	2.00	3.33	5.17	3.58	5.50	5.00	3.00	2.00	2.33	5.83	3.33	5.17	4.02	3.83
	S5	6.00	6.50	5.83	6.17	5.83	6.50	6.14	6.33	6.00	6.33	5.50	5.83	6.17	5.50	5.83	5.94	6.02
	S6	5.00	5.00	5.00	5.00	5.00	4.67	4.94	4.83	4.33	4.00	4.00	4.00	4.83	5.00	4.00	4.37	4.62
MEAN		5.31	5.14	5.00	4.78	5.03	5.28	5.09	5.50	5.39	4.78	4.39	4.53	5.47	4.94	5.00	5.00	5.04
OVERALL MEAN		5.09	5.19	4.77	4.62	4.94	5.20	4.97	5.24	5.37	4.81	4.43	4.43	5.24	4.87	4.95	4.92	4.94

Only the absolute level of visual performance for subject No. 4 was substantially different from that of the other five subjects; his *relative* visual performance was similar to theirs (see below). It would appear that subjects are quite different in their scaling responses yet similar in their visual performance responses. Further, only subject No 3 had scaling responses similar to *her* visual performance responses ($r = 0.982$, $p < 0.001$) and to the visual performance responses averaged across *all* subjects ($r = 0.974$, $p < 0.001$). Because only one subject's task evaluation scale responses correlated with performance, one might conclude that the significant correlation mentioned above between average performance and average task evaluation scalings was fortuitous. This point is amplified in Fig. 5. Deleting an individual subject's data can seriously affect the shape of the scaling curves and thus the conclusions one might reach about all subjects' subjective impressions of the

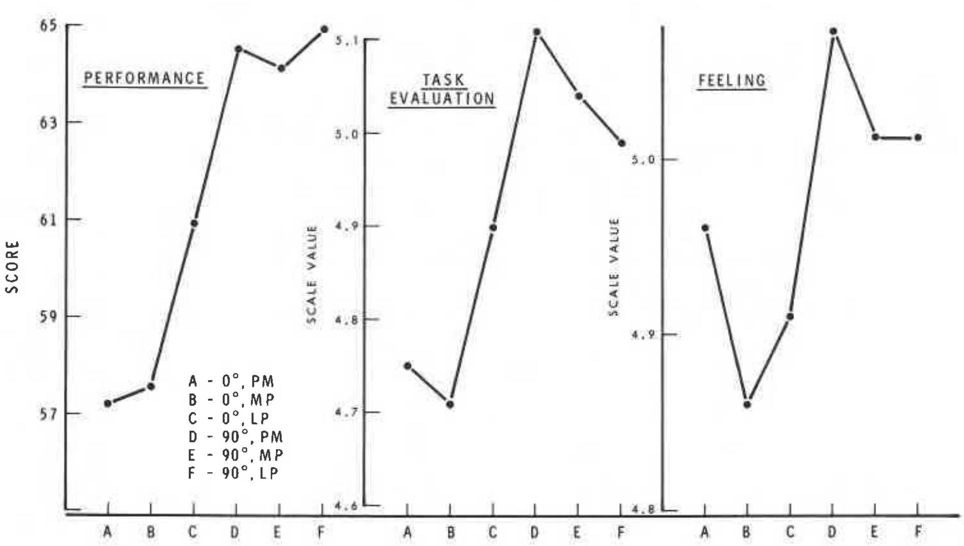

Fig. 2. Visual performance and subjective scaling data for the six experimental half-blocks averaged across subjects and scales.

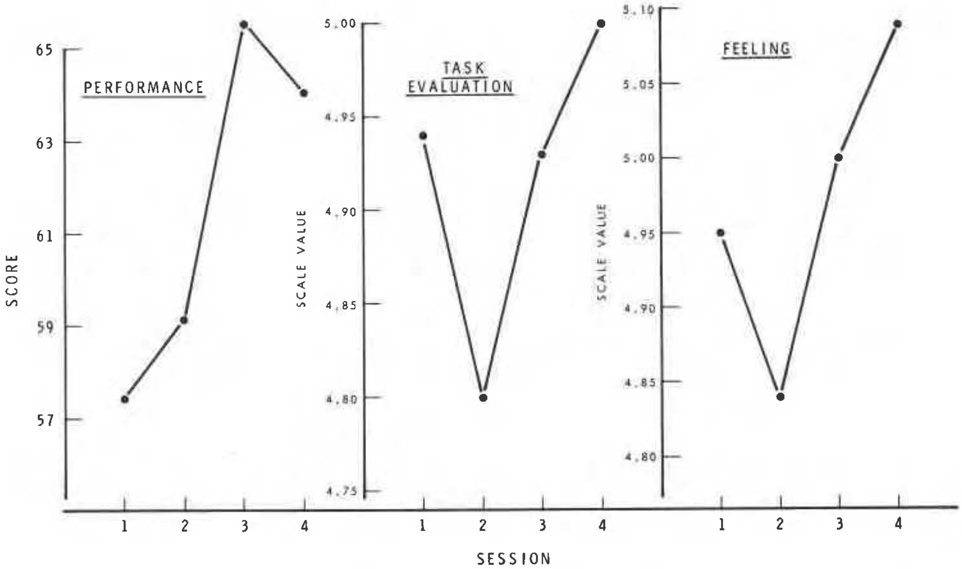

Fig. 3. Visual performance and subjective scaling data for the four experimental sessions averaged across subjects and scales.

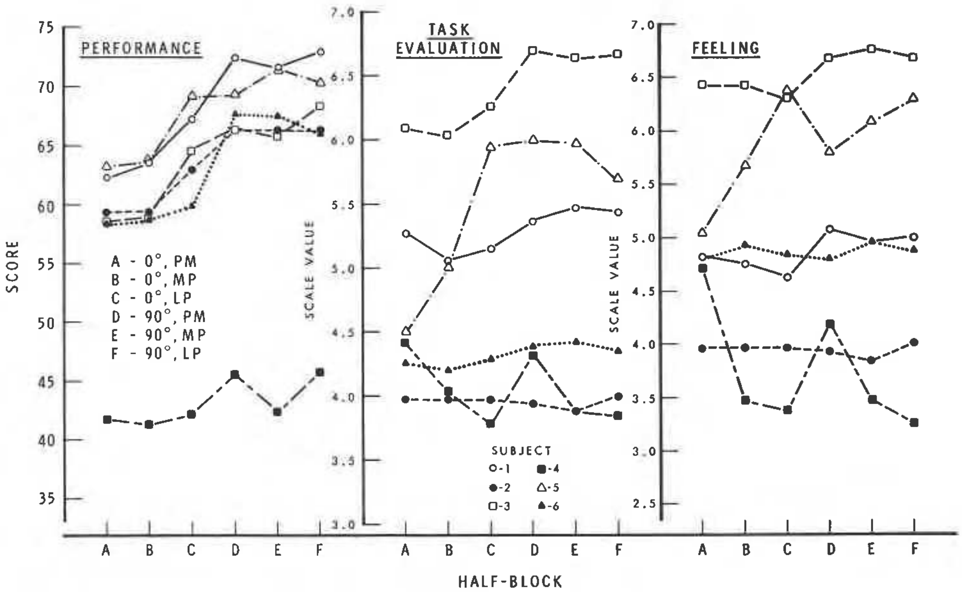

Fig. 4. Visual performance and subjective scaling data for the six experimental half-blocks averaged across scales but not across subjects.

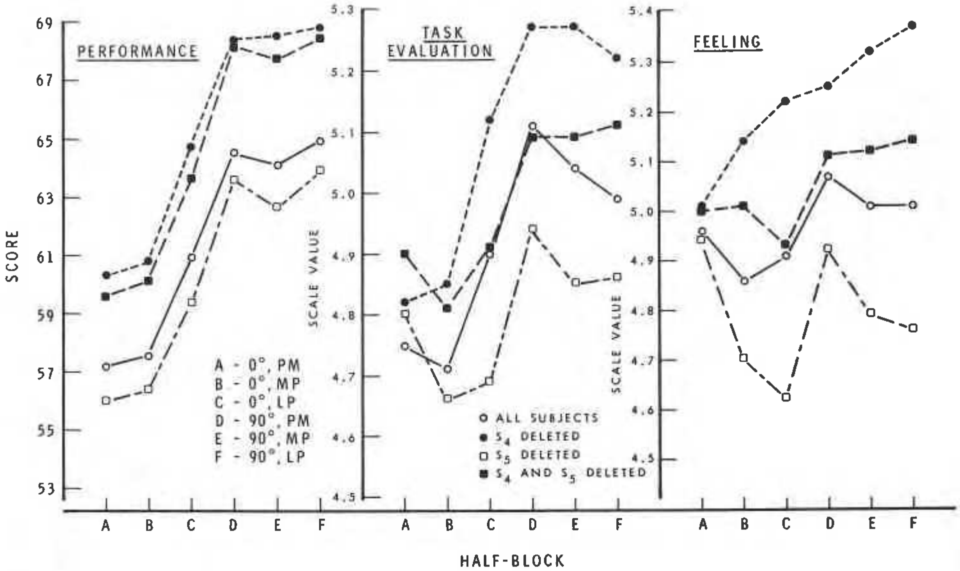

Fig. 5. Deletion of some subjects' data from the averaged data in Fig. 2.

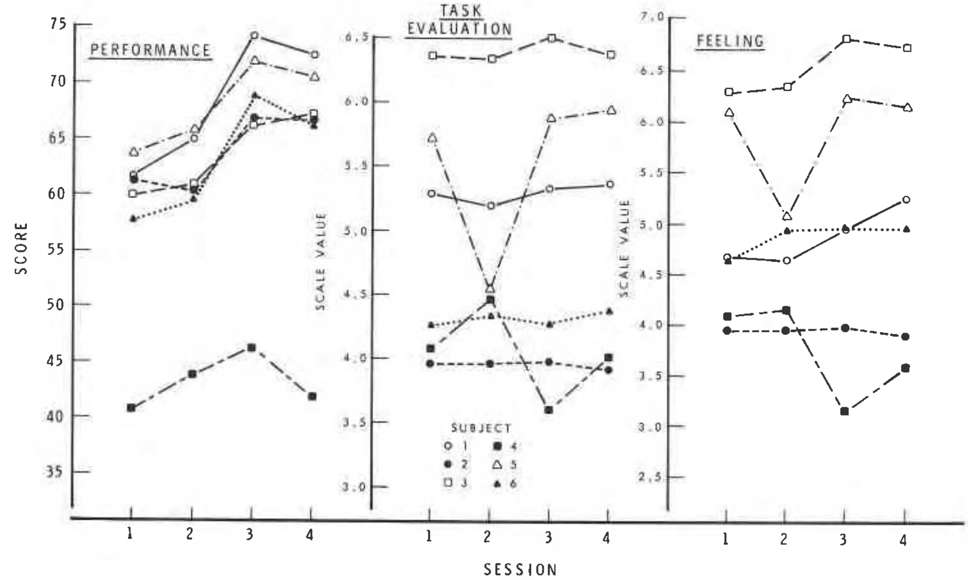

Fig. 6. Visual performance and subjective scaling data for the four experimental sessions averaged across scales but not across subjects.

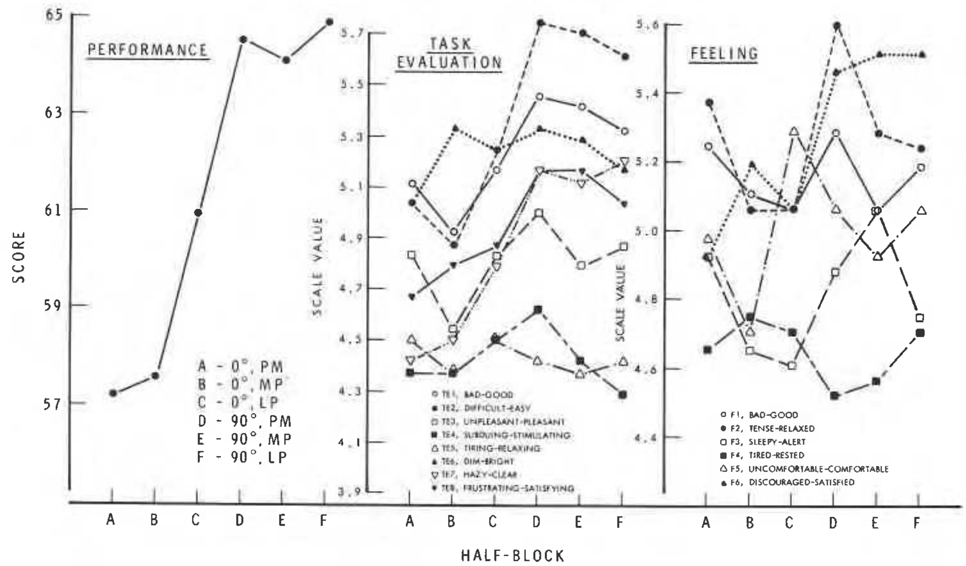

Fig. 7. Visual performance and subjective scaling data for the six experimental half-blocks averaged across subjects but not across scales.

experimental conditions. This is in contrast with deletion of the performance data for the same subjects; deletion shows little effect on the relative performance curves.*

Results similar to those obtained across half-blocks were also obtained across sessions. Performance data for the six subjects were quite similar, although the subjective scaling data for both task evaluations and feelings were idiosyncratic (Fig. 6). This examination also leads one to doubt the validity of averaging across subjects' scaling data because of the marked dissimilarities across sessions. As with the half-block data, only subject No. 3 had a significant correlation between performance scores and scale values across sessions. In this case, however, it was the average feeling scale values that correlated well with *her* visual performance during the sessions ($r = 0.970$, $p < 0.05$) and with the average visual performance for *all* subjects during the four sessions ($r = 0.994$, $p < 0.01$).

In sum, then, averaging subjects' visual performance data for the six half-blocks and the four sessions seems appropriate because the relative performance data are so similar. On the other hand, averaging subjective scaling data across subjects seems inappropriate because their data are so idiosyncratic. Further, only one of six subjects (S3) had subjective scaling responses that seemed to be influenced by the same parameters affecting visual performance. (For the six half-blocks this subject's perceptions were reflected in the over-all *task evaluation* averages; for the four sessions her perceptions were reflected in the *feeling* scale averages.) Averaging across subjects' scaling responses can, therefore, be risky; unless

*The absolute number of data points deleted in Fig. 5 are different for performance and for scaling responses, but the proportion of data deleted is the same.

unperceptive or uncooperative subjects' responses are averaged out the idiosyncratic behaviour of these subjects can alter the scaling response curves and thus the inferences drawn about the factors influencing subjective responses.

3.3 Breakdown by scale

Examination of individual scales, averaged across subjects, provides a result similar to that drawn for averaging across subjects. Fig. 7 shows performance scores with task evaluation and feeling rating scales for the six half-blocks. Most scaling curves were not similar to the performance curve, but four task evaluation scales show a significant correlation with performance. Specifically, the scales TE1, 'bad-good' ($r = 0.874$, $p < 0.05$), TE2, 'difficult-easy' ($r = 0.946$, $p < 0.01$), TE7, 'hazy-clear' ($r = 0.994$, $p < 0.001$), TE8, 'frustrating-satisfying' ($r = 0.905$, $p < 0.02$) have shapes similar to those for performance. None of the feeling scale curves resembled those for performance across half-blocks.

Only one feeling scale, 'tired-rested' (F4), is similar to that for performance across sessions ($r = 0.952$, $p < 0.05$). No task evaluation scales were similar to performance scales across sessions. It would therefore be inappropriate to average across scales in many cases because scales, like subjects, do not always appear to be replications.

3.4 Breakdown by scale and subject

Table 3 shows the important relations between visual performance and scaling for each subject and each scale. Entries are the significant correlations ($p < 0.01$) between scaling responses and visual performance responses under the same experimental conditions. The differences among subjects and their use of the seven-point scales are revealed in this table. Subjects 2 and 4 gave no scale values that could confidently be correlated with performance. Yet there were marked dissimilarities in the behaviour of the two. Subject No. 2 marked the fourth interval on the scales almost exclusively throughout the experiment (Tables 1 and 2); on the other hand, subject No. 4, whose results also lacked any clear correlation between scaling and performance responses, was extremely erratic in his scaling responses. It is probably by chance alone that his responses on F6 'discouraged-satisfied', Table 3, were correlated significantly with his visual performance scores across sessions. Two facts support this hypothesis. First, no other scaling response from this subject was correlated with performance, even for those scales that were effectively used by the other subjects (e.g. TE1 and TE2, Table 3). Second, it is unlikely that the factors leading to better performance by this subject would also lead to more 'discouragement' on his part, as implied by the negative correlation between responses in F6 and those in visual performance.

Two subjects, S3 and S5, used the scales very effectively, based upon the number of scale responses that were correlated with performance under the same conditions. Subject No. 3 was probably the 'best' subject because she exhibited several significant correlations between visual performance and scaling responses across the six half-blocks and across the four sessions. Her scaling behaviour was so consistent for the half-blocks that it was revealed in her average scaling responses (Section 3.2). As well, this subject was the only one to exhibit scaling responses (F4 and F5) similar to performance responses across the four sessions.

Table 3. Correlations and significance probabilities between visual performance and each rating scale for each subject*.

	Scale	Correlation	Significance
HALF-BLOCKS			
Subject 1	bad−good (F1)	0.9301	0.004
	bad−good (TE1)	0.9911	0.001
Subject 3	discouraged−satisfied	0.9224	0.004
	difficult−easy	0.9766	0.001
	unpleasant−pleasant	0.9109	0.006
	dim−bright	−0.9816	0.001
	hazy−clear	0.9821	0.001
	frustrating−satisfying	0.9885	0.001
Subject 5	bad−good (F1)	0.8804	0.010
	tense−relaxed	0.9797	0.001
	bad−good (TE1)	0.9345	0.003
	difficult−easy	0.9885	0.001
	tiring−relaxing	0.9571	0.001
	hazy−clear	0.9312	0.003
Subject 6	bad−good (TE1)	0.9874	0.001
	difficult−easy	0.9719	0.001
	dim−bright	−0.8876	0.009
SESSIONS			
Subject 3	tired−rested	0.9944	0.003
	uncomfortable−comfortable	0.9988	0.001
Subject 4	discouraged−satisfied	−0.9845	0.008

*Only the probabilities of chance less than or equal to 0.01 are included.

Subject 5's responses to several scales also correlated well with performance, but unlike those of subject 3 they were not consistent enough to produce a correlation between her average scaling responses and her performance responses. It should be noted that subjects 3 and 5 did not always use the same scales in the same manner. This indicates that while they were basing their scaling responses on the same experimental parameters, as indicated by common correlations with performance, they were using different vehicles (i.e. scales) to express them.

Although subjects 1 and 6 used some of the scales effectively, fewer scales correlated with performance than did those from subjects 3 and 5. The former two subjects seemed to prefer (based upon a high level of significance) the 'bad-good' task evaluation scale (TE1) to scale the experimental parameter (contrast) associated with half-block changes (as had subject 5). Subjects 1 and 5 used the 'bad-good' feeling scale (F1) in the same way. Like the averaged scaling responses across sessions (Section 3.1) this may be another indication that these subjects did not clearly differentiate between task evaluations and personal feelings.

This hypothetical inability of some subjects to be objective in separating 'feelings' from 'task evaluations' in scaling responses could have serious consequences for those assuming that scaling responses are unbiased. This has been a topic of discussion for many years[12] and one that must be considered when the responses of subjects in scaling experiments are applied. In short, without proper calibration procedures one cannot be sure that subjects' responses are unbiased.

A special note should be made of the 'dim-bright' scale (TE6). Subjects 3 and 6 had a high *negative* correlation between responses on this scale and performance across the six half-blocks. Before conducting the experiment the author had assumed that the subjects' scaling responses would pertain to their perceptions of *background* brightness (which was kept constant throughout the experiment). This assumption was based upon subjects' responses in an earlier experiment[1] in which background luminance was changed experimentally. It appears from the results obtained in this experiment, however, that the two subjects were evaluating the experimentally manipulated *target-brightness.* Importantly, then, an investigator's *intended* meaning for scales like brightness, spaciousness or comfort may not be interpreted in the same way by the subjects. Flynn *et al*[8] pointed out that the investigator must sometimes rely upon his own 'ingenuity and background knowledge' to ascertain the bases of subjective scaling responses. This reliance upon one's own creativity and experience for interpretation is actually a form of *subjective scaling by the investigator* and therefore can be potentially biased. (The responses on the 'dim-bright' scale in this experiment are examples of this investigator's incomplete background knowledge). A calibration procedure reduces the subjectivity needed by the investigator and thus allows a more objective assessment of subjects' reasons for responding.

4 Conclusions

This experiment indicates that subjects use scales differently. Some used scales effectively, although several used different scales to reflect the same perceptions. One subject did not use the scales to indicate any change in experimental parameters, but almost exclusively marked one interval on all scales throughout the experiment. Another subject was so erratic in his responses that he seriously altered the shape of the average response curves. Finally, in evaluating the *same* aspect of a task, two subjects exhibited a negative correlation between visual performance and responses on one scale and a positive correlation between performance and responses on other scales. Averaging these idiosyncratic scaling responses eliminated many of the subtle yet important differences in scaling behaviour. It was therefore concluded that in many cases it is erroneous to assume, without checks (i.e. a calibration procedure), that subjects and scales can be considered replications.

5 Discussion

The major advantage of subjective scaling is that aspects of the environment not readily measurable by electrophysiological or psychophysical techniques yet potentially influential to the over-all effectiveness of a room (like pleasantness or spaciousness) can be evaluated. In this experiment subjects were required to make assessments of task characteristics associated with contrast (across the six half-blocks) or with learning and fatigue (across the four sessions). Perhaps these are less interesting requirements for subjects, but they can be calibrated against a performance measure.

Without a calibration of scaling responses it is not possible to ascertain unequivocally whether subjects are influenced by experimental parameters or by extraneous factors. It may be difficult to find an independent measure of, say, pleasantness or spaciousness that could be used to calibrate subjective scaling responses to these environmental characteristics. Nevertheless, another set of scaling responses obtained under standard conditions might be used to compare (i.e. calibrate) subjective scaling responses under new, untested situations.

At the very least, it seems prudent in light of the large individual differences obtained in this experiment to cull subjects who cannot or will not use subjective scales with less esoteric, yet testable, changes like photometric contrast. Although, it is possible that subjective responses calibrated in this manner may not reflect more complicated environmental aspects like pleasantness or spaciousness, it is much more likely that these subjects can make subjective assessments than can subjects who do not demonstrate sensitivity by any testable criterion.

Finally, it seems worth considering the establishment of standard calibration procedures for subjective scaling. Firstly, it would eliminate contamination of results by uncooperative, unperceptive or misunderstanding subjects. Secondly, it would facilitate comparisons of results obtained by other laboratories. The specific criteria chosen to calibrate subjects, whether with performance measurements or other means of scaling, need to be addressed by a large community interested in lighting design engineering. It seems wise, meanwhile, to avoid using subjects as scaling instruments without demonstrating first that they are sensitive to the variables manipulated by the investigator.

Acknowledgments

The author would like to thank B. Guzzo for running subjects, A. Hsu for computer assistance, and P. Boyce for his comments on the manuscript.

This project was funded jointly by the Illuminating Engineering Research Institute of North America and by the National Research Council of Canada, Division of Building Research.

References

1. Smith, S. W. and Rea, M. S., Presented at 19th CIE Session, Kyoto, Paper No. P-79-35 (1979).
2. Flynn, J. E. and Spencer, T. J., J. Illum. Engng. Soc. **6** (3), 167 (1977).
3. Weston, H. C., Industrial Health Research Board, HMSO, London, Report 87 (1945).
4. Bodmann, H. W., Int. Light. Rev., **13**, 41 (1962).
5. Boyce, P. R., Ltg. Res. & Technol., **5** (3), 125 (1973).
6. Smith, S. W. and Rea, M. S., J. Illum. Engng. Soc., **8** (1) 47 (1978).
7. Rea, M. S., J. Illum. Engng. Soc., **10** (93) (1981).
8. Flynn, J. E., Hendrick, C., Spencer, T. and Martyniuk, O., J. Illum. Engng. Soc., **8** (2), 95 (1979).
9. Hawkes, R. J., Loe, D. L. and Rowlands, E., J. Illum. Engng. Soc., **8** (2), 111 (1979).
10. McKennan, G. T., Parry, C. M. and Tilic, M., Paper BM presented at CIBS National Lighting Conference, Canterbury (1980).
11. Nunnally, J. C., Psychometric Theory, New York, McGraw-Hill (1967).
12. Cain, W. S. and Marks, L. E., Stimulus and Sensation, Boston, Little, Brown and Co. (1971).

Publications of the Division may be obtained by mailing the appropriate remittance (a Bank, Express, or Post Office Money Order, or a cheque, made payable to the Receiver General of Canada, credit NRC) to the National Research Council of Canada, Ottawa. K1A 0R6. Stamps are not acceptable.

A list of all publications of the Division is available and may be obtained from the Publications Section, Division of Building Research, National Research Council of Canada, Ottawa. K1A 0R6.