

August 13, 1963

STAT

[Redacted]

Frank

Dear Frank:

STAT

STAT

Here are a few comments and suggestions regarding [Redacted] [Redacted] study of the correlation between subjective judgments and objective measures of photographic image quality. I am very much in agreement with the ultimate goals of the work [Redacted] is doing, so please consider my comments and suggestions in light of my sympathetic attitude toward his work. In other words, if I sound excessively critical, attribute the sound to my interest and desire to encourage and help him--assuming that I can help.

STAT

1. The magnitude of any obtained correlation will be limited not only by the discriminability of the stimuli, i.e., the subjects' skill in distinguishing among them, but also by the reliability of the subjects' judgments. If, for example, the test-retest reliability, i.e., the correlation between successive sets of judgments made by the same subjects with a month or so intervening between judgmental sessions, turns out to be zero, no statistically significant correlation between them and objective, physical measures can be expected. If the test-retest reliability of the judgments as indicated by the correlation between successive sets of judgments turns out to be .5, the maximum correlation that can be obtained between the judgments and the objective measures of image quality is about .7, assuming the objective measures are perfectly reliable.

The point of this discussion is that an effort should be made to determine the test-retest reliability of the judgments of quality. The same subjects should perform the judgmental task twice with a month or so between judgmental sessions so

STAT

[REDACTED]
August 13, 1963

Page 2

that they will not recall their first judgments of specific stimuli. The obtained reliability coefficient will indicate the extent to which the correlation between subjective and objective measures is attenuated by a lack of reliability. Further, it will indicate the worth of subjective judgments as criteria of image quality. As I recall, Frank, you too wanted to see a reliability study done.

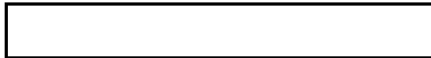
2. My second comment concerns the psychophysical procedure and the data analysis. The two cannot be considered entirely independently, for the psychophysical procedure used will limit the number of appropriate analysis techniques available. (I don't know how [REDACTED] intends to analyze the data or what kind of correlation he plans to compute, so to some extent I'm shooting in the dark here.)

STAT

As the procedure is currently designed, the subjects can use almost any number of judgmental categories in assessing photo image quality. And because of individual differences in discriminability and attitudes toward the task, you will find some subjects using 4 or 5 categories and others using 9 or 10 or possibly more categories. Consequently the resulting data cannot be combined too conveniently to arrive at a single regression equation and correlation coefficient based on the judgments of all subjects. Instead, as I see it, one equation and coefficient will have to be computed for each subject or for small groups of subjects that, fortuitously, ended up using the same number of judgment categories.

It would be more convenient from a statistical point of view if all subjects were required to sort the photos in a specific number of categories, say 7 or 9, ranging from "best" through "average" or "median" to "poorest" image quality. I am not aware of the number of physical steps used in generating the stimulus materials (or the approximate number of "just noticeable differences"--jnd's--from one end of the physical scale to the other, which could only be determined empirically) so I can't guess too well about the number of judgmental categories the subjects might be able to use reliably. But, in any case, if a specific number of judgment categories were used, the data from all subjects could be plotted as follows:

STAT




August 13, 1963
Page 3

Categorical
Judgments
of Image
Quality

7						1	5
6				1	1	4	1
5				2	3	1	
4			1	2	2		
3		2	3	1			
2	2	4	2				
1	4						
	1	2	3	4	5	6	7

Physical Measure of Quality
(in units of blur or graininess)

The total number of entries in the table would be the product of the number of subjects and the number of stimuli judged. From the table a coefficient of contingency could be computed. Or if the assumptions could be met (or reasonably ignored) a product-moment correlation could be computed.

(As I mentioned before, I don't know how  intends to analyze the data. He may be aware of some techniques that I'm not familiar with, or I may not understand his research goals. In either case, these comments might be entirely irrelevant.)

STAT

From a psychophysical point of view, the most sensitive judgmental technique in a study such as this is the method of pair comparisons. The subjects simply compare two stimuli at one time and respond by indicating which is poorer and which is better quality. Every stimulus is compared at least once with every other stimulus, and that's the limitation of the method. With n stimuli, there are $n(n-1) \div 2$ possible pairs, which can result in a burdensome judgmental task.

- I don't know precisely how the stimulus materials were prepared. I don't know whether blur and graininess were varied independently or simultaneously. If

STAT

 August 13, 1963

Page 4

they were varied independently, there might be, for example, 7 levels (magnitudes) of blur at the same level of graininess. There would then also be 7 levels of graininess at the same level of blur, yielding a minimum of 14 stimuli. On the other hand, if they were varied simultaneously, all levels of blur would occur at each level of graininess, and vice versa. With 7 levels of each variable, simultaneously varying them would produce a minimum of 49 stimuli.

This involved, and probably unclear, discussion has relevance, I think, for the manner in which the judgments of quality are obtained. If blur and graininess were varied independently, shouldn't they be treated independently in obtaining the judgments of quality? They are probably not equally related to judgments of quality; they are probably not equally difficult to discriminate; and, thus, judgments of quality as a function of blur, on the one hand, and graininess, on the other, are probably not equally reliable.

If blur and graininess were varied simultaneously in preparing the stimulus materials, an experimental design should be selected which would permit the experimenter to determine the relative contributions of each physical variable to judgments of quality. I would suggest an analysis of variance design, which implies the necessity of numerically scaled judgments of image quality rather than simple, ranked judgments, and further implies the necessity of developing the scale. Or, I would suggest a non-parametric analysis of variance such as that described by Garner and McGill.* The latter design is not as powerful statistically but is a convenient method of handling categorical, rather than scalar, data.

I realize, Frank, that this terminology may be foreign to you, but I find I can't express myself without using it, which is unfortunate. But, hopefully, you and can cut through all of the statistical jargon. My only excuse for the jargon is the fact that experimental psychologists

STAT

*Garner, W.R. & McGill, W.J. The relation between information and variance analysis. *Psychometrika*, 1956, 21, 219-228.

August 13, 1963

Page 5

are continually faced with the omnipresent and profound fact of individual differences and must approach and design experiments using statistical models.

4. Because of the large number of stimuli being judged, I don't see any way of experimentally controlling the effects due to the order in which they are presented. Significant effects due to order of stimulus presentation have been demonstrated repeatedly in psychophysical work. The only thing I can suggest is that about three anchor stimuli, one at the middle and one adjacent to either end of the judgmental scale be presented first and then the judgments of them not be used in the analysis. The remaining stimuli would be shuffled and presented in a different random order to each subject. In other words, the effects due to order of stimulus presentation would be randomized across all subjects.
5. I would consider whether or not the subjects should be told in the instructions that the photos differed in blur and graininess, when they are being asked to make a one-dimensional judgment of quality.

I am very much interested in, and in sympathy with the need for, the study is doing. I would like an opportunity to discuss it personally with him, for I fear that I may not have communicated too much here.

I realize that his work is a first step, but I think it is an important step toward a goal I feel is very much worth seeking: a specification of the relations among physical measures of photo image quality, subjective measures of image quality, and objective measures of PI performance. I feel that such a specification would be exceptionally valuable in all phases of our work: materials, collection, processing, and exploitation.

A final point of qualification. I showed these comments and suggestions to a fellow in our office here who has specialized in psychophysical research. His reaction was that I have over-simplified the problems. However, in general, he agrees with me, though he feels what I have said will make difficult reading. My apologies.

STAT

[redacted]
August 13, 1963
Page 6

Incidentally, Frank, I know I said I would write to you a week ago. Well, I simply have not had an opportunity to write. Work at the office and at home piles up when I'm in Washington. This time it piled up at a prodigious rate.

Best regards,

Woy

[redacted]

STAT

DNS:lo

STAT

cc: [redacted]