Perceptual and Motor Skills, 1980, 50, 472-474.

NOTE ON CHOOSING A RESPONSE SCALE[1]

DAVID J. WEISS

California State University, Los Angeles

Summary.—The functional measurement criterion for choosing a response scale involves joint validation of a judgmental model and a scale. This criterion was applied by Weiss (1972), whose grayness averaging data led him to accept graphic ratings and reject magnitude estimation. These same data were reanalyzed in light of a different criterion, that of scale sensitivity, using a statistical test developed by Schumann and Bradley (1959). Graphic ratings were significantly more sensitive.

Choosing among response scales is an old problem for psychologists. Guilford (1954) has argued that the scale with the highest inter-rater reliability be used, and Ramsay (1973) has used the precision of estimation as an index of rating scale performance. A criterion in terms of Weber functions was introduced by Eisler and Montgomery (1974). Perhaps the most commonly used approach is that of information measurement. Transmitted information (Garner, 1962) is taken as a measure of a scale's ability to convey perceived differences in the stimuli being judged. Unfortunately, there has been a difficulty in much of the information-based literature. MacRae (1970) has pointed out a number of biases in estimates of transmitted information and has shown that these biases have been serious enough to lead investigators to incorrect conclusions. The value of many investigations is thus limited (but for an unbiased exception see Garner, 1960) because usually only the information measures have been reported.

Working from a different perspective, functional measurement theorists (Anderson, 1970) have approached the response scale problem as one of validity. That scale is deemed valid which allows an appropriate judgmental model to be confirmed. For example, Weiss (1972) tested an averaging model for judgments of average grayness. Using statistical and graphical tests of additivity, he found that the model was upheld when the judgments were expressed as graphic ratings but was rejected when the same subjects made their responses using magnitude estimation. The conclusion was that graphic ratings were valid, but magnitude estimates were not.

This report presents the grayness averaging data of Weiss (1972), analyzed in terms of scale sensitivity. It utilizes an analysis of variance approach to sensitivity developed by Schumann and Bradley (1959). This paper, which included a psychological example, does not seem to be widely known. One purpose of this note is to call attention to its potential usefulness. The Schu-

[1]Request reprints from D. J. Weiss, Department of Psychology, California State University-Los Angeles, 5151 State University Drive, Los Angeles, CA 90032.

mann and Bradley test applies significance logic to a comparison of F ratios from similar experiments. Thus one can test a null hypothesis of the form that the differential effects of a common set of stimuli are of comparable magnitude in two experiments.

The details of experimental procedure are given in Weiss (1972). Briefly, 8 subjects judged the average grayness of pairs of Munsell neutral value chips. The stimulus pairs were constructed from a 5×5 factorial design, with the levels of each factor ranging from nearly black to nearly white. Four replications of the design were carried out for each response mode. Since the same set of stimulus pairs was used for both response conditions, comparison of the F ratios for the "stimulus pairs" source in an analysis of variance on the responses should indicate whether one scale is more sensitive. It is of interest to see whether a criterion based on scale sensitivity yields the same conclusion as one based on the functional measurement criterion.

Analysis of variance showed that the F ratios for stimulus pairs, tested against interaction with subjects, were 72.07 for graphic ratings and 32.30 for magnitude estimation, each with 24/168 *df*. The larger F ratio for graphic ratings suggests an advantage for this mode, and the Schumann-Bradley test confirms that this advantage is significant at the .05 level. In this test W', the ratio of the F ratios, was 2.23. The other parameters required for the test, a' and b, which determine the critical value for W', were 316.14 and 84, respectively. The conclusion to be drawn from this analysis is that the graphic ratings better demonstrated differences among the stimulus pairs.

According to two logically distinct criteria, graphic ratings have been shown to be superior as a response technique to magnitude estimation. This conclusion is in accord with that of Anderson (1974), who called magnitude estimation "biased and invalid." Of course, magnitude estimation has been widely used (Stevens, 1971); but beyond face validity, there is little to justify this popularity.

The test of sensitivity developed by Schumann and Bradley (1959) seems to merit attention. In the present context, it focused on the same property of the response scale as would an information measure, namely, the extent to which different stimuli generate different responses; but the test has the important advantage of providing an assessment of statistical significance.

The present results also provide a defense against a charge which has occasionally been leveled against users of functional measurement. The criticism is that because nonsignificant interaction is usually taken as evidence in favor of a proposed additive model, there is a temptation to conduct experimental analyses which are low in power. Failure to reject a model may in any given instance result from a weak experiment. The grayness averaging data offer an empirical refutation of this criticism. The response scale on which no inter-

action was detected was also the one which was more sensitive. While validity must remain the central issue for a response scale, agreement using other criteria is satisfying.

REFERENCES

ANDERSON, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970, 77, 153-170.

ANDERSON, N. H. Cross-task validation of functional measurement using judgments of total magnitude. *Journal of Experimental Psychology*, 1974, 102, 226-233.

EISLER, H., & MONTGOMERY, H. On theoretical and realizable ideal conditions in psychophysics: magnitude and category scales and their relation. *Perception & Psychophysics*, 1974, 16, 157-168.

GARNER, W. R. Rating scales, discriminability, and information transmission. *Psychological Review*, 1960, 67, 343-352.

GARNER, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

MACRAE, A. W. Channel capacity in absolute judgment tasks: an artifact of information bias? *Psychological Bulletin*, 1970, 73, 112-121.

RAMSAY, J. O. The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, 1973, 38, 513-532.

SCHUMANN, D. E. W., & BRADLEY, R. A. The comparison of the sensitivities of similar experiments: model II of the analysis of variance. *Biometrics*, 1959, 15, 405-416.

STEVENS, S. S. Issues in psychophysical measurement. *Psychological Review*, 1971, 78, 426-450.

WEISS, D. J. Averaging: an empirical validity criterion for magnitude estimation. *Perception & Psychophysics*, 1972, 12, 385-388.

Accepted December 12, 1979.