Empirical Tests of Scale Type for Individual Ratings

Rainer Westermann
University of Göttingen

This article describes eight studies that tested empirically the hypothesis that rating procedures lead to interval-scale measurements for each single subject. In order to enhance the probability of obtaining interval scales, subjects made numerical ratings and were deliberately instructed to choose their responses so that the algebraic differences between numbers represented the subjective differences between the corresponding objects with respect to the attribute under study. This approach is based on axiomatic measurement theory. It is exemplified by a study from clinical psychological research pertaining to the subjective fear aroused by each of 160 objects or situations. Any subject's ratings are regarded as interval-scale measurements of his or her individual degree of fear if the testable axioms of a finite, equally-spaced difference structure are satisfied empirically. These axioms pertain to ordinal judgments on differences, and they are tested empirically by deriving statistical hypotheses and using a refined significance-test method as an error theory. For the eight studies criteria were chosen primarily to avoid accepting false interval-scale hypotheses at the expense of relative high risks for false rejections. Nevertheless, empirical data allow acceptance of the hypothesis for 54 of the 114 subjects. As a consequence, for at least half of the subjects, this rating procedure seems to result in interval scales.

Both the use of rating scales and their (implicit) interpretation as interval-scale measurements are typical for a large part of psychological research in different areas. Dawes (1972, pp. 95–96) and Surber (1984) gave some illuminating examples from social and developmental psychology, respectively.

There are many different types of rating scales ranging from small sets of verbally labeled responses (e.g., agree, undecided, disagree) to numerical or graphical ratings. It is a common feature of all rating procedures, however, that they do not include checks for response consistency. For that reason, ratings are not representational measurements, but only index measurements in the sense given by Dawes (1972).

It is important to note that the same result holds for most magnitude or ''ratio'' scalings in the sense of Stevens (1960). These methods are frequently used not only in psychophysics, but also in social sciences (Lodge & Tursky, 1982). Usually, subjects are instructed to assign a number to each object so that subjective ratios are represented. Generally, no consistency checks can be made in these cases (cf. Dawes, 1972).

According to Dawes (1972), index measurements can be evaluated only in terms of their usefulness, for example, in terms of predictive validity; and he stressed that there is no sense in asking what scale type is attained by a particular index measurement technique. It seems obvious that this position may be detrimental for both measurement theory and substantial research. When rating and magnitude scales are excluded from any measure-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 3, September 1985, pp. 265–274
© Copyright 1985 Applied Psychological Measurement Inc.
0146-6216/85/030265-10$1.75

ment theoretical analysis, the common (implicit) interpretation of rating scale values as interval scales can be neither justified nor criticized from a methodological point of view. In other words, there would be no criteria for differentiating between a sound and a nonsensical statistical and numerical interpretation of ratings or magnitude estimations. In addition, this exclusion certainly lessens both the impact and importance of measurement theory for psychology as a whole, and it supports the tendency to consider measurement theory a somewhat esoteric discipline.

However, Dawes' position is not the only way to attack the problem. On the contrary, it is indeed possible to analyze ratings and other index measurements from a measurement theoretical point of view. To be concrete, a method is presented here to test the hypothesis that any subject's ratings or direct estimations are of interval-scale level. This method has been applied in eight studies, the results of which are presented below. In comparison with other approaches (cf. Anderson, 1982; Birnbaum, 1982; Orth & Wegener, 1983), this testing procedure tries to minimize the number of necessary assumptions and relies exclusively on qualitative and relatively simple judgments.

Method

Each of the eight studies consisted of two parts. First, subjects rated a set of items according to a prespecified attribute. On the basis of these responses, special sets of items were selected for each subject to test the interval-scale hypothesis. In the second part, each subject made judgments with respect to these sets of items.

In Table 1, the eight studies are characterized (1) by the number of subjects for which the interval-scale hypothesis was tested, (2) by the number of items rated by the subjects, and (3) by the attribute to be judged. In Study 1, nine subjects gave direct ratings of their personal fear with respect to 160 objects or situations. In Study 2, the likeableness of personality trait adjectives was rated (Westermann, 1984). Studies 3 and 4 were follow-up studies of Study 1. In Study 3 the same kind of fear-arousing items were assessed as in Study 1, whereas in Study 4 fear-arousing situations were described in more detail in one or two sentences. In Study 5 job characteristics of psychologists were judged with respect to their attractiveness for psychology students. In Study 6 different combinations of communicators and sources were scaled with respect to their trustworthiness. Studies 7 and 8 dealt with belief strength concerning political opinions and the respondent's personal situation at the university, respectively.

For a more detailed description of the method of testing hypotheses concerning the scale type of ratings and other index measurement, the first study is taken as an example.

Behavior therapists use fear inventories both as clinical and research instruments (Arrindell, 1980; Wolpe, 1973). The best known fear inventories are the various forms of the Fear Survey Schedule (FSS) presented, for example, by Wolpe and Lang (1964), Suinn (1969), Lawlis (1971) and Wolpe (1973) (for a review, see Mack & Schröder, 1977, and Seidenstücker & Weinberger, 1978). The FSS-III from Wolpe and Lang (1964), for example, consists of 72 items such as "noise of vacuum cleaner," "cats," and "looking foolish." The items "refer to things and experiences that may cause fear or other unpleasant feelings" (p. 28). Subjects are asked to place each item in one of five categories labeled verbally from "not at all" to "very much." Thus, this method results only in a rather coarse placement of fear-arousing stimuli in five ordered categories. To obtain a finer ordering and quantification, some authors recommend using rating methods with a larger number of possible responses (Baade, Burck, Koebe, & Zummvenne, 1980; Oswald, 1980). Then, the assumption is usually made that each subject's numerical responses can be interpreted as interval-scale measurements.

From various Fear Survey Schedules, 160 items were compiled. Each subject was asked to judge his or her personal magnitude of fear with respect to each object or situation. Judgments were given in terms of a rating scale ranging from 0 to 100. The two end points were defined verbally: 0 means to be completely free of fear, and 100 means the maximum fear the subject can imagine. In addition, subjects were asked to give their responses so that equal subjective fear differences were represented by equal numerical differences. For the sake of clarity, a ruler-like scale from 0 to 100 was presented graphically.

The resulting individual numerical assignments (scale values), *S*, were used to construct a set of items by means of which the interval-scale hypothesis could be tested for each subject (cf. Westermann, 1982, 1983, 1984).

To be concrete, the two critical necessary conditions for interval-scale measurement were tested, which are formulated as axioms of finite, equally-spaced difference structures: Equal-Spacing and Monotonicity. A complete axiomatic definition of a finite, equally-spaced difference structure was given by Krantz, Luce, Suppes, & Tversky (1971). The most important point for the present studies is the fact that axioms defining difference structures pertain to an order relation $\geqslant$ between pairs of stimuli. Thus, $(a, b) \geqslant (c, d)$ may mean, for example, that (according to a person's judgment) the difference between objects *a* and *b* with respect to a certain attribute is not less than the respective difference between the two other objects *c* and *d*.

Testing the Equal-Spacing Axiom

For numerical assignments resulting from a rating procedure, there is usually a large number of quadruples of stimuli with the property that the absolute difference between the scale values of the first two stimuli equals the corresponding differ-

Table 1
Characteristics of the 8 Empirical Studies
Testing the Interval-Scale Hypothesis
for Individual Ratings

	Number of		Attribute to be Judged
Study	Subjects	Items	
1	9	160	degree of personal fear with respect to objects and situations
2	10	110	likableness of personality trait adjectives
3	17	70	degree of personal fear with respect to objects and situations
4	18	50	degree of personal fear with respect to situations (described more detailed)
5	10	47	attractiveness of jobs for psychologists
6	9	42	credibility of communicators and sources
7	19	50	belief strength concerning political problems
8	22	40	belief strength concerning university problems

ence between the other two stimuli:

$$|S_a - S_b| = |S_c - S_d| \quad , \tag{1}$$

with S_a, S_b, S_c, and S_d denoting the numbers assigned to stimuli *a*, *b*, *c*, and *d*, respectively. If a subject's numerical responses are indeed interval-scale level, the subjective difference between the first two stimuli (*a*, *b*) must be equal to the subjective difference between the other two stimuli (*c*, *d*), in which case the set of stimuli can be thought of as being part of a so-called standard sequence. A standard sequence is an ordered set of stimuli satisfying the Equal-Spacing Axiom of finite, equally-spaced difference structures, that is, having equal (subjective) intervals between adjacent stimuli (Krantz et al., 1971).

As a consequence, testing whether the subjective difference between *a* and *b*, on the one hand, and *c* and *d*, on the other, are equal is a test of the Equal-Spacing Axiom. In Studies 1 and 2, 20 different quadruples satisfying Equation 1 were randomly chosen for each subject in order to test the Equal-Spacing Axiom;[1] the numbers and scale values of the stimuli selected for Subject 1 are listed in Table 2. Note that the sum of the scale values for the first pair is always not less than the corresponding sum for the second pair. This allows a test of the conflicting hypothesis that a numerical difference of, say, 10 points means a smaller sub-

Table 2
Test of the Equal Spacing Axiom
for Subject 1 in Study 1

Quad. No.	Items								Response Type
	Pair H				Pair L				
	No.	S	No.	S	No.	S	No.	S	
1	132	40	143	50	147	0	104	10	H
2	133	70	151	80	83	60	125	70	L
3	111	40	93	60	149	0	127	20	H
4	159	80	142	100	146	20	89	40	H
5	127	20	144	50	152	0	154	30	H
6	148	50	131	80	62	40	136	70	H
7	129	20	96	60	153	0	44	40	H
8	71	30	155	100	74	0	87	70	L
9	113	20	92	70	160	0	157	50	H
10	133	70	134	80	90	50	50	60	L
11	24	20	95	80	123	0	51	60	H
12	14	40	146	100	116	20	105	80	L
13	71	30	121	100	126	0	57	70	H
14	73	10	108	80	130	0	78	70	H
15	117	20	138	100	137	0	109	80	H
16	45	20	101	100	139	0	86	80	L
17	67	10	118	100	102	0	142	90	L
18	158	50	29	60	35	20	79	30	L
19	16	30	25	70	2	20	38	60	L
20	135	50	88	100	103	0	122	50	L

Responses of Type H: Frequency f = 11
Proportion $\hat{p}$ = .55

[1] Random samples of items and quadruples are appropriate for overall tests of the axioms and the interval-scale hypothesis. For more specific tests, researchers may restrict the corresponding populations, or they may deliberately choose the sets and subsets in which they are primarily interested.

jective difference in upper regions of the scale than in lower regions. Actual presentation, however, was determined randomly for each subject.

To avoid any problems with indifference or equality judgments, subjects in Studies 1 and 2 were forced to judge which of the two differences was larger. These responses were coded as "H" or "L," indicating whether the difference between the pair with the higher or the lower scale values was judged larger. If the Equal-Spacing Axiom holds, both possible responses have the same probability, that is,

$$p(\mathrm{H}) = p(\mathrm{L}) = .5 \quad . \qquad (2)$$

This hypothesis can be tested by a binomial or sign test. However, with $n = 20$ replications, conventional significance levels, such as .05, lead to very high Type 2 error probabilities. It is important to note that any Type 2 error leads to falsely accepting the axiom as valid and, eventually, to falsely interpreting the numerical assignments as interval-scale measurements. Any Type 1 error, on the other hand, can only result in interpreting interval-scale values as ordinal scales. Thus, small risks of Type 2 errors seem to be more important than small significance levels for the problem at hand (cf. Westermann & Hager, 1983a, 1983b). As a consequence, it was decided to reject the null hypothesis of equal probabilities when the frequencies of responses of either type were not greater than 7 or not less than 13, which corresponded to proportions of .35 and .65, respectively. These cutoff points led to probabilities of .26 and .10 for Type 1 and Type 2 errors, respectively (for a "large" effect size in the sense of Cohen, 1977).

In Studies 3 through 8 subjects were allowed to judge the differences to be equal. To test the Equal-Spacing Axiom, half of these responses were assigned to both the response categories H and L. The number of quadruples used to test this axiom was reduced from 20 to 15. The Equal-Spacing Axiom was considered satisfied if the proportion of Type H responses was between .33 and .67. These critical values led to error probabilities of $\alpha = .12$ and $\beta = .16$ for "very large" effect sizes of $g = .30$.

Testing the Monotonicity Axiom

Monotonicity, as formulated both in finite and infinite difference structures, is a necessary condition for interval-scale level measurement pertaining to sets of six stimuli, which are symbolized as e, f, g, l, m, and n. In the case of only two possible responses, the Monotonicity Axiom reduces to the following condition:

$$\begin{array}{ll} \text{If} & (e, f) > (l, m) \quad , \\ \text{and if} & (f, g) > (m, n) \quad , \\ \text{then,} & (e, g) > (l, n) \quad . \end{array} \qquad (3)$$

If the difference between e and f is judged larger than the difference between l and m, and if the same is true for f and g in comparison with m and n, then the difference between e and g must be judged larger than the difference between l and n.

As a consequence, three quadruples have to be judged for a single test of the Monotonicity Axiom with respect to a set of six stimuli $[e, f, g, l, m, n]$. For each subject in Studies 1 and 2, three such sets of six stimuli were selected randomly under the constraint that both

$$\begin{array}{l} |S_e - S_f| > |S_l - S_m| \text{ and} \\ |S_f - S_g| > |S_m - S_n| \end{array} \qquad (4)$$

holds. In Studies 3 through 8, five such sets were selected to test monotonicity. Table 3 shows, as an illustration, the numbers and scale values of the three sets of stimuli selected for Subject 1 in Study 1. The first pair is always the pair with the larger difference in scale values, that is, (e, f), (f, g), or (e, g). When the subject judged the corresponding difference to be larger than the difference between the other two stimuli, the response was coded as "T," or otherwise, as "U."

If the Monotonicity Axiom holds, the same type of responses should be given in the case of all three quadruples derived from one set of six stimuli, that is, a T-T-T or a U-U-U sequence of responses is a positive result with respect to the validity of the Monotonicity Axiom. When there are Type T responses for both the first and the second quadruple in Expression 3, but a Type U response for the third one, or vice versa, this is a negative result. When there is one Type T response for one of the first two quadruples and a Type U response for the

Table 3
Test of the Monotonicity Axiom
for Subject 1 in Study 1

		Items								
Set	Quad.	Pair T				Pair U				Response
No.	No.	No.	S	No.	S	No.	S	No.	S	Type
1	21	107	0	140	50	110	0	73	10	T
	22	140	50	63	80	73	10	2	20	T
	23	107	0	63	80	110	0	2	20	T
2	24	112	0	106	50	120	0	67	10	U
	25	106	50	61	100	67	10	35	20	T
	26	112	0	61	100	120	0	35	20	T
3	27	84	0	82	50	85	0	98	10	U
	28	82	50	64	100	98	10	24	20	T
	29	84	0	64	100	85	0	24	20	T

Responses of Type T: Frequency f = 7

Proportion $\hat{p}$ = .78*

*Significantly larger than $p = .5$ ($\alpha = .10$).

other, the Monotonicity Axiom is not testable. In these cases the premise of the axiom is not given, and it does not make sense to ask whether the conclusion is satisfied empirically.

In the present studies, there were, at most, five opportunities to test the Monotonicity Axiom. Therefore, it was not possible to derive decision rules in terms of a significance test. Instead, in Studies 1 and 2, this axiom was considered satisfied if there was a positive result for at least one of the three sets, and if there were no negative results for the remaining sets. In Studies 3 through 8, five sets of stimuli were used to test monotonicity, and the axiom was considered satisfied if there were more positive than negative results.

Testing the Consistency Condition

Each quadruple presented to test monotonicity led to another kind of consistency check between the direct ratings in the first part of the study and the ordinal judgments on differences in the second part. If differences between numerical ratings do indeed represent (the order of) subjective differences, then the differences between the stimuli e, f, and g were expected to be judged larger than the differences between l, m, and n. Taking random fluctuations into account, Type T responses were expected to be more probable than Type U responses, that is,

$$p(\mathrm{T}) > .5 \quad . \tag{5}$$

This prediction was complementary to the prediction derived from the Equal-Spacing Axiom for stimulus pairs with equal differences in scale values. Although the test of this ordering condition was dependent on the test of the Monotonicity Axiom, it can be considered as a third way of testing the interval-scale hypothesis.

It is clear that Expression 5 can again be tested by a binomial test. In this case, however, any Type 1 error may lead to an unwarranted interpretation of the ratings as interval-scale measurements. For that reason the significance level should be chosen relatively low at the expense of power. For the current example there were only nine replications, however. Under these circumstances it was decided to set α equal to .09, which resulted in a power of $1 - \beta = .60$ for "large" effect sizes of $g = .25$ (cf. Cohen, 1977). In Studies 3 through 8 the Consistency Condition could be tested with 15 replications. The condition was accepted if the proportion of Type T responses was at least .72, which resulted in error probabilities of $\alpha = .06$ and $\beta = .31$ (for an effect size of $g = .25$).

In order to avoid erroneous interpretations of ratings as interval-scale measurements, the interval-scale hypothesis for a subject was accepted only if all three conditions (Equal-Spacing, Monotonicity, and Consistency) were considered as satisfied. This is a very strict criterion, but it seems difficult to give justifications for more lenient criteria on the basis of significance tests or probability theory. The total set of quadruples selected for any subject to test equal-spacing, monotonicity, and consistency was presented in written form and in random order. In addition, the order of the two pairs of stimuli forming a quadruple and the order of the items within each pair were determined randomly.

Results

As an illustration, the results for Subject 1 of Study 1 are reviewed. Table 2 contains the responses of this subject with respect to those 20 quadruples selected to test the Equal-Spacing Axiom. The proportion of responses of Type H is .55. So, the Equal-Spacing Axiom is regarded as satisfied for this individual.

The responses pertaining to monotonicity and order consistency are shown in Table 3. There is a positive result for the first set of stimuli with respect to the Monotonicity Axiom. Subject 1 has given Type T responses to all three quadruples. Set 2, however, does not allow a test of monotonicity because there is a T response to the first, but a U response to the second quadruple. The same is true for the third set of stimuli. According to the criteria specified above, the Monotonicity Axiom is considered as satisfied empirically. The Consistency Condition is satisfied, also. As shown in Table 3, Subject 1 gave Type T responses in seven out of nine cases. Taken together, all three tests speak in favor of the interval-scale hypothesis for this subject.

The results of Study 1 are summarized in Table 4. According to the criterion described above, the Equal-Spacing Axiom was not satisfied for Subjects 2, 4, 7, and 8. In addition, orderings of differences did not appear to be consistent for Subjects

Table 4
Tests of the Interval-Scale Hypothesis for All Subjects in Study 1

Subj. No.	Equal Spacing Axiom		Consistency		Monotonicity Axiom				Total Eval.
	$\hat{p}(H)$	Eval.	$\hat{p}(T)$	Eval.	1	2	3	Tot.	
1	.55	+	.78	+	+	o	o	+	+
2	.75	-	.89	+	-	+	+	-	-
3	.60	+	.78	+	+	o	o	+	+
4	.30	-	.56	-	o	o	-	-	-
5	.50	+	.78	+	+	o	o	+	+
6	.45	+	.78	+	o	+	+	+	+
7	.75	-	.67	-	o	+	o	+	-
8	.30	-	.89	+	+	+	o	+	-
9	.50	+	.89	+	o	+	+	+	+

Note. The Axioms or conditions are considered as satisfied (+), not satisfied (-), or not testable (o) according to the criteria specified in the text.

4 and 7, and there were deviations from monotonicity for Subjects 2 and 4. Adopting the strict criterion specified above, the interval-scale hypothesis was rejected for these four subjects, with at least one rejection in the tests of the three conditions. For the other five of the nine subjects, however, there was no negative result. As a consequence, the ratings for these five subjects can be tentatively considered as interval-scale values.

The results of all eight studies are summarized in Table 5. A total number of 114 subjects were run. The Equal-Spacing Axiom was satisfied by 80% of the subjects, the Consistency Condition by more than 70%, and the Monotonicity Axiom by 67%. All three conditions can be regarded as satisfied for 54 of the 114 subjects.

Discussion

According to the criteria specified above, about 50% of the subjects seemed to have given ratings at the interval-scale level. This proportion may seem low; two facts, however, have to be taken into account. First, all these studies were conducted with students participating in an undergraduate course for experimental psychology, so that the subjects' motivation to give careful responses might have been only moderate. Second, decision criteria were chosen in order to hold the risk of falsely accepting interval-scale level for any subject's ratings at reasonable low levels, so that the risk of falsely rejecting the interval-scale hypothesis must be considered relatively high. In other words, these tests were rather severe (Meehl, 1967; Popper, 1975; Westermann & Hager, 1983a, 1983b).

Therefore, these results seem to corroborate the hypothesis that a considerable proportion of subjects were able to give ratings at the interval-scale level. Because of the limited scope of these studies, further severe tests of this hypothesis would be worthwhile.

In applied psychological research, however, there is not generally the opportunity to check scale properties for each of the subjects by the methods presented here and to discard those subjects that failed the test. Nevertheless, this result may have some consequences for applied psychological researchers who plan to use rating scales.

Beginning from the result that at least about 50% of the subjects in these studies seemed to be able to give ratings at the interval-scale level, researchers can try to maximize the corresponding proportions in their own studies. First, subjects should be provided with a proper response format, such as the ruler-like scale from 0 to 100. Second, subjects must be deliberately instructed how to use numbers and numerical differences to represent subjective magnitudes and differences. Third, researchers should try to motivate their subjects to judge carefully in accordance with the instructions. Under

Table 5
Results of the Studies Testing the Interval-Scale
Hypotheses for Individual Ratings

		Number of Subjects Fulfilling			
Study No.	No. of Subjects	Equal Spacing	Consistency	Monotonicity	All three Conditions
1	9	5	7	7	5
2	10	6	6	7	4
3	17	14	13	10	7
4	18	16	15	14	12
5	10	8	5	3	2
6	9	7	7	6	4
7	19	17	14	15	10
8	22	19	15	14	10
Total	114	92	82	76	54

these conditions, researchers can reduce their risk of erroneous conclusions when interpreting all resulting ratings as interval-scale measurements, and they will have a better justification for using parametric statistical tests, for averaging individual ratings to derive mean scale values, and so forth.

As has been shown in this article, contrary to the position held by Dawes (1972), it is possible to answer questions concerning the scale type of ratings in a reasonable manner. Thus, usefulness is not the only criterion to evaluate ratings, and Dawes' classification of ratings as index measurements may be somewhat misleading.

There remains a basic difference, however, between such interpretations of ratings as, say, interval-scale measurements and genuine fundamental and representational measurements at the interval-scale level in the sense given by Suppes and Zinnes (1963) and by Dawes (1972). Interval-scale interpretations of ratings usually are based on the results of *other* studies testing the interval-scale hypothesis for other individuals' rating. In addition, according to the present procedure, the axioms and necessary conditions for the desired interval-scale representation are tested, not in the course of the scaling procedure, but post hoc, after the ratings have been given.

For these reasons it would be preferable to speak of a ''quasi-representational'' interval-scale measurement when ratings are interpreted as interval scales. This new category of ''quasi-representational measurements'' is introduced to encourage a clear distinction between ratings and their interpretation, on the one hand, and the well-known classes of index, derived, representational, and fundamental measurements, on the other hand.

References

Anderson, N. H. (1982). Cognitive algebra and social psychophysics. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 123–148). Hillsdale NJ: Erlbaum.

Arrindell, W. A. (1980). Dimensional structure and psychopathology correlates of the Fear Survey Schedule (FSS-III) in a phobic population: A factorial definition of agoraphobia. *Behaviour Research and Therapy, 18,* 229–242.

Baade, F. J., Burck, J., Koebe, S., & Zummvenne, G. (1980). *Theorieen und Methoden der Verhaltenstherapie.* Tübingen: Deutsche Gesellschaft für Verhaltenstherapie.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale NJ: Erlbaum.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Dawes, R. M. (1972). *Fundamentals of attitude measurement.* New York: Wiley.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Vol. I.* New York: Academic Press.

Lawlis, G. F. (1971). Response style of a patient population on the Fear Survey Schedule—III. *Behaviour Research and Therapy, 9,* 95–102.

Lodge, M., & Tursky, G. (1982). The social-psychophysical scaling of political opinion. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 177–198). Hillsdale NJ: Erlbaum.

Mack, B., & Schröder, G. (1977). Entwicklung ökonomischer Angst-Symptom-Listen für die klinische Diagnostik. *Psychologische Beiträge, 19,* 426–445.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Orth, B., & Wegener, B. (1983). Scaling occupational prestige by magnitude estimation and category rating methods: A comparison with the sensory domain. *European Journal of Social Psychology, 13,* 417–431.

Oswald, W. D. (1980). Zur Operationalisierung von ''State''-Angst, ''Trait''-Angst und Anspannung mit Hilfe individueller Ankersituationen. *Diagnostica, 26,* 21–31.

Popper, K. R. (1975). *The logic of scientific discovery* (8th ed.). London: Hutchinson.

Seidenstücker, G., & Weinberger, L. (1978). Entwicklung einer Angstliste. *Diagnostica, 24,* 78–88.

Stevens, S. S. (1960). Ratio scales, partition scales, and confusion scales. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications* (pp. 49–66). New York: Wiley.

Suinn, R. M. (1969). The STABS, as measure of test anxiety for behaviour therapy; Normative data. *Behaviour Research and Therapy, 7,* 335–339.

Surber, C. F. (1984). Issues in using quantitative rating scales in developmental research. *Psychological Bulletin, 95,* 226–246.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1–76). New York: Wiley.

Westermann, R. (1982). Empirical test of scale type resulting from the power law for heaviness. *Perceptual and Motor Skills, 55,* 1167–1173.

Westermann, R. (1983). Interval-scale measurement of attitudes: Some theoretical conditions and empirical testing methods. *British Journal of Mathematical and Statistical Psychology, 36,* 228–239.

Westermann, R. (1984). Zur empirischen Uberprüfung des Skalenniveaus von individuellen Einschätzungen und Ratings. *Zeitschrift für Psychologie, 192,* 122–133.

Westermann, R., & Hager, W. (1983a). On severe tests of trend hypotheses in psychology. *Psychological Record, 33,* 201–211.

Westermann, R., & Hager, W. (1983b). The relative importance of low significance level and high power in multiple tests of significance. *Perceptual and Motor Skills, 56,* 407–413.

Wolpe, J. (1973). *The practice of behavior therapy* (2nd ed.). New York: Pergamon.

Wolpe, J., & Lang, P. (1964). A fear survey schedule for use in behaviour therapy. *Behaviour Research and Therapy, 2,* 27–30.

Acknowledgments

Thanks are due to Heinz Liebeck for a list of fear-arousing stimuli; S. Gunkel, C. Kordowski, Y. Missios, S. Anton, B. Giermann, H. Thölking, K. Lampe, A. Bettges, S. Drangmeister, C. Schäfer, W. Storch, S. Floer, D. Kudlacek, E. Andres, S. Schemat, M. Schützwohl, K. Engel, G. Czechatz, S. Pollmann, H.-J. Fromm, and B. Rumpf for conducting the empirical studies; and to two anonymous reviewers for helpful comments on an earlier draft of this article. Special appreciation is expressed to Kornelia Junge and Eckehard Kuhlmei for computer programming and assistance in collecting and analyzing the data.

Author's Address

Send requests for reprints or further information to Rainer Westermann, Institut für Psychologie, Universität Göttingen, Goβlerstr. 14, D-3400 Göttingen, West Germany.