Linked Raters' Judgments: Combating Problems of Statistical Conclusion Validity

George S. Howard
University of Notre Dame

Fernando H. Obledo, David A. Cole, and Scott E. Maxwell
University of Houston

The traditional procedure for obtaining judged ratings, to ascertain if treatment-related change has occurred, involves the randomization of the materials to be rated. An alternative approach (linked judgments) is investigated as a potential solution to certain instrumentation-related threats to statistical conclusion validity of the incumbent rating procedure. Data from a weight reduction study are presented which suggest that linked raters' judgments provide both a more powerful and a more valid index of treatment effectiveness than the traditional procedure.

Methods for the accurate evaluation of the effectiveness of treatment and training interventions have been of interest to a broad spectrum of psychologists. The research design which is most frequently employed to evaluate treatment or training intervention effectiveness is the two group pretest-posttest design (Design 4; Campbell & Stanley, 1963). With the random assignment of subjects to treatment and control conditions, potential threats to internal validity are thought to be adequately controlled by Design 4. One of the threats to internal validity is instrumentation, which is defined as "changes in the calibration of a measuring instrument or changes in the observers or scorers used [which] may produce changes in the obtained measurements" (Campbell & Stanley, 1963, p. 5). The purposes of this article are to consider an instrumentation-related threat to statistical conclusion validity when judges' ratings are employed in a Design 4 manner, and to investigate the validity of an alternative judgment procedure called linked judgments.

Campbell and Stanley (1963) recommended that instrumentation effects can be controlled by the randomization of materials to be rated by judges: "Judges may judge a series of randomized sections of pretest, posttest, experimental, and control group transcriptions, [which] helps to control instrumentation in research . . ." (p. 14). Typically, the materials to be judged (e.g., tapes, test protocols) are randomized so that any systematic effects of instrument decay will be equally distributed through pretest and posttest ratings for both treatment and control subjects. This procedure yields mean pre and post ratings for treatment and control subjects which are approximately equally influenced by instrument decay in the rater's judgments. These data thereby yield an unbiased estimate of the treatment effect. While these treatment with control group comparisons are unbiased, instrumentation effects are still present, and they serve to reduce the reliability of the pre to post estimates of change. Simply stated, instrumentation effects, while controlled, are still present and influence the error term of the statistic employed, thereby altering the power of the test.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 1, Winter 1983, pp. 57–62
© Copyright 1983 Applied Psychological Measurement Inc.
0146-6216/83/010057-06$1.30

Linked Judgments

If it were possible to decrease the error in judges' estimates of pre to post change without reintroducing instrumentation bias into raters' judgments, the statistical power of the test would be increased, thereby yielding a more sensitive estimate of treatment effects. Cook and Campbell (1979) underscored the hazards of a statistical test having insufficient power by viewing statistical conclusion validity as a special case of internal validity. Bias refers to factors that systematically alter mean values which might lead to erroneous inferences. Biased means might lead to the conclusion that A does not affect B (when in fact it does), which is a clear threat to internal validity. Error, on the other hand, refers to factors which increase variability and decrease the chance of obtaining statistically significant results. If uncontrolled variability obscures true mean differences, it is again erroneously concluded that A does not affect B; this time the error in inference is produced by threats to statistical conclusion validity.

The idea behind linked judgments is simple: If pre and post ratings could be made simultaneously (or as close together in time as possible), instrumentation effects would have a negligible influence on this index of change. Therefore, it is proposed that the pre and post materials for each subject be presented to raters simultaneously to be judged. The position (order) of pre and post material will be randomized across subjects with raters not being cued as to which is pre or post. Further, instrument decay in raters will be controlled in treatment with control group comparisons by randomizing the order of presentation of materials of treatment and control subjects.

Viewed from a slightly different perspective, randomized versus linked judgments can be seen as parallels to the differences between subjects' judgment capacity when asked to make absolute versus relative judgments. In the randomized judgment procedure, judges are asked to make a series of absolute judgment ratings of materials for various subjects along some specified dimension. Miller (1958) has pointed out that there is a clear and definite limit to the accuracy with which the absolute magnitude of a unidimensional stimulus variable, which he calls the *span of absolute judgment*, can be identified. Any judgment made in the randomized procedure might be contaminated by instrumentation effects and imprecision due to the span of absolute judgment. The resulting measure of treatment-related change is influenced by both these sources of error in both pre and post ratings. Because measures of change would have error associated with all these sources, Cronbach and Furby (1970) have shown that such indices of change lack appropriate power. Miller (1958) has suggested a solution to the problem of the span of absolute judgment which prefigures the linked judgment procedure: "We are not completely at the mercy of this span, however, because we have a variety of techniques for getting around it and increasing the accuracy of our judgments. [One of the] . . . most important of these devices . . . [is] to make relative rather than absolute judgments . . ." (p. 104). In the linked judgment procedure judges are encouraged to make relative rather than absolute judgments, thereby allowing for more sensitive judgments and at the same time avoiding the problems associated with instrumentation effects.

Howard and Maxwell (1981) have demonstrated that linked judgments can find significant treatment-related change in assertiveness in subjects who participated in a group designed to increase assertiveness, while randomized judgments found nonsignificant differences between treatment and control group subjects. More powerful techniques do not necessarily imply more accurate measures. Because of the lack of a universally accepted criterion measure of change in assertiveness, Howard and Maxwell (1981) were unable to assess the relative validity (and accuracy) of linked versus randomized judgments. The present study considered the judged evaluation of an intervention in a weight reduction group, where a valid criterion measure was available. Actual weight loss was compared with linked versus randomized judgments of participants' pictures taken before and after the intervention, to ascertain if linked judgments of change were both more powerful *and* more valid than their randomized counterparts.

Method

Subjects

Students in 12 undergraduate courses responded to a request for subjects to take part in a weight reduction study. Students who were registered in a course which allowed credit for research participation (about 60% of the students) were given credit for their participation. Thirty-two students expressed an interest in the study and 16 were randomly assigned to the weight reduction group, while the remainder served as a wait-list control group. Complete data were gathered on 25 students (12 treatment, 13 control). Those in the control group were offered the weight reduction treatment at the conclusion of the formal study.

Treatment

The treatment package developed for this study included components which had been successful in previous studies, such as self-monitoring procedures (e.g., Gottman & McFall, 1972; Johnson & White, 1971), physical activities (e.g., Mahoney & Mahoney, 1976; Stuart & Davis, 1972), and cognitive ecology techniques (e.g., DiLoreto, 1971; Meichenbaum, Gilmore, & Fedoravicious, 1971). The treatment lasted 7 weeks and sessions were 2 hours per week. Students were reminded of their commitment to lose weight via a ''buddy system'' wherein each group member telephoned one other member and was in turn contacted by some other individual every day of the week except the day which the group met. The group was cofacilitated by two of the authors, both of whom had previous experience conducting weight reduction groups.

Measures

Actual weight was assessed by means of a Detecto-Medic balance beam scale. Students were fully clothed when weighed, and each weighing was performed at the same time of the day. Photographs were taken before and after the treatment program with a Nikon 35mm SLR camera. The students stood against a wall which was blank except for a measuring scale which went from 4′6″ to 6′6″. The height of each student was obvious in each picture due to the inclusion of the scale. The students were fully clothed and were instructed to wear the same clothes at both picture-taking sessions (83% of treatment and 85% of control subjects did so). Pictures were taken from a distance of 7 feet and included the entire body of each student with the exception of ankles and feet. The students were instructed to stand casually with their arms at their sides.

Judges

Ten students enrolled in a course in Research Methods in Psychology volunteered to serve as judges in the study. Judges were randomly assigned to either the linked or unlinked judgment conditions. Judges in the unlinked condition were presented each photograph separately and asked to assess the student's weight, which they recorded. Judges in the linked condition were presented both photographs for each student simultaneously and asked to assess the student's weight in each picture.

Procedure

Immediately before the first session, the students were weighed and photographed. Control subjects were instructed to not return until one week after the last group meeting. Treatment subjects returned each week for the treatment sessions. Students were weighed and photographed the week after the last group meeting. Photographs were developed by a commercial developer. Approximately four weeks later, 10 judges were randomly assigned to linked and unlinked ratings groups. Each judge individually studied the photographs and recorded his/her ratings on a standard data sheet. Students and raters were debriefed immediately after their participation in the study.

Results

The effectiveness of the weight reduction intervention was assessed in three ways: actual weight loss, unlinked raters' judgments, and linked raters' judgments. Although analysis of covariance is generally the preferred method of analysis in randomized pretest/posttest designs (Huck & McLean, 1975), when pretest with posttest correlations are as high as in the present study, analysis of change scores yields a slightly more powerful test of the treatment effect. However, treatment effects were not significant when actual weight loss was considered (*t*(23) = 1.91). There was also a nonsignificant treatment effect for the change in the average of the five linked judges' ratings (*t*(23) = .79), as well for the change in the average of the five unlinked judges' ratings (*t*(23) = .67).

The issue of accuracy was approached by forming the absolute value of the difference between an actual weight (or a weight loss) and a particular judge's rating of weight (or the difference between weight ratings). This procedure yielded a total of 250 accuracy scores of pre weight, 250 accuracy scores of post weight, and 250 measures of weight change accuracy. Accuracy scores were deemed preferable to correlation coefficients because correlations represent the degree to which scores on two measures are proportional when expressed as deviations from their means. Accuracy implies the further condition that the absolute magnitude of these scores on these measures be equivalent.

Table 1 presents the absolute mean accuracy for each of the five linked judges and each of the five unlinked judges for pre weight, post weight, and weight change. Inspection of the data reveal that unlinked Judge 5 was an outlier. A more conservative statistic was therefore employed (see below) in order to guard against spurious results due to this outlier. The purpose of comparing linked with unlinked judgment procedures is to determine if either judgment technique has superior accuracy when generalizing across subjects and judges. The design of the present study involved crossing subjects with judges who were nested under conditions (i.e., linked versus unlinked). Since treatment effects were nonsignificant, the treatment factor was not included. Because the desire was to generalize across both subjects and judges, both facets represent random effects. Therefore, the appropriate statistic is a quasi-*F* statistic (Myers, 1979). Three analyses were conducted, one each for pre, post, and change in absolute mean accuracy scores. Although accuracy scores of this sort tend to be slightly

Table 1
Mean Absolute Accuracy in Pounds
for Pre, Post, and Change Judgments

Condition and Judge	Pre		Post		Change	
	Mean	SD	Mean	SD	Mean	SD
Linked						
1	10.65	6.77	11.24	8.08	5.62	6.27
2	13.22	9.81	13.52	8.77	3.45	3.53
3	13.74	11.67	11.90	9.69	3.17	4.41
4	16.88	10.92	17.04	13.92	4.99	5.14
5	12.14	7.90	12.00	7.72	3.11	3.00
Unlinked						
1	14.14	12.76	11.07	7.49	9.38	8.48
2	13.16	12.20	14.78	7.51	10.18	7.90
3	16.40	14.80	15.00	10.50	7.76	8.82
4	15.40	11.59	14.36	9.20	4.84	4.84
5	28.89	22.56	20.94	19.42	16.55	13.44

positively skewed, Santa, Miller, and Shaw (1979) found the quasi-F statistic to be robust to violations of normality. There was no significant difference between linked and unlinked judgment procedures for pre ($F'(1,15) = 1.54$) or post ($F'(1,16) = 1.12$) absolute mean accuracy scores.[1] However, the two judgment procedures were reliably different on mean absolute accuracy scores of weight change ($F'(1,8) = 7.17, p < .05$) such that linked judgments were more accurate (see Table 1).

To further gauge the magnitude of the differences between the accuracy of linked and unlinked approaches, an alternative technique for estimating linked versus unlinked accuracy was developed. An average judged change score was computed for each subject separately for linked and unlinked judges. For 20 of the 25 subjects, the linked estimate of change was closer to actual weight loss than was the unlinked estimate. This trend was significantly different from the expected 50% ($Z = 3.00, p < .01$).

Howard and Maxwell (1981) found that linked judgments possessed greater statistical power than unlinked judgments. Greater power can result from differences in the size of the treatment effect obtained by the two procedures or differences in the within-group variability. The size of the treatment effect was virtually identical for the linked and unlinked judgments as well as the treatment effect for the actual weight change. Inspection of within-groups variability revealed some striking differences. Pooled within-groups standard deviation of actual weight change was 3.19; for linked judges the within-groups standard deviation was 3.02; but for unlinked judges the within-groups standard deviation was 8.52. Given the comparable treatment effect, the unlinked approach would require a larger sample size to obtain the same power. In particular, the needed ratio is $(8.52/3.02)^2 = 7.95$, in this instance. That is, assuming the same treatment effect, almost eight times as many subjects would be needed to obtain a significant effect if the unlinked approach is used.

Discussion

These findings suggest that, in addition to being a more powerful index of change (cf. Howard & Maxwell, 1981), linked judgments might also be more valid than the traditional unlinked (randomized) method of obtaining judged ratings. If subsequent evidence corroborates these findings, researchers might consider employing linked judgments to avoid potential problems of statistical conclusion validity due to the apparent reduced power of unlinked judgments of treatment-related change.

The value of more powerful research designs and measures transcends merely avoiding errors of overconservatism. Few researchers are really interested in whether a training or treatment intervention is superior to no treatment at all or to a placebo control (cf. O'Leary & Borkovec, 1978). Rather, the concern is to find interventions that are superior to other equally plausible alternative interventions. Often, however, when two similar interventions are considered, no reliable differences can be observed. To the extent that this failure to reject the null hypothesis is due to the inability of the design to detect real but small differences between the interventions, researchers are subtly encouraged to avoid the more meaningful comparisons because they stand a reasonable chance of showing differences that are, in fact, trivial.

Looking more closely at why linked judgments might produce preferred measures of treatment-related change, some recent literature in measurement theory might be relevant. When particular values of stimuli are being estimated, the context in which the stimuli are presented affects the values assigned to the stimuli (see Parducci, 1982). With linked judgments, both stimuli are given to the rater simultaneously and the judge is asked to focus on how much change has occurred. In addition, with linked judgments, raters have more explicit information about where differences are expected to

[1] The numerator and denominator degrees of freedom for the quasi F are in general fractional. In this study both degree of freedom values were rounded down to the nearest integer in order to assess statistical significance.

occur (i.e., between pairs of stimuli). However, there are situations (e.g., instances where most of the pairs of stimuli show no differences) where context effects might be contaminating and represent a threat to external validity.

Overall, linked judgments appear to be a promising method for obtaining judged estimates of change. Researchers involved in evaluating treatment and training effectiveness in a broad array of content areas of psychology might find linked judgments a more sensitive and accurate evaluation approach. Finally, since judgments are easily obtained by both the linked and unlinked procedures, researchers are encouraged to employ a few more judges and to perform analyses in both the linked and traditional manner. Comparisons of these sets of data will serve to further clarify the relative strengths and limitations of both approaches to measuring change.

References

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.

Cronbach, L. J., & Furby, L. "How we should measure 'change'—or should we?" *Psychological Bulletin*, 1970, *74*, 68–80.

DiLoreto, A. O. *Comparative psychotherapy: An experimental analysis*. Chicago: Aldine-Atherton, 1971.

Gottman, J. M., & McFall, R. M. Self-monitoring effects in a program for potential H. S. Dropouts. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 273–281.

Howard, G. S., & Maxwell, S. E. Linked raters' judgments: A more sensitive measure of change. *Evaluation Review*, 1981, *6*, 140–146.

Huck, S., & McLean, R. Using a repeated measure ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 1975, *82*, 511–518.

Johnson, S. M., & White, G. Self-observation as an agent of behavior change. *Behavior Therapy*, 1971, *2*, 488–497.

Mahoney, M. J., & Mahoney, K. *Permanent weight loss*. New York: Norton, 1976.

Meichenbaum, D. H., Gilmore, J. B., & Fedoravicious, A. Group insight versus group desensitization in treating speech anxiety. *Journal of Consulting and Clinical Psychology*, 1971, *36*, 410–421.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. In D. C. Beardslee & M. Wertheimer (Eds.), *Readings in perception*. Princeton NJ: Van Nostrand, 1958.

Myers, J. L. *Fundamentals of experimental design*. Boston: Allyn & Bacon, 1979.

O'Leary, K. D. & Borkovec, T. D. Conceptual, methodological, and ethical problems of placebo groups in psychotherapy research. *American Psychologist*, 1978, *78*, 821–830.

Parducci, A. Category ratings: Still more contextual effects. In B. Wegener (Ed.), *Social attitudes and psychological measurement*. Hillsdale NJ: Erlbaum, 1982.

Santa, J. L., Miller, J. J., & Shaw, M. L. Using quasi-*F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin*, 1979, *86*, 37–46.

Stuart, R. B., & Davis, B. *Slim chance in a fat world: Behavior control of obesity*. Champaign IL: Research Press, 1972.

Author's Address

Send requests for reprints or further information to George S. Howard, Department of Psychology, University of Notre Dame, Notre Dame IN 46556, U.S.A.