

Comparing Holistic and Disaggregated Ratings in the Evaluation of Scientific Presentations

HAL R. ARKES^{1*}, VICTORIA A. SHAFFER¹ and ROBYN M. DAWES²

¹*The Ohio State University, USA*

²*Carnegie-Mellon University, USA*

ABSTRACT

The National Institutes of Health refused to switch to disaggregated ratings as a method for evaluating proposals, because no contest between disaggregated and holistic ratings had ever used scientific materials as the to-be-rated stimuli. We designed two studies to fill this research void. Participants rated scientific convention presentations either using a holistic procedure in which one overall rating was given or a disaggregated procedure in which one rating was given to each of five criteria. In four of the six convention sessions the disaggregated ratings led to higher inter-rater reliability than did the holistic ratings; three of these differences were statistically significant. The inter-rater reliabilities between the two types of ratings collapsed across all sessions differed significantly. In a second experiment, participants rated posters using either disaggregated or holistic ratings. The disaggregated ratings again led to higher inter-rater reliability, but not significantly so. In 35 of the 43 sessions in which disaggregated and holistic ratings were compared, the variance of the disaggregated ratings was smaller. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS disaggregated ratings; holistic ratings; evaluation; inter-rater reliability

INTRODUCTION

Paul Meehl's (1954) book *Clinical Versus Statistical Prediction* began a controversial dialogue between clinicians and members of the psychological research community which has continued to this day. Although his book addressed topics principally within the psychological domain, the controversy has fueled subsequent investigations in such varied domains as liver disease (Einhorn, 1972), loan applications (Stillwell, Barron, & Edwards, 1983), violence predictions (Werner, Rose, & Yesavage, 1983), and graduate school applications

* Correspondence to: Hal R. Arkes, Department of Psychology, The Ohio State University, 240N Lazenby Hall, 1827 Neil Avenue Mall, Columbus, OH 43210-1222, USA. E-mail: Arkes.1@osu.edu

Contract/grant sponsor: National Science Foundation, Program in Decision, Risk, and Management Science, award 0109250.

(Dawes, 1971). The present research attempts to further Meehl's work by examining the benefits of a more routinized, mechanical method for evaluating scientific materials such as research presentations at professional conventions or proposals submitted to federal funding agencies.

The impetus for the current manuscript was a series of meetings held at the National Institutes of Health (NIH) in 1995. The senior author was a member of the Review of Grant Applications (RGA) Committee, whose mission was to suggest changes in the method by which NIH evaluated grant applications. A co-author, Robyn Dawes, was a consultant to the RGA Committee. Based on research summarized by von Winterfeld and Edwards (1986, pp. 362–369), we suggested that reviewers of NIH grant proposals should perform disaggregated rather than holistic evaluations. In other words, reviewers should rate each proposal on each of several criteria rather than providing one single rating which reflects one's overall opinion of each proposal.

NIH reviewers had always been asked to give one overall number—a holistic rating. However, beginning with Meehl (1954) and Raiffa (1968), researchers have advocated the disaggregated procedure: Ask evaluators to rate each criterion and then amalgamate these disaggregated ratings. Dawes and Corrigan (1974) were among the first to show that simply adding standardized disaggregated ratings resulted in a sum which was more accurate in a number of prediction and judgment tasks than was the holistic rating. A large number of studies have provided support for this conclusion (e.g., Dawes, Faust, & Meehl, 1989; Fischer, 1977; Kleinmuntz, 1990; Sawyer, 1966, Stillwell et al., 1983). However the support is not quite universal. For example, Morera and Budescu (1998) found that the disaggregated strategy was not superior to a holistic one when applied to judgments derived from an analytical hierarchy process (Saaty, 1970). Similarly, Cornelius and Lyness (1980) found that the disaggregated strategy was not universally superior to the holistic one. For example, when incumbents rated their common job, those using a disaggregated procedure had superior inter-rater reliability compared to those using a holistic procedure on 70% of the jobs, the holistic procedure being superior on 20%, and the procedures being equal on 10%.

The general superiority of disaggregated over holistic judgments has also been found in tasks which are less closely related to the evaluation of NIH grant proposals. Armstrong, Denniston, and Gordon (1975) and MacGregor, Lichtenstein, and Slovic (1988) have demonstrated the superiority of the disaggregated strategy in making point estimates for unknown quantities. Hora, Dodd, and Hora (1993) and Kleinmuntz, Fennema, and Peecher (1996) found the same result in assessing subjective probability distributions for unknown quantities, although Henrion, Fischer, and Mullin (1993) reported contrary results. Considering the research literature as a whole, we felt that our suggestion to consider disaggregated ratings had sufficient empirical support to warrant consideration by the RGA Committee.

The reaction to our suggestion was intensely negative. NIH personnel raised two primary objections. First, the prior research repeatedly demonstrating the superiority of the disaggregated method had not included NIH proposals as the to-be-rated stimuli. How could we be sure that the prior results would generalize to the domain of scientific stimulus materials?

Second, many studies comparing holistic to disaggregated methods used a target number or "gold standard" which the two methods competed to match. For example, Stillwell et al. (1983) compared the accuracy of holistic and disaggregated methods in predicting credit worthiness. The gold standard was the credit worthiness of each potential borrower as calculated by the bank's formula based upon a discriminant function analysis. In rating grant applications, however, there is no readily available gold standard of merit. Although Dawes (1977) had suggested that disaggregated ratings would be superior to holistic ratings even in a context with no external gold standard, NIH officials were unconvinced. With no gold standard many interested parties at NIH asked, how it could be proved or even suggested that one method was superior to the other?

The purpose of this research is to respond to these two objections. Thus this research is quintessentially practical: Are disaggregated ratings superior to holistic ones when the items to be evaluated are scientific

materials such as NIH grant applications? Tens of billions of dollars are spent every year based on such evaluations; it would behoove us to find the best way of performing them.

PSYCHOMETRIC CONSIDERATIONS

How does one demonstrate the superiority of one method of evaluation over another? If grant applications were the stimulus to be evaluated, perhaps one could wait for a decade or so until the research supported by NIH had come to fruition and its results could be fully assessed. This is not a pure measure, however, because the investigators who received NIH funding would enjoy an advantage over those who were denied funding. Compared to non-funded investigators, funded ones can more easily hire talented personnel and buy superior equipment. These advantages would enable funded researchers to have more successful research programs than non-funded researchers, even if the funding decision were arbitrary. This confound would seriously contaminate the evaluation of the method by which funding was decided. In a preference task, the problem is even more difficult, because there is no obvious criterion. The lack of a criterion in a preference task has previously been discussed by several authors (e.g., Ubel & Loewenstein, 1997, pp. 651–652).

However there may be a way to use reliability to finesse the validity problem. Validity of a judgment can be assessed by obtaining the correlation between a judgment and some criterion. However a judgment should not correlate more highly with any other variable than it correlates with itself. In fact, the maximum validity coefficient between two variables is the square root of the product of the two reliabilities (Kaplan & Saccuzzo, 1997, p. 148). Therefore the reliability of a judgment obviously constrains the validity of that judgment. Hence many researchers have investigated the comparative reliability of holistic and disaggregated judgments, rather than assessing their comparative validities. In fact, empirical studies have verified that disaggregated judgments are more reliable, in the sense that their test-retest reliability surpasses that of holistic judgments (Ravinder, 1992). We will follow this same investigative tactic in this article, although we will assess reliability by means other than test-retest reliability. We will use inter-rater reliability.

To address the second objection, we will use scientific materials as our to-be-rated stimuli. In Experiment 1, we asked participants at a scientific convention to rate the oral presentations; in Experiment 2, we asked graduate students to rate posters presented at an undergraduate research colloquium. Of course, some raters used a holistic method, whereas others used a disaggregated one. The two questions that can be answered are whether one method achieved higher inter-rater reliability than the other and whether one method produced smaller amounts of rating variability.

EXPERIMENT 1

Method

Participants

The participants were 101 members of the Society for Medical Decision Making attending the group's 2001 annual meeting. The membership of this Society includes physicians, psychologists, pharmacists, nurses, and other professionals in health related fields.

Materials

For each of the seven sessions of oral presentations, two types of rating sheets were used. At the top of the holistic rating sheet were printed the criteria to be used in making one holistic rating per oral presentation.

The five criteria were: 1. Significance: Is the topic significant? Does it concern a scientifically important subject or is it relevant for health policy? 2. Methods: Are the methods scientifically sound? 3. Results: Are actual results presented in enough detail and in an understandable way? 4. Conclusions: Do the conclusions follow from the results? Are they justified? Are the results generalizable? 5. Innovation: Is there something innovative about the presented material?

The holistic rating sheet contained the time, title, and author of each oral presentation followed by one 9-point rating scale anchored at “poor” and “excellent” with “average” typed under the mid-point (5) of the scale.

The second type of rating sheet was the disaggregated one, which was identical to the holistic one with a single exception. Rather than having one rating scale placed after the time, title, and author of each oral presentation, five 9-point scales were placed there, with the name of one criterion positioned above each scale.

Procedure

Participants were recruited in three ways. First, an e-mail was sent to all members of the Society for Medical Decision Making seeking volunteers to be in an experiment at the upcoming annual convention. Second, at the meeting itself a table was placed near the registration booth at which the first author was stationed in order to recruit more participants. Third, the first author or one of the convention coordinators made an announcement before each of seven sessions asking for more participants. The titles of the seven sessions were the following: 1. Plenary Presentation of Top Abstracts, 2. Cost-Effectiveness, 3. Health Economics, 4. Technology/Outcomes, 5. Psychiatry/Mental Health, 6. Pharmacoeconomics, 7. Shared Decision Making.

During recruitment all potential participants were informed that in order to take part in the research it would be necessary for a rater to be present for and rate every oral presentation within one session; they could not shift from one concurrent session to a second one being held in a different room. All participants were promised and paid \$35.

The rating forms were either handed in after each session or were mailed to the first author after the convention.

Results

The data from the Shared Decision Making session could not be used because there were several last minute scheduling changes due to the terrorist attacks of September 11, 2001. The rating sheets were created based on the announced program, but some members of the Society who were scheduled to present papers at this particular session chose not to fly to San Diego the following month. As a result, the program was changed at the last moment, and the format of the rating sheets prepared for the Shared Decision Making session before the meeting no longer corresponded to the actual presentations. Although some raters tried to modify their sheets to accommodate the new order and new papers, it was apparent that the raters were inconsistent in their adaptation to the schedule change. Thus we did not have confidence that we could ascertain which rating corresponded to any article in this session; therefore, we discarded all the data from this session. In addition, one participant rated two sessions. We discarded the data from the second session he rated but used the data from the first session he rated. Another participant who received a holistic rating sheet nevertheless disaggregated the criteria and rated each one separately. We discarded his data. The ratings of one of the participants in the Cost-Effectiveness session had no variance, so this person's ratings were discarded. This left 86 usable sets of ratings for six sessions.

Of the 33 oral presentations evaluated by the 86 raters, 8 of their 471 rating opportunities (1.7%) were missed, because the rater had to depart momentarily for one reason or another. Therefore, we imputed data

for those 8 observations.¹ The imputed scores were distributed as follows: Cost-Effectiveness session holistic rating-1; Health Economics session disaggregated rating-1; Mental Health session disaggregated rating-1; Mental Health session holistic rating-2; Pharmacoeconomics session holistic rating-2; Technology/Outcomes session holistic rating-1.

For each person who performed disaggregated ratings, we calculated the mean rating over the five criteria—this was the re-aggregated rating—and calculated the mean Spearman correlation between that person's re-aggregated rating and the re-aggregated ratings of every other person rating the same session using the disaggregated method. For each person who did holistic ratings, we calculated the mean Spearman correlation between that person's holistic rating and the holistic ratings of every other person rating the same session using holistic ratings. After we calculated the Spearman correlations, we noted that three participants were highly deviant from the other 83 participants; these observations were classified as outliers based on a method suggested by Tukey, and they were excluded from analyses (as cited in Hogg & Tanis, 2001). (Two were disaggregated raters; one was a holistic rater.) The data from the remaining 83 subjects are presented in Table 1.

Because the appropriate significance test to compare the holistic and disaggregated ratings within each session (Palachek & Schucany, 1984) required the use of the Spearman correlations, we performed the above calculations. However such correlations are not normally distributed. Therefore, Table 1 contains two additional measures of association to help summarize the data.² First, we used an r-to-z transformation on the Spearman correlations between each person's ratings and the ratings of every other person in that group.³ The mean of each group's z-scores was then converted back to a correlation, and these correlations are presented in Table 1. Second, we calculated the intraclass correlation (ICC) (Shrout & Fleiss, 1979) among each group's raters.

As can be seen in the Table 1, in four of the six sessions the disaggregated method was superior, and in two of the sessions the holistic method was superior. However, in only three of the six sessions were the

Table 1. Inter-rater reliabilities within each group as represented by mean spearman rank-order correlations, mean r-to-z transformed Spearman correlations, and intraclass correlations (ICC)

Session	Disaggregated ratings				Holistic ratings			
	<i>n</i>	Spearman	r-to-z	ICC	<i>n</i>	Spearman	r-to-z	ICC
Opening plenary*	10	0.43	0.50	0.36	8	0.01	-0.02	-0.01
Health economics*	6	0.59	0.64	0.35	9	0.02	0.04	-0.04
Cost-effectiveness	8	0.16	0.19	0.11	7	0.27	0.33	0.25
Mental health	4	0.51	0.57	0.48	4	-0.16	-0.34	-0.07
Pharmacoeconomics*	4	0.85	0.88	0.79	5	0.52	0.58	0.49
Technology-outcomes	10	0.03	0.05	0.12	8	0.29	0.35	0.31

*Within these sessions the disaggregated ratings had significantly higher inter-rater reliability than did the holistic ratings ($p < 0.06$).

¹We imputed the missed ratings according to the following strategy. For each person who had an omitted rating, we used all of the other ratings completed by that person to ascertain how many standard deviations above or below the average that rater was on these completed ratings compared to other raters within that group and that session. This z-score was then used to impute the rating that person would have given to the rating which had been omitted. For example, if a holistic rater provided ratings which were one standard deviation above the mean of the ratings given by other members of the holistic group within that session for the presentations during which all members were present, then we imputed a rating for the missed presentation which was one standard deviation above that group's mean rating for that particular oral presentation.

²We would like to thank the reviewers of this article for providing several useful suggestions for additional analyses.

³Because an r-to-z transformation of a correlation of 1.0 is undefined, whenever two raters correlated 1.0 we changed this correlation to 0.99.

differences significant, and all three favored the disaggregated method. The mean Spearman correlation for the disaggregated method in the Opening Plenary Session was 0.43, 95% CI (0.33, 0.54); for the holistic method it was only 0.01, 95% CI (-0.13, 0.16). Using the method suggested by Palachek and Schucany (1984), which corrects for the non-independence of the correlations within a group, we ascertained the former was significantly greater than the latter, $t(40) = 2.60$, $p < 0.05$. In the Health Economics Session, the disaggregated method again produced significantly higher Spearman intercorrelations, 0.59, 95% CI (0.40, 0.78) versus 0.02, 95% CI (-0.17, 0.21): $t(18) = 2.51$, $p < 0.05$. Finally, in the Pharmacoeconomics Session, the disaggregated method had higher Spearman intercorrelations 0.85, 95% CI (0.74, 0.96) than did the holistic ratings 0.52, 95% CI (0.32, 0.73), $t(6) = 1.87$, $p < 0.06$. To compare the disaggregated raters in all six sessions with the holistic raters in all six sessions we used the mean r-to-z transformed interrater correlations within each of the 12 groups (6 sessions \times 2 modes of evaluation). Weighting these 12 means by the number of raters in their respective groups, the correlation for the disaggregated method 0.44, 95% CI (0.33, 0.56) exceeded that of the holistic method 0.18, 95% CI (0.09, 0.27), $t(81) = 3.90$, $p < 0.001$. (These are the mean z-scores transformed back into correlations.)

We also examined the dispersion of the ratings for each individual presentation. Because each re-aggregated rating represents a mean of the disaggregated component ratings, the re-aggregated ratings should have a smaller standard deviation than the holistic ratings due to the natural statistical advantage associated with the averaging process. For each oral presentation (34 total), we calculated the standard deviation of holistic ratings and the re-aggregated ratings. For 28 of the 34 conference presentations, the re-aggregated ratings had a smaller standard deviation than did the holistic ratings ($p < 0.01$, sign test).

In order to examine the inter-rater reliability of the individual criteria used by the disaggregated raters, we used r-to-z transformations of all inter-rater Spearman correlations within each group and averaged across groups. Those mean z-scores, transformed back to correlations, are presented in Table 2. As can be seen, the inter-rater reliability was highest for the criterion of 'innovation.' The inter-rater reliabilities of the other four criteria were similar and more modest.

Finally, in addition to the traditional comparisons within rating groups, we examined the inter-rater correlations between the holistic and the disaggregated raters. If the disaggregated rating procedure was greatly superior to the holistic one, then the intercorrelations between the disaggregated raters and the holistic raters might be higher than the intercorrelations among just the holistic raters, despite the fact that the raters contributing to the latter correlation matrix would have the advantage of a common method. We compared the mean holistic-holistic (H-H) intercorrelation in the six sessions, each session contributing one unweighted datum to the mean (0.16) with the mean disaggregated-holistic (D-H) intercorrelation in the six sessions, each session contributing one unweighted datum to the mean (0.13). (These are the mean transformed z-scores converted back to correlations.) The D-H inter-rater reliability did not exceed the H-H inter-rater reliability, while the D-D inter-rater reliability was superior to both (0.43). Apparently, whatever advantage the disaggregated method enjoys is not enough to overcome the advantage of the common evaluation method reflected in the H-H intercorrelations.

Table 2. Inter-rater reliability of the individual criteria used by the disaggregated rating groups in both experiments

Criterion	Experiment 1	Experiment 2
Significance	0.27	0.34
Methods	0.21	0.18
Results	0.29	0.25
Conclusions	0.23	0.42
Innovation	0.45	

Note: Only four criteria were used in Experiment 2.

EXPERIMENT 2

Method*Participants*

The participants were 27 psychology graduate students from the Ohio State University. The graduate students were in various stages of their graduate training (pre-masters, post-masters, and doctoral candidates) and were specializing in a variety of areas of psychology.

Materials

Each of 9 undergraduate honors student posters was evaluated by the graduate student raters using either a holistic or disaggregated rating method.

Separate rating sheets were created for the holistic and disaggregated raters. Both groups were asked to use the following set of criteria to evaluate the posters: 1. Significance: Is the topic significant? Has a convincing case been made that the topic is important? 2. Methods: Are the methods scientifically sound? 3. Results: Are the results presented in enough detail and in an understandable way? 4. Conclusions: Do the conclusions follow from the results? Are they justified? The following instructions were presented to both groups: 'Here are the criteria to be included in your rating of each presentation. On each scale merely circle the rating number which reflects your opinion.'

The holistic rating form required the raters to make one overall rating for each poster using the four criteria. The posters were evaluated using a 9-point rating scale anchored at "poor" and "excellent" with "average" as the mid-point (5) of the scale. The title of the poster and the name of the student author were given above each of the 9-point rating scales.

The disaggregated ratings sheets used the same set of instructions and the same four criteria. However, the raters using the disaggregated method were asked to give four separate ratings to each poster; the 4 ratings corresponded to the four criteria (each criterion was rated on a separate 1–9 scale). The rating scale was identical to those included in the holistic rating form. For the disaggregated ratings forms the title of the poster and the student author were provided above each of the nine sets of ratings.

Two versions of each rating form were created. The first version presented the posters in alphabetical order by student author, while the second version listed the posters in reverse alphabetical order.

Procedure

The participants were recruited via e-mail. The e-mail was sent to all graduate students in the Psychology Department at The Ohio State University except those in the quantitative section. The quantitative students were excluded due to their possible familiarity with the research hypothesis. Additionally, a small subset of the social psychology graduate students was disqualified because they had participated in a similar previous experiment. The graduate students were told that they would be paid \$50 to participate in an experiment in which they would be asked to rate posters presented at the psychology undergraduate research colloquium. Twenty-eight graduate students responded to the e-mail. They were randomly assigned to either the holistic or disaggregated rating groups.

Eighteen posters were presented at the psychology undergraduate research colloquium. An e-mail was sent to each of the 18 student authors requesting their permission to allow their posters to be used in our study. Permission was received from nine of the authors. On the day of the research colloquium, 27 of the 28 graduate student volunteers participated in the experiment. Each graduate student retrieved their packet of materials at the beginning of the colloquium, which included a set of instructions and a rating form. After the graduate students rated each of the posters, the rating sheets were personally returned, and the graduate students each received \$50 for their participation.

Results

For the holistic raters, a Spearman rank-order correlation was calculated between their set of ratings and the ratings of the other holistic raters. An average of these correlations was computed for each holistic rater. In order to compare the disaggregated raters in the same manner, we needed to integrate their four ratings to create one overall rating, just as we had done in Experiment 1. Therefore, for each disaggregated rater, their four ratings (significance, methods, results, and conclusions) were averaged together to obtain one re-aggregated rating. Then, as with the holistic raters, the re-aggregated ratings scores from each individual were correlated (using Spearman rank-order correlations) with the ratings of every other disaggregated rater.

In examining the average correlations for each of the group members, it was noted that one participant in the disaggregated group was on average correlated -0.38 with the other members of the disaggregated group! Again, using the criterion suggested by Tukey, we excluded the ratings of this participant from the analyses. This left 13 graduate students supplying ratings for each group (total $n = 26$). The disaggregated group's average Spearman rank-order correlation after the removal of the 14th participant was 0.43 , 95% CI ($0.38, 0.49$); the average correlation for the holistic group was 0.32 , 95% CI ($0.27, 0.42$). Using the method suggested by Palachek and Schucany (1984), we found that inter-rater reliability of the disaggregated group did not differ significantly from that of the holistic group. We again point out that the rank-order correlations between all members within one group are not independent, and the Palachek and Schucany (1984) method corrects for non-independence. Due to this correction, a large difference between groups is required in order to obtain statistical significance.

As we did in Experiment 1, we also calculated the r-to-z transformed correlations for the two groups (0.47 for the disaggregated group and 0.36 for the holistic group) as well as the ICC for the two groups (0.50 for the disaggregated group and 0.30 for the holistic one).

We also examined the dispersion of the ratings; for 7 of the 9 posters the disaggregated ratings had a smaller standard deviation than did the holistic ratings. This was not significant. However, the variance of the disaggregated group's intercorrelations was significantly smaller than the variance of the holistic group's intercorrelations ($p < 0.05$, Levene's test).

As we did in Experiment 1, we examined the inter-rater reliability of the four criteria used in the disaggregated ratings. Those data are presented in Table 2. Note that the innovativeness of the research was not one of the criteria used by disaggregated raters in Experiment 2. Note also that the inter-rater reliability of the methodology was lowest in Experiment 2 as it had been in Experiment 1.

As we did in Experiment 1 we compared the H-D and H-H correlations; both comparisons resulted in correlations of 0.36 , while the D-D correlations were again superior to both (0.43). (Again, these represent means of the z-scores transformed back to correlations.)

GENERAL DISCUSSION

Using scientific material as the target stimuli, we found that, on average, disaggregated ratings were more reliable than holistic ratings of presentations at the 2001 convention of the Society for Medical Decision Making. Furthermore, despite the extremely small sample size, not an uncommon characteristic of field research, the disaggregated ratings fostered significantly higher inter-rater reliability than did holistic ratings in three of the six sessions. In Experiment 2, graduate students using the disaggregated rating method provided more reliable ratings than did those using holistic ratings when evaluating posters at an undergraduate research colloquium, thus replicating the pattern of the first experiment. However, the small sample size and the substantial statistical correction required by the Palachek and Schucany (1984) method constrained our ability to find a statistically significant result.

A second finding, and an additional advantage of using a disaggregated method, is that the ratings contain significantly smaller amounts of dispersion. Between the two experiments, for 35 of the 43 scientific

stimulus materials, disaggregated ratings had smaller standard deviations than holistic ratings. In this case, a reduction of dispersion can be viewed as a reduction of measurement error. In one sense, this result is compelled by the fact that the disaggregated raters had their ratings of each criteria amalgamated. Any extreme rating given to one criterion would be “diluted” by ratings given by the same rater to other criteria upon which that same oral presentation would be evaluated. Any extreme rating given by a holistic rater would not have any other ratings on that oral presentation with which the extreme rating might be combined and diluted. However, proponents of the holistic rating method might argue that this should not be an advantage for the disaggregated method, because holistic raters should have the capacity and resolve to mentally combine all of the criterion ratings as they were asked to do. As a result, the explicit amalgamation advantage enjoyed by the disaggregated raters would be implicitly duplicated by the holistic ones. Nonetheless, the fact that the disaggregated ratings had less dispersion than the holistic ones suggests that the holistic raters were not attending to or combining their consideration of all of the criteria. Therefore, empirically demonstrating the smaller dispersion of disaggregated ratings is a useful addition to the argument that disaggregated ratings are superior to holistic ratings.

We also should note that in Experiment 1 we were allowed to perform the experiment only on papers accepted for presentation at the convention. Thus the range of quality was necessarily restricted compared to the situation in which we could have evaluated all submitted presentations, not just those deemed of sufficient merit to be accepted for presentation. Due to this range restriction, a significant difference between the methods was more difficult to detect. In addition, the posters evaluated in Experiment 2 were subject to a different type of range restriction: the grad students rated only those posters whose honors student authors were confident enough in the quality of their posters to agree to have them evaluated. Although this probably caused some restriction in terms of the range of poster quality, Experiment 2's range restriction was probably less than that of Experiment 1.

Finally, disaggregated ratings enjoy the practical advantage of flexibility during re-aggregation. In these two experiments, each of the criteria was equally weighted in the re-aggregation process. However, one could differentially weight the multiple criteria depending upon the goal of a specific project. For example, as Lord and Novick (1968) have demonstrated, the criteria can be weighed to maximize the reliability of the re-aggregation. Furthermore, if an objective criterion existed, one could weigh the multiple criteria to maximize predictive validity. Additionally, weights could be constructed to reflect the perceived importance of the criteria as deemed by appropriate experts. Therefore, in many real-world tasks, the flexibility of the disaggregated rating system could result in an increase in reliability above and beyond what is presented in this article.⁴

While we have demonstrated that using a disaggregated rating scale is beneficial in that it can increase inter-rater reliability and decrease the amount of error, in neither study did we directly assess the comparative validity of the ratings. Although it would be ideal to use validity to demonstrate the disaggregated method's superiority, it is practically impossible to identify a ‘gold-standard’ for these types of scientific stimuli. However, due to the relationship between the two constructs—reliability constrains validity—reliability can be useful proxy for validity.

Finally, an intriguing question is why the holistic ratings were higher (albeit not significantly so) on two occasions in Experiment 1. We can offer only a speculation, which we term the “showmanship hypothesis.” Suppose that one speaker presented a paper characterized by very controversial conclusions. However, he or she had truly dazzling multi-media visual aids. The disaggregated rater might dutifully give appropriate ratings to the official five criteria, quality of visual aids not being among them. Because the disaggregated raters might disagree on various aspects of the controversial conclusions, inter-rater reliability might be modest.

⁴Perhaps some readers fear that evaluators in our studies might have “fudged” their ratings of the disaggregated dimensions in order to achieve a desired overall score. Any raters who did this would have worked against our hypothesis, making the disaggregated-holistic difference less likely to be significant.

The holistic raters, however, are not asked to rate individually each of the official criteria. According to the showmanship hypothesis, all of the holistic raters are therefore less likely to concentrate on the official criteria and are more likely than the disaggregated raters to have their single rating heavily influenced by the brilliant visual aids of the speaker. As a result the holistic raters show a higher inter-rater reliability, although the ratings are not based upon the appropriate criteria. Note that the showmanship hypothesis is reminiscent of the “Dr. Fox study” (Naftulin, Ware, & Donnelly, 1973), in which an engaging but otherwise deficient speaker was given a very high evaluation. If such a speaker can attract reliable holistic ratings, and if the unauthorized but salient criterion “entertainment value” can be more easily disregarded by those performing disaggregated ratings, then the holistic procedure may manifest higher reliability.

Given the stupendous magnitude of the national investment in scientific and medical proposals, consideration should be given to using disaggregated rating scales during proposal review, due to the fact that this method is likely to foster higher reliability. Of course, it would be prudent to test our recommendation more thoroughly by comparing the two types of ratings using scientific materials in many different subject domains and evaluation contexts. Given the extreme difficulty in finding a consensus gold standard for scientific merit, testing our recommendation by using reliability rather than validity may be the most practical course of action.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation, Program in Decision, Risk, and Management Science, award 0109250. We have been looking forward to the opportunity to express our sincere appreciation to the members of the Society for Medical Decision Making who participated in our research and to the Society’s executive board who kindly granted us permission to do the study. For their extremely valuable cooperation we are grateful to the President of the Society, M. G. Myriam Hunink, and to the coordinator for the 2001 convention, Michael J. Barry. We also thank Barbara Mellers for her suggestions on data analysis and Robert Wigton for his formulation of the showmanship hypothesis.

REFERENCES

- Armstrong, J. S., Denniston, W. B., & Gordon, M. M. (1975). The use of a decomposition principle in making judgments. *Organizational Behavior and Human Performance*, *14*, 257–263.
- Cornelius, E. T., III, & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analysis by job incumbents. *Journal of Applied Psychology*, *65*, 155–163.
- Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist*, *26*, 180–188.
- Dawes, R. M. (1977). Predictive models as a guide to preference. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-7*, 355–357.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.
- Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*, 86–106.
- Fisher, G. W. (1977). Convergent validation of decomposed multi-attribute utility assessment procedures for risky and riskless decisions. *Organizational Behavior and Human Performance*, *18*, 295–315.
- Henrion, M., Fisher, G. W., & Mullin, T. (1993). Divide and conquer: Effects of decomposition on the accuracy and calibration of subjective probability distributions. *Organizational Behavior and Human Decision Processes*, *55*, 207–227.
- Hogg, R. V., & Tanis, E. A. (2001). *Probability and statistical inference*. Upper Saddle River, NJ: Prentice-Hall Inc.
- Hora, S. C., Dodd, N. G., & Hora, J. A. (1993). The use of decomposition in probability assessment of continuous variables. *Journal of Behavioral Decision Making*, *6*, 133–147.

- Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications, and issues* (4th ed.). Pacific Grove, CA: Brooks-Cole.
- Kleinmuntz, D. N. (1990). Decomposition and the control of error in decision analytic models. In R. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 107–126). Chicago: University of Chicago Press.
- Kleinmuntz, D. N., Fennema, M. G., & Peecher, M. E. (1996). Conditioned assessment of subjective probabilities: Identifying the benefits of decomposition. *Organizational Behavior and Human Decision Processes*, *66*, 1–15.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- MacGregor, D., Lichtenstein, S., & Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes*, *42*, 303–323.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Morera, O. F., & Budescu, D. V. (1998). A psychometric analysis of the “divide and conquer” principle in multicriteria decision making. *Organizational Behavior and Human Performance*, *75*, 187–206.
- Naftulin, D. H., Ware, J. E., Jr., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, *48*, 630–635.
- Palachek, A. D., & Schucany, W. R. (1984). On approximate confidence intervals for measures of concordance. *Psychometrika*, *49*, 133–141.
- Raiffa, H. (1968). *Decision analysis: Introductory lessons on choices under uncertainty*. Boston: Addison-Wesley.
- Ravinder, H. V. (1992). Random error in holistic evaluations and additive decompositions of multiattribute utility—An empirical comparison. *Journal of Behavioral Decision Making*, *5*, 155–167.
- Saaty, T. L. (1970). *The analytic hierarchy process*. New York: McGraw-Hill.
- Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178–200.
- Shrout, P. E., & Fliess, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Stillwell, W. G., Barron, F. H., & Edwards, W. (1983). Evaluating credit applications: A validation of multiattribute weight elicitation techniques. *Organizational Behavior and Human Performance*, *32*, 87–108.
- Ubel, P. A., & Loewenstein, G. (1997). The role of decision analysis in informed consent: Choosing between intuition and systematicity. *Social Science and Medicine*, *44*, 647–656.
- Von Winterfeld, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge: Cambridge University Press.
- Werner, P. D., Rose, T. L., & Yesavage, J. A. (1983). Reliability, accuracy, and decision-making in clinical predictions of imminent dangerousness. *Journal of Consulting and Clinical Psychology*, *51*, 815–825.

Authors' biographies:

Hal R. Arkes, Ph.D. is a Professor in the Department of Psychology and a Senior Scholar in the Center for Health Outcomes, Policy, and Evaluation Studies at Ohio State University. His primary research interests are in medical and economic decision making.

Victoria A. Shaffer, (Ph.D.) is an assistant professor in the Psychology Department at Wichita State University. Her research interests include economic decision making and the evaluation of holistic versus disaggregated ratings.

Robyn M. Dawes, (Ph.D.) is the Charles J. Queenan, Jr. University Professor at Carnegie Mellon University. His work is in behavioral decision making—specializing in clinical judgment, irrationality (not based wholly on introspection), and cooperation. While deeply suspicious of words as conveyors of (substitutes for?) ideas, he publishes a lot.

Authors' addresses:

Hal R. Arkes, Department of Psychology, The Ohio State University, 24 ON Lazenby Hall, 1827 Neil Avenue Hall, Columbus, OH 43210-1222, USA.

Victoria A. Shaffer, Department of Psychology, Wichita State University, 1845 N. Fairmount, Wichita, KS 67260-0034, USA.

Robyn M. Dawes, Department of Social and Decision Science, Carnegie-Mellon University, Porter Hall, Pittsburgh, PA 15213, USA.