A RAND NOTE

Direct Measurement of Three Stylistic Dimensions of National Institutes of Health Consensus Statements

James P. Kahan, Peter C. Noehrenberg, Hilary H. Farris, Elizabeth M. Yano

RAND

The research described in this report was sponsored by the National Institutes of Health under Contract No. 263-MD-140911.

Published 1992 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

A RAND NOTE

N-3456-NIH

Direct Measurement of Three Stylistic Dimensions of National Institutes of Health Consensus Statements

James P. Kahan, Peter C. Noehrenberg, Hilary H. Farris, Elizabeth M. Yano

Prepared for the
National Institutes of Health

RAND

PREFACE

This Note reports an experiment conducted for the National Institutes of Health (NIH) Office of Medical Applications of Research to see whether direct ratings of the statements of NIH consensus conferences could replicate the factorial structure yielded by a detailed content analysis of those statements. The Note should be of interest to those persons concerned with the NIH Consensus Development Program as well as to students of social psychological measurement methodology.

SUMMARY

The Consensus Development Program of the National Institutes of Health (NIH) publicly evaluates scientific information concerning biomedical technologies. Since 1977, the program has issued over 80 consensus statements on medical topics, spanning virtually the entire universe of health care in the United States.

RAND conducted a major research study of the effect of consensus statements on medical practice; that research included a content analysis of the style of 24 consensus statements. That analysis showed that consensus statements varied on three dimensions labelled (1) *discursive*, (2) *directive*, and (3) *scholarly*. Discursive statements attempt to present a full description of the technology and the context surrounding its use. Directive statements attempt to formulate the findings of the conference in terms of guidelines for health care practitioners. Scholarly statements attempt to review the research on the technology.

NIH is interested in extending the stylistic ratings beyond the original set of 24 to all consensus statements, to see whether any of these characteristics is related to effective communication and technology transfer in the practice of medicine. Because it is not feasible to do a content analysis of every consensus statement, the present study attempts to examine whether the qualities of discursiveness, directiveness, and scholarliness can be directly measured by a straightforward rating scale.

Three raters were asked to independently rate each of the 24 original content-analyzed consensus conference statements as a whole on each of the three dimensions, using a scale of one to five, where one represented a low amount of the dimension and five represented a high amount. The definitions of discursiveness, directiveness, and scholarliness were based on the factors defining those qualities from the earlier content analysis. Following their individual ratings, the raters met to discuss the ratings and reconcile differences of opinion.

The initial, independent ratings by the three raters were found to be only moderately consistent. Of the total of 72 ratings, 50 were within acceptable limits of agreement. Although the correlations between raters for discursiveness varied between 0.65 and 0.76, the correlations for directiveness were lower (ranging from 0.37 to 0.67) and those for scholarliness lower still (0.31 to 0.49). However, in a meeting to discuss differences of opinion, the raters were able to revise their judgments to reach agreement for all but one of the 72 total ratings.

The mean (revised) ratings by the panel were uniformly positively related to the factor ratings obtained from the content analysis, with correlations of 0.79 for discursiveness, 0.57 for directiveness, and 0.60 for scholarliness. However, although these results indicate that the two measurement methods produce ratings that share common definitions, that sharing is only partial; the amount of the variance of the factor scores accounted for by the ratings (R-squared) is over 50 percent only for discursiveness and is closer to 30 percent for the other two dimensions. A further comparison of the content analysis and direct ratings by an alternative method that divided each scale into high, middle, and low thirds showed that the direct rating method could achieve the same results as the factor score method only for the discursiveness dimension.

The conclusion drawn from this experiment is that the direct ratings of discursiveness, directiveness, and scholarliness do not replicate the ratings based on the content analysis, but these results do not necessarily mean that NIH should abandon direct ratings. Arguably, the definitions used by the raters, although different from those of the content analysis, could suffice for the needs of the NIH. The global ratings appear to be of sufficient reliability that they may be used to assess whether or not stylistic qualities of the consensus statements relate to the impact of those statements on medical practice.

If direct ratings are to be obtained, our study suggests that they should be done by group discussion rather than by individual raters. Barring that possibility, the rating task should be preceded by an extensive group discussion of the meaning of each of the terms, as well as examination of a full range of sample ratings (possibly of earlier consensus statements), thus laying out in fairly concrete detail the basis for the ratings. Raters should be at least partially familiar with the entire set of medical technologies to be rated and should be comfortable with the concept of Likert-type rating scales.

ACKNOWLEDGMENTS

We wish to thank Charles Sherman, Deputy Director of the Office of Medical Applications of Research, and David Kanouse of RAND for valuable discussions regarding the design and analysis of this experiment and for their critical reviews of earlier drafts of this Note.

CONTENTS

FIGURES

TABLES

1. INTRODUCTION

The Consensus Development Program (CDP) of the National Institutes of Health (NIH), conducted through the NIH Office of Medical Applications of Research (OMAR), is one of the best established and most visible technology assessment activities in the United States. The program conducts consensus conferences to assess the safety and efficacy of medical technologies. Its stated purpose is "to publicly evaluate scientific information concerning biomedical technologies and arrive at consensus statements that will be useful to health care providers and the public at large and that will serve as contributions to scientific thinking about the technologies under consideration." By fulfilling this purpose, the CDP aims to promote timely adoption of beneficial technologies, encourage the abandonment of obsolete and ineffective technologies in favor of those that are safer or more efficacious, discourage the adoption of technologies that have little value, and inform public policy choices that affect the use of medical technologies. From its inception in 1977 through 1990, the CDP has conducted 83 consensus development conferences, spanning virtually the entire universe of health care in the United States.

A consensus conference is held when OMAR and an institute decide that sufficient interest exists in a topic and that sufficient data exist to conduct an assessment of the topic. A panel is assembled of scientists, medical practitioners, and other interested parties such as consumers who are knowledgeable about the topic. Generally, the panelists are themselves not experts; parties who have taken public positions with regard to the issues that might arise and who might therefore not be regarded as objective assessors are excluded. However, such experts and stakeholders are invited to present their points of view to the panel. The panel meets for about two and a half days to hear invited experts present information on the state of the science and on the safety and efficacy of the procedures, drugs, or devices under consideration. Presentations are followed by open discussion by panelists, speakers, and the audience. When the information has been heard, the panel convenes in executive session to draft a *consensus statement* about the topic, which is widely disseminated. This statement is intended to inform the public and to guide clinical practice and the direction of future research on the topic.

RAND conducted a major research study of the effect of consensus statements on medical practice.[1] As part of that study, the style of 24 consensus statements published

[1]D. E. Kanouse, et al., *Changing Medical Practice Through Technology Assessment*, Association for Health Services Research and Health Administration Press, Ann Arbor, Michigan, 1989.

between 1979 and 1983 was subjected to a sentence-by-sentence content analysis.[2] The content analysis showed that consensus statements varied a good deal in the type of message they seek to convey and the audience they seem to address. Statements varied on three dimensions, which we label here (1) *discursive*, (2) *directive*, and (3) *scholarly*.[3] Directive statements attempt to formulate the findings of the conference in terms of guidelines for health care practitioners. Scholarly statements attempt to review the research on the technology. Table 1 lists the 24 statements analyzed in the RAND study.

To date, the importance of these stylistic dimensions has not received an adequate test. The RAND study found no direct association of any of the dimensions with the effect of

Table 1

24 Consensus Conferences Examined in RAND Content Analysis

Seq.[a]	Year	Title
12	1979	*Antenatal diagnosis*
13	1979	Transfusion therapy in pregnant *sickle cell* disease patients
15	1979	The treatment of *primary breast cancer:* management of local disease
16	1979	*Steroid receptors* in breast cancer
17	1979	*Intraocular lens* implantation
18	1979	*Estrogen use* in postmenopausal women
19	1979	*Amantadine*: Does it have a role in the prevention and treatment of influenza?
22	1980	*Thrombolytic therapy* in thrombosis
23	1980	*Febrile seizures*
24	1980	*Adjuvant chemotherapy* of breast cancer
25	1980	Cervical cancer screening: the *Pap smear*
26	1980	*Endoscopy* in upper GI bleeding
27	1980	Childbirth by *cesarean delivery*
28	1980	Carcinoembryonic antigen (*CEA*) as a cancer marker
29	1980	*Coronary artery bypass* scientific and clinical aspects
30	1981	*Reye's syndrome*: diagnosis and treatment
31	1981	Computed tomographic (*CT*) *scanning* of the brain
32	1982	*Defined diets* and childhood hyperactivity
33	1982	Total *hip joint replacement*
35	1983	*Critical care* medicine
36	1983	*Liver transplantation*
37	1983	Treatment of *hypertriglyceridemia*
38	1983	Precursors to malignant *melanoma*
39	1983	Drugs and *insomnia*

[a]Sequence number is OMAR's identification number.

[b]Short title used in this Note is given in *italics*.

[2]J. P. Kahan, D. E. Kanouse, and J. D. Winkler, "Stylistic Variations in National Institutes of Health Consensus Statements, 1979–1983," *International Journal of Technology Assessment in Health Care*, Vol. 4, No. 2, 1988, pp. 289–304; see also Chapter 2 of Kanouse et al., op. cit.

[3]In the original, "directive" was called "didactic"; the term was changed in the present research when it became apparent that some people—dictionary definitions notwithstanding—embued "didactic" with a negative affect more appropriately associated with "dogmatic."

the statements, but the generalizability of this finding is limited because at the time of that study, hard outcome measures were available for only four conferences. Even though the empirical evidence is lacking, OMAR has treated the stylistic characteristics of consensus conferences as potentially important[4] and has tried (through the issuance of guidelines, etc.) to influence the process of drafting statements so as to produce statements written more clearly and persuasively. NIH is presently interested in extending the stylistic ratings beyond the original set of 24 consensus statements to the 84 consensus statements issued through 1990 and future statements. The purpose of having quantified ratings of these fundamental characteristics of each published statement is to see whether any of these characteristics is related to other, perhaps yet to be developed, measures of effective communication and technology transfer into the practice of medicine. In addition to being useful for correlative studies in the future, the ratings may be useful to describe the products and historical changes in the consensus program, now over 13 years old.

Because it is not feasible to do a content analysis of every consensus statement, the present study attempts to examine whether the qualities of discursiveness, directiveness, and scholarliness can be directly measured by a Likert-type instrument. If such "global" ratings can reliably measure these three dimensions, the instrument could be used as a tool in the assessment of whether quality control of consensus statement style improves the diffusion of conference recommendations and changes clinical practice. Positive results from such an assessment could in turn help OMAR define protocols for future consensus statements.

---

[4]See, e.g., F. Mullan and I. Jacoby, "The Town Meeting for Technology: The Maturation of Consensus Conferences," *Journal of the American Medical Association*, Vol. 254, No. 8, 1985, pp. 1068–1072.

2. METHOD

Three raters were recruited to rate 24 consensus statements on the dimensions of discursiveness, directiveness, and scholarliness. The raters were all RAND staff who work within the Health Policy program; all have a master's degree and are currently in doctoral programs. The raters were not familiar with the original study and were blind to the originally published ratings.

Each rater was asked to rate each conference statement, based on his or her impression of the statement taken as a whole, on each of the three dimensions on a Likert-type scale of one to five, where one represented an absence of the quality and five represented a very high amount. Raters were told that the three qualities are independent in the sense that any consensus statement can have a high or low rating on any one dimension irrespective of the scores on the other dimensions. It is possible that a consensus statement could be high, moderate, or low on all three qualities, as well as any other permutation of scores.

Raters were cautioned that the consensus statements had been written eight to 12 years ago; therefore, some of the findings may have appeared outdated. It was emphasized that the purpose of the ratings was the dimensions of discursiveness, directiveness, and scholarliness, not the medical correctness of the findings.

To eliminate any possible sequencing or chronological effects, the statements were randomly divided into three sets, and each rater read the sets in a different order. Raters could revise ratings of individual statements within a set, but once a new set was begun, could not revise ratings from previously read sets.

Following their individual ratings, the raters met to reconcile differences of opinion. If the three raters agreed or if there was a span of two numbers among the raters (i.e., if two raters agreed and the third differed by one digit), then the ratings were accepted without discussion. If there was a span of more than two digits, the raters discussed their reasons for their assessments and could (but were not required to) alter their judgments in light of their colleagues' reasoning.

The three qualities were defined for the raters on the basis of the factor scores of the content analysis, except that items referring to the nature of the topic (centered on a treatment, a condition, or a public health issue), locus of where the technology was performed (hospital vs. office vs. community based), the type of technology (a device, drug, or procedure), and the state of the medical art (mainstream, developing, or new) were not

incorporated into the definitions. We believed that these items did not reflect characteristics of the conference *statement*, but rather had to do with the choice of topic; therefore, they would be difficult to rate when looking at a consensus statement in isolation. Examining the factor loadings of the original analysis,[1] we noted that the nature of the topic was a component of the factors identified as discursiveness and scholarliness, the type of topic a component of directiveness, and the state of the medical art was a component of scholarliness. In total, characteristics of the topic accounted for one out of five variables defining discursiveness, one out of four variables defining directiveness, and two out of four variables defining scholarliness. Thus, a priori, we might anticipate less success replicating the scholarliness ratings than the other two dimensions.

The specific definitions provided to raters were as follows:

Discursive: Consensus statements with high positive ratings on this quality try to tell a story. That is, they describe the biomedical technology or issue addressed by the conference and its surrounding context. If there are multiple sides to the story, discursive consensus statements will attempt to give space to each side; because of this, consensus statements will offer opinions as well as make recommendations. In general, the longer a consensus statement is, the more discursive it is.

Directive: Highly directive consensus statements have a practical orientation. They present the findings of the conference in terms of guidelines for medical practitioners. The tone of the statement is one of providing information in a direct form, so that the reader can use the information without having to interpret the conclusions of the conference. As a consequence, directive statements tend to be straightforward, focusing less on details and more on results.

Scholarly: Consensus statements with high ratings on this quality are focused on the scientific evidence regarding the topic and on the implications of that evidence for practice or for future research. This leads them to concentrate more on the technological state of the art than on established clinical methods.

---

[1]See Table 6 of Kahan et al., op. cit.

3. RESULTS

Our interest focuses first on the agreement among the raters and then on the relationship of their ratings to the factor scores of the content analysis.

RATER AGREEMENT

Table 2 shows the agreement of the initial, independently done, ratings, by conference quality. The first two rows—unanimity and minimal discrepancy—represent what we believe to be acceptable agreement among the raters. The table shows that for the discursive, directive, and scholarly ratings, 19, 16, and 15 conferences, respectively, had acceptable agreement, for a total of 50 out of 72 overall ratings, which is a 69 percent agreement rate. The remaining rows all have spans of three values or more and were considered unacceptable; these ratings were discussed by the raters in a search for compromise to reach agreement.

A second examination of rater agreement is the correlation between pairs of raters over the 24 conferences. Table 3 presents these correlations separately for each dimension. The table shows that although the correlations for discursiveness indicate marginal but acceptable agreement, the corresponding figures for directiveness and scholarliness, although statistically significantly different from zero, were lower and indicative of a moderate degree of rater agreement that is not within the range of acceptable inter-rater reliability figures. As Table 3 also shows, there was no tendency for two of the raters to agree and for the third to be the outlier; each pair of raters had the highest correlation on one of the dimensions.

The raters discussed the 22 ratings that did not have agreement, presenting their reasoning behind the ratings. The discussion revealed that raters' perceptions of the conferences were more similar than their ratings would indicate but that there was some disagreement about the meaning of the ratings, and in particular, the requirements

Table 2

Inter-Rater Agreement (Initial Individual Ratings)

Agreement Category	Number of Conferences		
	Discursive	Directive	Scholarly
Unanimity	7 (29%)	5 (21%)	2 (8%)
Two raters agree, third one digit off	12 (50%)	11 (46%)	13 (54%)
All raters different, span three digits	4 (17%)	5 (21%)	4 (17%)
Two raters agree, third two digits off	1 (4%)	2 (8%)	3 (13%)
All raters different, span four digits	0 (0%)	1 (4%)	2 (8%)
Total	24	24	24

Table 3

Inter-Rater Correlations, by Dimension

Dimension	Raters 1 and 2[a]	Raters 1 and 3	Raters 2 and 3	Avg.[b]
Discursiveness	0.76	0.65	0.75	0.72
Directiveness	0.64	0.67	0.37	0.57
Scholarliness	0.43	0.31	0.49	0.41

[a]Rater numbers were randomly assigned to the three raters.

[b]The average correlation was obtained by transforming r to Z, taking the mean normal deviate, and retransforming back to r.

necessary to assign an extreme score (rating of one or five). The discussion clarified the understanding of the definitions to such an extent that, although consensus was not a requirement of the discussion, raters revised their judgments sufficiently to reach agreement (the first two categories of Table 2) in all but one instance (the directiveness rating of the CT scanning conference), where the grouping remained "all raters different with a span of three digits." For all of the 72 ratings, including the one without a consensus, the mean rating of all three raters after discussion was taken as the group judgment score. The group judgment scores were used in place of individual ratings in all subsequent analyses.

CORRESPONDENCE OF CONTENT ANALYSIS AND GLOBAL RATINGS

Table 4 shows, for all 24 consensus statements, the original estimated factor scores from the content analysis[1] and the mean ratings from the present study. Figures 1 through 3 present the same information in graphical form. The data show a moderate agreement between the two ratings. This is reinforced by the correlations across conferences: 0.79 for discursiveness, 0.57 for directiveness, and 0.60 for scholarliness. All three of these correlations are significantly different from zero (two-tailed test) at a confidence level of $p < 0.01$. Although these results indicate that the two measurement methods produce ratings that share common definitions, this sharing is only partial; the amount of the variance of the factor scores accounted for by the ratings (R-squared) is over 50 percent only for discursiveness and is closer to 30 percent for the other two dimensions.

In analyses not shown here, we also compared the original ratings of each rater and the mean of the unrevised ratings with the factor scores; in all cases, the correlations obtained were less than for the mean of the revised ratings.

[1]The factor scores in Table 4 are taken directly from Table 7 of Kahan et al., op. cit.

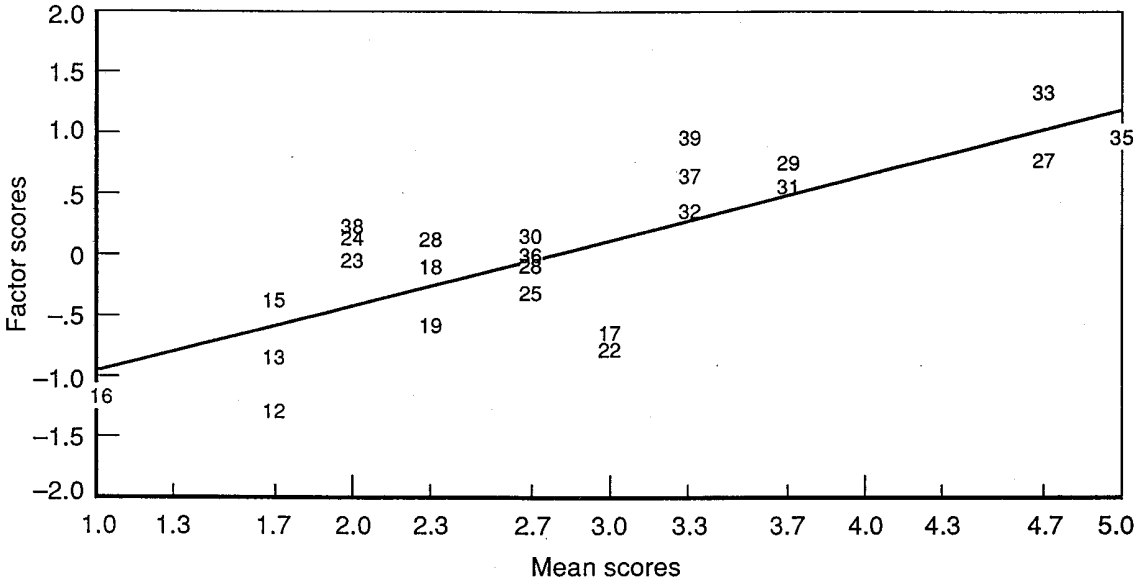
Table 4

Factor Scores and Mean Rating Scores for 24 Consensus Statements

	Discursive		Directive		Scholarly	
Conference	FS	MS	FS	MS	FS	MS
Antenatal diagnosis	–1.31	1.7	0.19	4.7	–0.34	2.0
Sickle cell	–0.86	1.7	–1.19	1.7	1.70	4.3
Primary breast cancer	–0.41	1.7	–0.81	2.3	0.54	3.7
Steroid receptors	–1.19	1.0	–0.42	3.7	–0.18	3.7
Intraocular lens	–0.67	3.0	0.99	3.0	0.37	3.0
Estrogen use	–0.11	2.3	0.44	1.7	0.92	3.3
Amantadine	–0.60	2.3	0.65	4.7	0.08	3.0
Thrombolytic therapy	–0.81	3.0	1.10	4.0	0.31	3.3
Febrile seizures	–0.07	2.0	–0.23	3.7	0.26	4.0
Adjuvant chemotherapy	0.12	2.0	–0.22	2.7	0.50	4.3
Pap smear	–0.33	2.7	–0.09	4.3	–0.98	2.3
Endoscopy	–0.06	2.7	0.94	3.3	–0.61	2.0
Cesarean delivery	0.72	4.7	–0.53	3.0	–0.13	4.7
CEA	0.10	2.0	–1.02	2.3	–1.04	3.7
Coronary artery bypass	0.71	3.7	0.24	3.0	0.08	3.7
Reye's syndrome	0.11	2.7	0.00	3.7	0.14	2.3
CT scanning	0.51	3.7	1.18	4.0	–1.30	1.7
Defined diets	0.31	3.3	–0.78	2.0	1.33	3.7
Hip joint replacement	1.29	4.7	0.32	3.7	–0.39	3.3
Critical care	0.93	5.0	–1.72	2.3	–0.72	1.7
Liver transplant	–0.03	2.7	–0.35	4.0	0.38	3.3
Hypertriglyceridemia	0.59	3.3	0.33	4.3	0.02	3.3
Melanoma	0.16	2.0	0.01	3.0	–0.54	2.7
Insomnia	0.93	3.3	1.00	4.3	–0.40	1.7

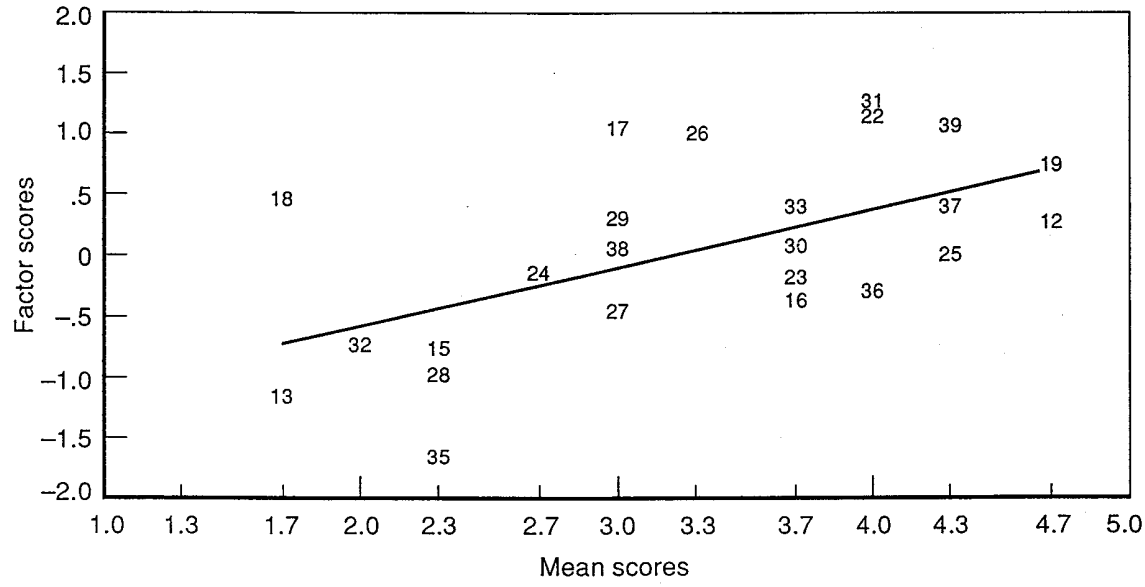
It is possible that the weakness of the correlations between rating methods was because raters were in agreement on the general quality of the dimensions (i.e., whether they were high, medium, or low), but disagreed about the finer degree of the rating. To test this possibility, we conducted an ordinal comparison of the content analysis and direct ratings by dividing each into high, middle, and low thirds on each scale and examining how well the division into thirds on the group ratings matched the corresponding division on the factor scores. These results are presented in Table 5, where it can be seen that fairly good correspondence for the discursiveness dimension, moderately good correspondence for the directness dimension, but poor correspondence for the scholarly dimension were obtained. Although the X^2 values for all three matrices are statistically significant at $p < 0.05$, the three "extreme" differences for directiveness and the large lack of agreement for the medium classification for scholarliness argue against any claim that the two methods of rating were measuring identical concepts.

We also examined the relationships between the dimensions. In the original study, the three dimensions were largely unrelated, with correlations of –0.04 for



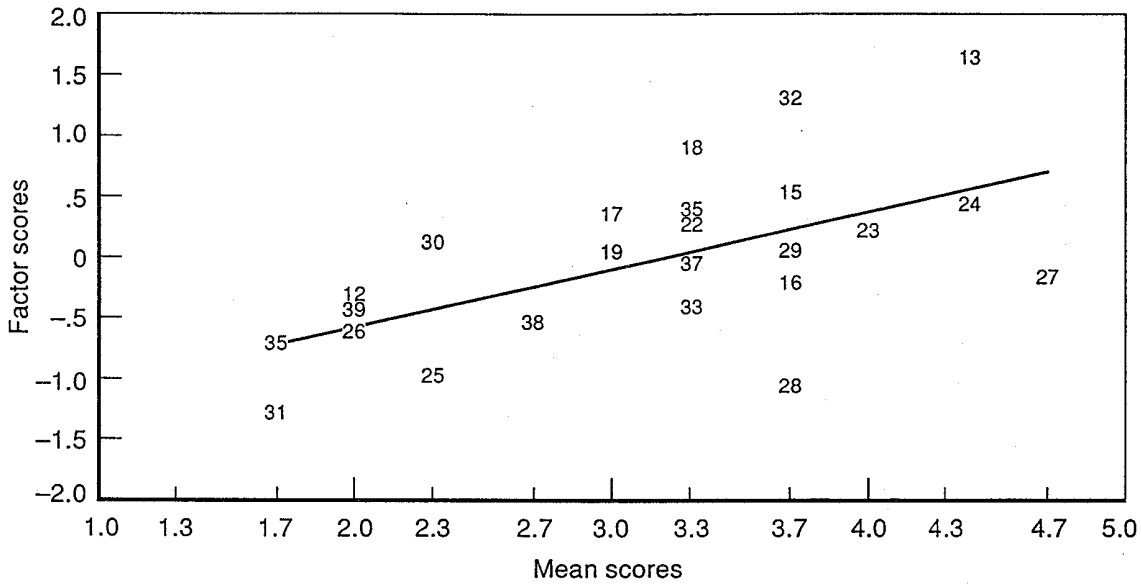
NOTE: Plot entries refer to NIH statement numbers (see Table 1).

Fig. 1—Discursiveness: Scatterplot of Factor Scores and Mean Group Ratings



NOTE: Plot entries refer to NIH statement numbers (see Table 1).

Fig. 2—Directiveness: Scatterplot of Factor Scores and Mean Group Ratings



NOTE: Plot entries refer to NIH statement numbers (see Table 1).

Fig. 3—Scholarliness: Scatterplot of Factor Scores and Mean Group Ratings

Table 5

Correspondence of Factor Scores and Ratings, by Ranked Thirds

	Discursiveness			Directiveness			Scholarliness		
Factor Scores	High	Med	Low	High	Med	Low	High	Med	Low
High	8.0	0.0	0.0	5.0	1.0	2.0	3.6	4.4	0.0
Med	0.0	5.5	2.5	2.0	4.5	1.5	3.6	1.4	3.0
Low	0.0	2.5	5.5	1.0	2.5	4.5	0.8	2.2	5.0

discursiveness/directiveness, –0.29 for discursiveness/scholarliness, and –0.22 for directiveness/scholarliness. The mean ratings of the present study had respective correlations of 0.01, –0.19, and –0.42. Although none of the correlations for mean ratings differs statistically from its content analysis counterpart, the last correlation for the mean ratings is significantly different from zero (two-tailed test) at $p < 0.05$. This indicates, the rating instructions notwithstanding, that the raters saw directiveness and scholarliness as somewhat opposite in nature.

4. DISCUSSION

SHOULD GLOBAL RATINGS BE OBTAINED?

The fairly strong conclusion to be drawn from this experiment is that for 24 NIH consensus statements published between 1979 and 1983, the method of direct ratings of discursiveness, directiveness, and scholarliness do not perfectly replicate the factor analysis ratings based on a content analysis of those statements. The results suggest some commonality but not enough to conclude that these alternative rating methods are measuring the same thing.

In retrospect, a strong replication should not have been expected. The two methods are really very different; whereas the content analysis systematically combines measures of features of the statements that are fairly objective (e.g., number of sentences, type of technology, number of medical recommendations) and were not initially selected to tap the hypothesized dimensions, the global ratings ask for subjective judgments explicitly on those dimensions. The method variance associated with either method suffices to reduce the expected correlations. In addition, a major reason for the lack of correspondence between the two methods is that some of the characteristics of the consensus conferences used in the factor analysis could not be replicated in the ratings. Recall that in the original study, the content categories were supplemented by information about the characteristics of the conference *topic*. These characteristics of the conference topic were not appropriate parts of a rating of the style of the statement and hence did not enter into the definition of the qualities to be rated. For these reasons, then, the two ratings were inherently not the same, although they appear to be measuring somewhat the same things differently.

Do our results mean that NIH should not obtain the global ratings? Not necessarily. The value of the global ratings depends more on whether the ratings capture important dimensions of the statement than on how well they correlate with the ratings developed by content analysis. In the present study, the raters were able, after discussion, to agree on the degree of discursiveness, directiveness, and scholarliness of statements, even if their understanding of those terms was to some degree different from the meaning as defined by the content analysis. Arguably, the definitions used by the raters suffice for the needs of the NIH and could replace the implicit definitions of the content analysis.

The usefulness of ratings, no matter how obtained, is not yet determined. At the time RAND did its evaluation, it was difficult to evaluate the relationship of stylistic qualities of

the statements to their effect. With more consensus statements issued, more time for adoption, and more measures of effect, such an evaluation is both feasible and timely. For example, some statements have recommended changes in practice that should lead to changes in drug prescription patterns, insurance reimbursement claims, or other phenomena that may be measured by publicly available indicators. Global ratings of the appropriate consensus statements may be used to see if there is an effect of the style of the statement; if so, then continued global ratings and the development of a protocol for consensus statement style may be warranted.

HOW SHOULD GLOBAL RATINGS BE OBTAINED?

Our study suggests that if direct ratings are to be obtained, the ratings should be done by group discussion rather than by individual raters, to provide the raters a common definition of the dimensions. Our raters, although all sophisticated in the social sciences and familiar with medical technologies, interpreted the dimensions differently when rating as individuals but were able to achieve a common definition when discussing their ratings. Ratings by a group process will avoid this problem.

If group ratings are not possible, the rating task should be preceded by an extensive group discussion of the meaning of each of the terms, as well as examination of a full range of sample ratings (possibly of earlier consensus statements), thus laying out in fairly concrete detail the basis for the ratings.

Should NIH undertake to rate consensus conferences, there are a number of questions that our study does not provide answers for, but for which we can provide some perhaps useful advice. Because of the effort involved in teaching the ratings scale, it is probably not worthwhile for NIH to have raters assess single conferences. Rather, raters should assess sets of a dozen or more conferences, to have some idea of the range of discursiveness, directiveness, and scholarliness in the statements. It might be a good idea to include in the set of statements to be rated some conferences that have already been rated, to determine whether or not the current set of raters are calibrated with previous raters. Although our study cannot directly recommend *who* should perform the ratings, we can note that our raters could successfully assess the statements. This suggests that the ratings task does not require experienced M.D.s or Ph.D.s. Our observations of the interactions among our raters lead us to believe that raters should be at least partially familiar with the entire set of medical technologies to be rated and should be comfortable with the concept of Likert-type rating scales. It is unlikely that persons with less than a bachelor's degree in the sciences or

social sciences would have this familiarity and comfort. If these suggestions are incorporated into the ratings method, NIH should then have a reasonably consistent set of measures for comparing present and future consensus statements.