1

ARI Research Note 91-82

Effects of Stress on Judgment and Decision Making in Dynamic Tasks

Kenneth R. Hammond and Cynthia M. Lusk
University of Colorado
Center for Research on Judgment and Policy

for

Contracting Officer's Representative
Michael Drillings

Office of Basic Research
Michael Kaplan, Director

June 1991

ARMY RESEARCH INSTITUTE

United States Army
Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON Technical Director	JON W. BLADES COL, IN Commanding

Research accomplished under contract for the Department of the Army

University of Colorado

Technical review by

Michael Drillings

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY -- 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE --	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) --	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Note 91-82

6a. NAME OF PERFORMING ORGANIZATION Center for Research on Judgment and Policy	6b. OFFICE SYMBOL *(If applicable)* --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute Office of Basic Research
6c. ADDRESS *(City, State, and ZIP Code)* Campus Box 344 University of Colorado Boulder, CO 80309-0344		7b. ADDRESS *(City, State, and ZIP Code)* 5001 Eisenhower Avenue Alexandria, VA 22333-5600
8a. NAME OF FUNDING / SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences	8b. OFFICE SYMBOL *(If applicable)* PERI-BR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-86-C-0142
8c. ADDRESS *(City, State, and ZIP Code)* Office of Basic Research 5001 Eisenhower Avenue Alexandria, VA 22333-5600		10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61102B	74F	N/A	N/A

11. TITLE *(Include Security Classification)*
Effect of Stress on Judgment and Decision Making in Dynamic Tasks

12. PERSONAL AUTHOR(S)
Hammond, Kenneth R.; and Lusk, Cynthia

13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM 88/09 TO 89/12	14. DATE OF REPORT *(Year, Month, Day)* 1991, June	15. PAGE COUNT 183

16. SUPPLEMENTARY NOTATION
Michael Drillings, Contracting Officer's Representative

17. COSATI CODES			18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB-GROUP	Judgment Forecasting Decision making Experts stress

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Three empirical studies on judgment and decision making in dynamic tasks were carried out during the period 1 September 1988 to 31 December 1989. Subjects were expert research meteorologists. Topics were forecasting (a) hail, (b) microbursts, and (c) convection initiation (thunderstorms) at an airport approach. Primary findings were as follows:

- in the hail study, meteorologists' forecasts were closely approximated by a weighted-sum model;
- in the microburst study, experts who worked together for years, when tested in work conditions, did not agree on the judgments of principal cues;
- in the convection study, more accurate forecasts were made on high stress than low stress days, thus contradicting the conventional wisdom.

Two annotated bibliographies were produced: the effects of stress on judgment and decision making, and the effects of variation of display formats on judgment and decision making.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Michael Drillings	22b. TELEPHONE *(Include Area Code)* (703) 274-8722	22c. OFFICE SYMBOL PERI-BR

Table of Contents

EFFECTS OF STRESS ON JUDGMENT AND DECISION MAKING IN DYNAMIC TASKS

1 September 1988 - 31 December 1989

Kenneth R. Hammond (Principal Investigator)
and Cynthia M. Lusk
University of Colorado, Boulder

EXECUTIVE SUMMARY

The primary goal of the project was to increase understanding of the effects of stress on judgment and decision making under changing conditions. This report covers work carried out during the period September 1, 1988 to December 31, 1989. The context of the research was aviation weather forecasting at the National Center for Atmospheric Research (NCAR) and the National Oceanic and Atmospheric Administration (NOAA) research sites that provide the necessary circumstances for generalization of results with respect to both (a) professional persons as subjects and (b) conditions involving changing information of sufficient complexity to be of interest to the military. In addition, substantial reviews of the literature on (a) the effects of stress on judgment and decision making and (b) the effects of variation in display formats were carried out and annotated bibliographies constructed.

Empirical studies of these topics of expert judgment and decision making in both static and dynamic tasks were carried out in relation to three weather forecasting problems, (a) hail, (b) microbursts, and (c) the effects of stress on forecasts of convection initiation at an airport approach.

Hail

A paper describing the results of this research was published in the journal *Weather and Forecasting.* The abstract states:

> This study compared meteorologists, an expert system, and simple weighted-sum models in a limited-information hail forecasting experiment. It was found that forecasts made by meteorologists were closely approximated by an additive model, and that the model captured most of their forecasting skill. Furthermore, the additive model approximated the meteorologists' forecasts better than the expert system did. Results of this study are consistent with the results of extensive psychological research on judgment and decision making processes. Potential implications are discussed.
>
> See Appendix I for full report.

Microbursts

Two manuscripts describing the work were completed; one was sent to a psychological journal, the other to a meteorological journal. Both were accepted conditionally upon some revision, now underway. A paper describing the microburst research was presented at an international aviation weather conference.

Abstract from Manuscript Submitted to Psychological Journal

> The major goals of this research are to (a) study professionals engaging in dynamic, and thus representative, task conditions, (b) apply lens model theory to these dynamic conditions, (c) learn how judgments are changed in response to changing conditions, and (d) utilize a hierarchical judgment model to investigate the judgment process from perception of data to final judgment. The results indicate that (a) agreement regarding secondary cue values is modest, not because of

differences in perception of primary cue values but because of differences in inferences drawn from them, (b) magnitude of agreement for each secondary cue is related to the proximity of the cue to primary cue data, and (c) agreement in probability judgments is higher when secondary cue values are specified. There was some evidence to suggest an increase in agreement among forecasters' judgments as more information relevant to judgment was received. Finally, increased information over time resulted in more extreme probability judgments for half the forecasters.

See Appendix II for full report

Abstract from Manuscript Submitted to Meteorological Journal

Two studies of microburst forecasting were conducted in order to demonstrate the utility of applying theoretical and methodological concepts from judgment and decision making to meteorology. A hierarchical model of the judgment process is outlined in which a precursor identification phase is separated from the prediction phase. In the first study, forecasters were provided with specific, perfectly reliable precursor values and were asked to provide judgments regarding the probability of a microburst. Results indicated that the microburst forecasts were adequately represented by a linear model. Modest agreement was observed among the forecasters' judgments. In the second study forecasters viewed storms under dynamic conditions representative of their usual operational setting. They made judgments regarding precursor values, as well as the probability of a microburst occurring. The forecasters' agreement regarding microburst predictions was found to be even lower than in the first study. In addition, agreement regarding microburst predictions was found to be even lower than in the first study. In addition, agreement regarding the (subjectively) most important precursor value was near zero. These results suggest that opportunities to improve forecasting would result from a better understanding of the precursor identification and prediction phases of the forecasting process.

See Appendix III for full report

Convection Initiation and the Effects of Stress

A field study of the effects of high and low stress on expert meteorologists forecasting convection at Stapleton Airport in Denver, Colorado was carried out. The summary states:

> In sum, all of the results presented indicate a decrement in performance on low stress (activity) days compared to high stress (activity) days. The bias measures (Table 4) indicate that the decrement may be due, in part, to larger judgmental biases occurring during low stress days. In addition, there is some evidence (Table 6) that forecasters use a higher criterion (β) under low stress than high stress conditions. More research is necessary to clarify and expand these findings. Although the present data indicate forecasters may introduce bias into their judgments or a different decision criterion may be operating on low stress days, the processes accounting for the differences are unknown.
>
> Note: Further analyses of these data were carried out and a manuscript is now in preparation for publication.
>
> See Appendix IV for the full report.

Field Study of the Effects of Stress on the Use of Various Information Displays, Cognitive Processes and Accuracy of Inference

In order to provide baseline data for a field study of the effects of stress in naturalistic conditions at the National Weather Service forecasting conditions office in Denver, a study was undertaken of three meteorologists making forecasts of convection (thunderstorms) over four regions and six forecasting occasions during a one hour period of data display under changing conditions. These data have been collected and analysis is underway. However, the proposal submitted to ARI was not approved for funding, thus eliminating our study of the effects of stress.

Annotated Bibliography for the Effects of Stress on Judgment and Decision Making (Revised)

A letter of inquiry about recent research was sent to 60 authors whose work was included in the draft bibliography. As a result, 25 new citations and annotations were added and the conclusions in the draft manuscript were updated and revised. This review was included with Report No. 13.

Annotated Bibliography for the Effects of Display Format on Judgment and Decision Making

This annotated bibliography is nearing completion. The goal of this bibliography was to review the literature on the effects of display format on the cognitive processing of that information. The annotations are complete and the final touches on the bibliography are in process, including the writing of an introduction. A major portion of the annotations are included as Appendix V.

Overall Conclusions

1. The methodology used to study *static* tasks can be applied to the study of *dynamic* decision making with useful results, a conclusion which has far-reaching methodological consequences.

2. Research on dynamic decision making led to many of the same results found in relation to decision making in static tasks, namely, (a) a difference was found between experts' description of their cognitive activity and their cognitive activity as observed and analyzed by quantitative procedures; (b) only moderate agreement was found within and between expert judges; (c) psychologists were more accurate than expert forecasters in predicting which conditions would enhance accuracy of forecasts.

3. The search of the literature on the effects of stress on judgment and decision making led to the following conclusion:

> No generalization regarding the effects of stress on judgment and decision making can be readily justified on the basis of the articles annotated here. No general principle explaining the effect of stress on judgment and decision making is supported

by a conclusive set of empirical studies. It has not been clearly demonstrated that stress impairs, enhances, or has no effect on cognitive activity. Predictions about the effects of stress on judgment and decision making in specific circumstances cannot be defended by reference to this literature.

4. The results of the field study reported in Appendix IV contradict conventional wisdom; performance improved under stressful conditions. In addition, our analysis showed that although forecasting accuracy (as defined in terms of Signal Detection Theory) improved under stressful conditions, the decision criterion (β) used by the forecasters implicitly changed; although the forecasters were unaware of it the ratio of false positives to false negatives increased. The implications of these results are the topic of a manuscript in preparation.

Publications

Hammond, K. R. (1988). Judgment and decision making in dynamic tasks. *Information and Decision Technologies, 14,* 3-14.

Potts, R., Lusk, Cynthia, Hammond, K., & Stewart, T. (1988). Expert judgment in the nowcasting of microbursts. In *Preprints: Third International Conference on the Aviation Weather System* (pp. 190-195). Boston: American Meteorological Society.

Stewart, T. R., Moninger, W. R., Grassia, J., Brady, R. H., & Merrem, F. H. (1989) Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting, 4,* 24-34.

Technical Reports

* Lusk, C. M., & Hammond, K. R. (1989). *Judgment in a dynamic task: Microburst forecasting* (Tech Rep. No. 288). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

Lusk, C. M., Mross, E. F., & Hammond, K. R. (1989). *Judgment and decision making under stress: A preliminary study of convection forecasts* (Tech Rep. No. 287). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

** Lusk, C. M., Stewart, T. R., Hammond, K. R., & Potts, R. (1988) *Judgment and decision making in dynamic tasks: The case of forecasting the microburst.* (Tech Rep. No. 284). Boulder: University of Colorado, Center for Research on Judgment and Policy.

Mross, E. F., & Hammond, K. R. (1990). *Annotated bibliography for cognition and stress* (Tech Rep. No. 295). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

* Under revision for *Journal of Behavioral Decision Making*
** Under revision for *Weather and Forecasting*

Appendix I

Stewart, T. R., Moninger, W. R., Grassia, J.,
Brady, R. H., & Merrem, F. H.

1989

Analysis of expert judgment in a hail forecasting experiment

Weather and Forecasting, 4, 24-34.

Analysis of Expert Judgment in a Hail Forecasting Experiment

THOMAS R. STEWART,[†,**] WILLIAM R. MONINGER,[*] JANET GRASSIA,[†]
RAY H. BRADY[*] AND FRANK H. MERREM[*]

[*] *Environmental Research Laboratories, National Oceanic and Atmospheric Administration, Boulder, Colorado*
[†] *Center for Research on Judgment and Policy, University of Colorado at Boulder, Boulder, Colorado*

(Manuscript received 15 March 1988, in final form 20 December 1988)

ABSTRACT

This study compared meteorologists, an expert system, and simple weighted-sum models in a limited-information hail forecasting experiment. It was found that forecasts made by meteorologists were closely approximated by an additive model, and that the model captured most of their forecasting skill. Furthermore, the additive model approximated the meteorologists' forecasts better than the expert system did. Results of this study are consistent with the results of extensive psychological research on judgment and decision making processes. Potential implications are discussed.

1. Introduction

The future in weather forecasting is a partnership between person and machine (Snellman 1977; Schlatter 1985; Tennekes 1988), and an understanding of the capabilities and limitations of both is critical to making that partnership effective. Although computer models and algorithms help aggregate weather information for operational forecasters, the human forecaster remains the primary information processor. While a great deal of effort has been devoted to the development of advanced weather forecasting workstations, there has been little study of how forecasters aggregate the information provided by the workstations. The human information processing system is the least understood, yet probably the most important, component of forecasting accuracy.

Human information processing has been a major topic of study by psychologists and others interested in judgment and decision making, and that research has produced a substantial body of knowledge, theories, and techniques that are relevant to the design and implementation of person-machine weather forecasting systems. Three major conclusions drawn from judgment and decision research may have particular relevance for weather forecasting: 1) the results of systematic studies of human information processing yield insights into this process that often contradict people's introspective observations; 2) human information processing is limited and subject to systematic errors and biases; and 3) cognitive assistance can overcome some of the limitations of the judgment process and improve the quality of judgment. For reviews of the research, see Einhorn and Hogarth (1981), Hammond et al. (1980), Hogarth (1980), Sjoberg (1982), Slovic and Lichtenstein (1973), and Slovic et al. (1977).

In this paper we describe an experiment which illustrates how research techniques that have been used by psychologists for over 30 yr can be used to study information processing by weather forecasters. The next section explains how this experiment fits into an overall strategy for investigating the cognitive processes of weather forecasters. Then we describe the experiment, present the results, and discuss the implications.

2. Overview of research strategy

The cognitive processes used in weather forecasting can be divided into three categories: information acquisition, information integration, and output (see Hogarth 1980). Information acquisition is the process of obtaining the information about past and current weather. Each feature of past and current weather (e.g., radar signatures such as reflectivity, rotation, tilt) is a "cue" for the forecast of future weather. Information integration is the activity of assimilating and organizing the cues into a judgment, or set of judgments, about future weather. Output is the process of formulating the forecast into its final form to be issued to the public.

In cognitive psychology, as in most other areas of research, it is necessary to simplify a phenomenon in order to study it. In the present study, we chose to simplify by excluding the perceptual processes involved

[**] Present affiliation: Center for Policy Research, The University at Albany, State University of New York, Albany, New York.

Corresponding author address: Dr. Thomas R. Stewart, Center for Policy Research, Milne 300, The State University of New York—Albany, Albany, New York 12222.

in information acquisition and limiting the forecasters' cognitive activity to information integration and output. As a result, this study concerns only the integration of information to form a forecast, not the perceptual processes involved in acquiring information. Our method (described below) assured that all forecasters in the study used exactly the same information. Consequently, some aspects of forecast skill were necessarily excluded from the study, and a somewhat unrealistic forecasting situation was created because the meteorologists were not able to acquire information as they would in an operational setting.

In the study of complex cognitive processes, there is an inherent trade-off between realism and control that gives rise to a difficult dilemma. We can study cognitive processes in highly realistic situations (e.g., operational forecasters making actual forecasts) where we have very little control, and are therefore not able to draw strong conclusions about the results, or we can conduct controlled studies by introducing constraints (as we did in the present study) so that we can be clear about the results of the experiment, at the expense of introducing doubt about the generality of the results.

The resolution of this dilemma is to include studies representing various points on the realism/control continuum in a research program. When the results of controlled studies are consistent with what is observed in natural settings, we can be confident in our findings. The study to be reported here falls near the low realism/high control end of the continuum. As a result, we can expect to draw relatively clear conclusions about how forecasters integrate information in the experiment ("internal validity") but we must be cautious in generalizing to the cognitive activity of forecasters in operational settings ("external validity"). Despite their limitations, such simplified studies of judgment and decision making have provided important insights into the nature of human cognition (Brown 1972; Kirwan et al. 1983; Dawes 1986). When they are combined with results of more realistic studies (which we have currently planned) the generality of the results can be systematically investigated. Furthermore, when the results of a limited study are consistent with a larger body of theory and research, confidence in generalizations increases. Thus, this study should be viewed as an initial step in the systematic study of human information processing in weather forecasting.

3. Method

Information derived from Doppler radar volume scans of 75 storms was presented to seven meteorologists who then made probability forecasts of hail and severe hail. Two different models, representing alternative ways of describing the meteorologists' subjective judgment processes, were compared with the forecasts. The radar volume scan data, the procedure for obtaining forecasts, and the models are described below.

a. Data

The raw data for the study consisted of 644 Doppler radar volume scans of 156 storms. The data were collected in the summer of 1985 during a forecasting exercise (Haugen 1986) conducted by NOAA's Program for Regional Observing and Forecasting Services (PROFS). The radar was operated by the National Center for Atmospheric Research (NCAR). This radar (CP-2) produced volume scans of reflectivity, Doppler velocity, and differential reflectivity every 5 min, but scans included in the dataset were separated by 10-min intervals. The cues used were determined as part of an earlier project to develop an expert system for hailstorm diagnosis (Merrem and Brady 1988). For that study, a meteorologist (RHB) played back the radar data, and then visually estimated seven cues. The cues were maximum reflectivity at 1) low, and 2) middle levels of the storm, 3) maximum echo gradient within the storm, 4) rotation or convergence within the storm, and 5) tilt of the storm between low and middle levels. The optional cues, which were available for only some of the radar data, were 6) hail signature based on differential reflectivity (ZDR) and 7) upper-level divergence. The severity of each storm was determined from the logs of PROFS chase teams who observed the storms in situ, or from public reports telephoned to the local National Weather Service office. It was necessary to modify the original dataset because data were missing in many volume scans, and only volume scans with complete data could be used in this study. Therefore, upper-level divergence information was not used because it was missing in 67% of the volume scans. In addition, 191 volume scans were dropped because the ZDR signature was not available. The dataset used in this study consisted of six cue variables for the remaining 453 volume scans. Examination of these cases showed they were similar to the original set. The cues and the scoring criteria are listed below.

1) *Reflectivity of core at low level.* From the low-level (0.7 deg) reflectivity PPI scan, estimate the average reflectivity of the storm's core, assuming it consists of at least seven–ten pixels. (Note: In the summer of 1985, a pixel of data displayed on the monitors of the PROFS workstation corresponded to a 500 m × 500 m square.)

2) *Reflectivity of core at middle level.* From the middle-level (6.4 km AGL) reflectivity CAPPI (constant altitude) scan, estimate the average reflectivity of the storm's core, assuming it also consists of at least seven–ten pixels.

3) *Strong echo gradient.* Is there an area of echo (i) at low or middle-levels, (ii) a few kilometers or more in length, and (iii) situated on the SE, S, SW, or advancing flank of the storm where the reflectivity gradient exceeds 8 dBZ km^{-1}?

4) *Tilt.* Comparing the middle-level CAPPI and low-level PPI scans, (i) Is the middle-level high reflectivity core situated over the strong low-level reflectivity

gradient? or (ii) does a horizontal distance of approximately 4 km or more separate the centers of the two cores?

5) *Rotation.* In terms of velocity difference, what is the magnitude of the strongest (cyclonic or anticyclonic) shear or convergence signal observed within the echo at either low or middle levels?

6) *Favorable ZDR signature.* Do the low-level (0.2 deg) differential reflectivity data show a coherent (several pixels) hail signal with this cell?

Verification. In the set of 453 volume scans, either significant (diameter ⩾ 0.25 in. or small hail ⩾ 1 in. deep) or severe (diameter ⩾ ¾ in.) hail was verified within 30 min after 16.1% of the observations, and severe hail was verified after 6.6% of the observations. The problems associated with the verification of severe weather events have been discussed by Hales (1987). Severe storms which track across densely populated urban areas are more likely to be verified as such than are severe storms which remain over sparsely populated rural areas. Although potentially severe storms occurring over rural areas generally had a PROFS chase team assigned to them, it is likely that some of the significant or severe hail events accompanying these storms were not observed by chase teams. In addition, all hail reports were strictly interpreted; i.e., a storm reported as producing hail at 1539 LST was not assumed to be a hail producer at 1540 LST unless it was reported as hailing at the later time. Even though the majority of potential hail-producing storms were observed by chase teams, a few storms were undoubtedly missed. Although we consider our verification dataset to be one of the most complete ever assembled during a real-time forecast experiment, these inherent problems remain.

b. The forecasts

Seven meteorologists made 30-min probabilistic hail forecasts for a sample of 75 volume scans drawn from the original 453. The participants were all research meteorologists who had participated in one or more real-time forecasting experiments using the PROFS workstation. A stratified random sampling procedure was used to select the 75 volume scans to ensure that the base rate (proportion of volume scans for which hail was verified) in the sample matched that in the population of 453 volume scans. Because an error was discovered in the verification data after the study was run, the base rate in the sample turned out to be 14.7% for significant orsevere hail and 5.3% for severe hail only.

On the basis of the six cue variables for each volume scan, the meteorologists estimated probabilities both for any hail (significant or severe) and for severe hail only. Figure 1 illustrates how the volume scans were presented to the meteorologists. For reasons described in section 2, the levels of the cues for each volume scan were specified; i.e., meteorologists did not perceive them directly from the radar display as they would in operational forecasting.

The meteorologists expressed concern about the limited information they were given. They said that to forecast hail they would need additional information, for example, about the evolution of the storm, the storm's relation to the surrounding environment, and its location relative to the radar. We explained that the information provided was determined by the availability of data and that we recognized that forecasting skill exhibited in this study could be substantially different from the skill of forecasters in the field.

The 75 volume scans were presented in random order. After judging the first 50 volume scans, participants took a brief break and then judged the remaining 25 volume scans plus an additional 25 volume scans consisting of the even-numbered volume scans from the first set of 50, presented in random order. Repetition of 25 volume scans makes it possible to assess the consistency of the forecasts. No meteorologist reported noticing the repeated volume scans. All meteorologists evaluated 100 volume scans and filled out a questionnaire about their forecasting strategy in less than 2 h.

c. The models

Cognitive processes can be studied in the same way that other natural processes are studied, i.e., by developing alternative models and evaluating those models. Two information processing models were used in this study, and they were evaluated with regard to two criteria: 1) How well does the model reproduce the judgments of the meteorologists? and 2) How well does the model capture forecasting skill? i.e., How accurately does it forecast hail probability? Each model is described below.

1) MULTIPLE REGRESSION

A technique called "judgment analysis," which uses multiple regression analysis to model the judgments of experts, has been used extensively in psychology (Hammond et al. 1975; Stewart 1988). The effectiveness of this technique is based on a pervasive finding in research on judgment and decision making: in many domains of expertise, simple algebraic models can be used to reproduce the judgments of experts (Slovic and Lichtenstein 1973). Often a simple linear model works as well as or better than more complex models (Dawes and Corrigan 1974).

Using judgment analysis, models of the following form were statistically fit to the forecasts made by each meteorologist:

$$\begin{aligned} Y_{ij} = c_j + b_{j1}X_{i1} + b_{j1}^{*}(X_{i1})^2 + b_{j2}X_{i2} \\ + b_{j2}^{*}(X_{i2})^2 + b_{j3}X_{i3} + b_{j4}X_{i4} \\ + b_{j5}X_{i5} + b_{j5}^{*}(X_{i5})^2 + b_{j6}X_{i6} + e_{ij}, \end{aligned}$$

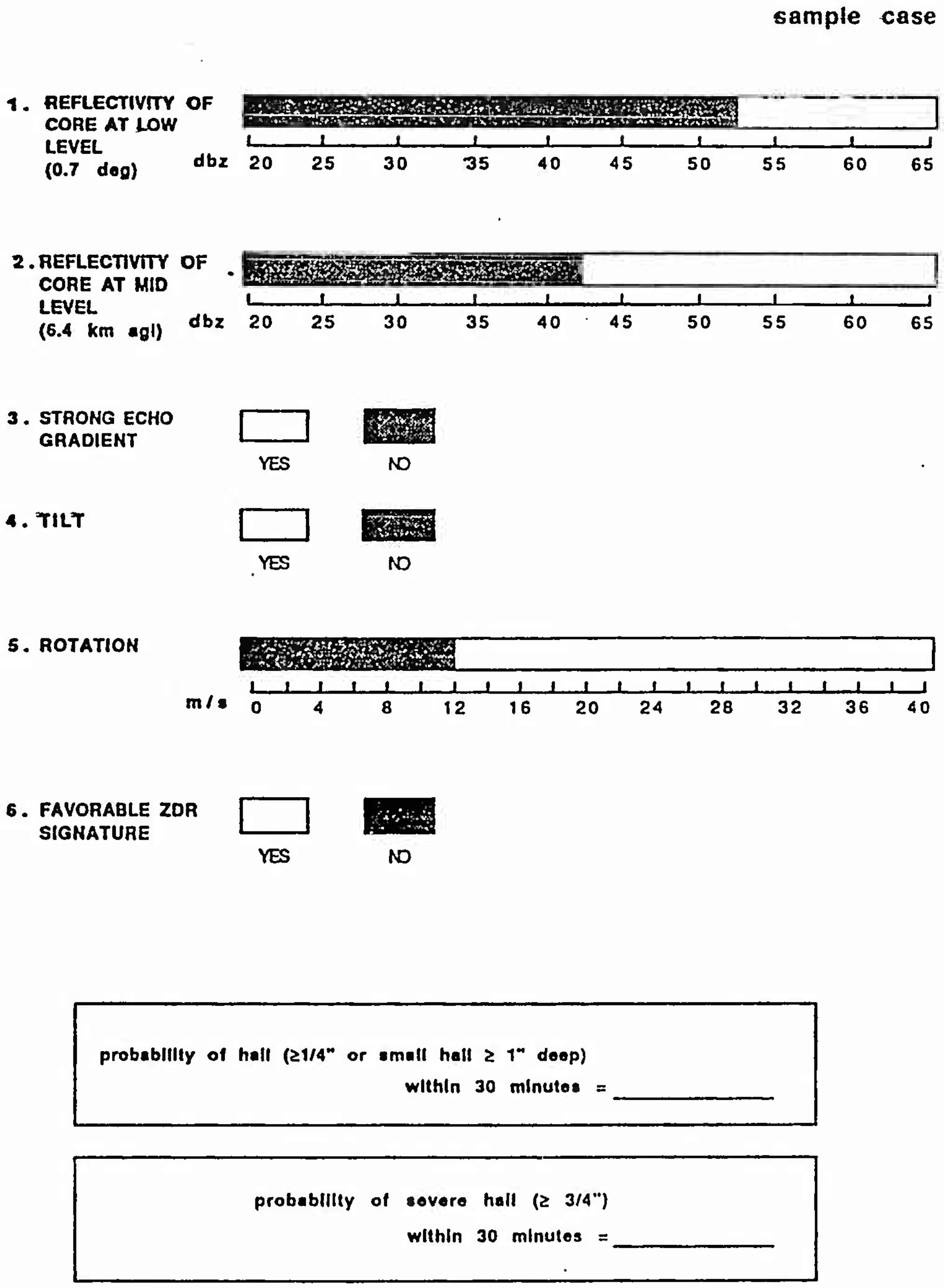

sample case
1. REFLECTIVITY OF CORE AT LOW LEVEL (0.7 deg)
dbz
20 25 30 35 40 45 50 55 60 65
2. REFLECTIVITY OF CORE AT MID LEVEL (6.4 km agl)
dbz
20 25 30 35 40 45 50 55 60 65
3. STRONG ECHO GRADIENT
YES
NO
4. TILT
YES
NO
5. ROTATION
m/s
0 4 8 12 16 20 24 28 32 36 40
6. FAVORABLE ZDR SIGNATURE
YES
NO
probability of hail (≥1/4" or small hail ≥ 1" deep) within 30 minutes = ______
probability of severe hail (≥ 3/4") within 30 minutes = ______

FIG. 1. Sample of a representation of a volume scan.

where

Y_{ij} the forecast made by meteorologist j based on volume scan i
c_j a constant for meteorologist j
b_{jk} the weight for cue k
b^*_{jk} the weight for the square of cue k
X_{i1} the low-level reflectivity for volume scan i
X_{i2} the middle-level reflectivity for volume scan i
X_{i3} the strong echo gradient for volume scan i (0 = no, 1 = yes)
X_{i4} the tilt for volume scan i (0 = no, 1 = yes)
X_{i5} the rotation for volume scan i
X_{i6} the ZDR for volume scan i (0 = no, 1 = yes) and
e_{ij} the residual for meteorologist j on volume scan i

The parameters (c_j, b_{jk}'s and b^*_{jk}'s) of the model were determined so that the sum of the squared differences between the predictions of the model and the actual forecasts were a minimum; that is, for meteorologist j, the sum of the $(e_{ij})^2$ over all of the cases is minimized.

The squares of low- and midlevel reflectivity and rotation were included in the model because plots of the meteorologists' judgments vs these cues suggested that most meteorologists used them in a nonlinear fashion, particularly when they judged the probability of severe hail. The plots indicated that, in many cases, the slope of the curve relating probability forecasts to cue values increases as the cue increases, as if the meteorologists were using the cues exponentially. This occurred much more frequently for the low- and middle-level reflectivity cues than for rotation. This may reflect meteorologists' awareness that dBZ, the measure of reflectivity, is a logarithmic scale. The quadratic approximation to the exponential was used because, in an additive model, the use of exponential transformations of the cues results in a statistically intractable model.

The correspondence between the statistical model and the actual forecasts is given by the multiple correlation (R), which can range from 0 to 1, with 1 indicating perfect fit. The squared multiple correlation (R^2) indicates the proportion of variance of the forecasts that is accounted for by the model.

2) Expert system

The goal of research on expert systems has been the development of computer programs that can emulate the behavior of experts. Expert systems contain a knowledge base that can be thought of as a model of how the expert aggregates information. Thus, an expert system is a model of human information processing. For reviews of expert systems research, see Waterman (1986) or Winston (1984). The relation between research on expert systems and judgment and decision research has been discussed by Hammond (1987a), Stewart and McMillan (1987), and Carroll (1987).

An expert system called HAIL, developed by Merrem and Brady (1988), was used in this study. HAIL consists of 250 rules based on the seven cue variables described in section 3a. Input to the system is provided by an experienced meteorologist. Output consists of statements ordered from 1 to 5 (see Table 1). In addition to diagnosing the presence of hail, the system provides information about the possibility of tornadoes and strong winds. As is typical of expert systems, the 250 rules were derived by discussion with only one person. The rules were designed to represent as closely as possible the thinking process used by the chosen expert meteorologist as he diagnoses storm severity. Since development of an expert system is extremely time consuming, it was not possible to develop one for the other meteorologists in the experiment.

Since the meteorologists made 30-min probability forecasts whereas HAIL was designed to provide categorical diagnoses of hailstorms, it was necessary to transform the output of HAIL so that it could be compared with the probability forecasts. This transformation was accomplished by computing the relative frequency, in the original 453 volume scans, of hail or severe hail within 30 min, given each categorical output (Table 1). These relative frequencies, which are estimates of the conditional probability of hail given the diagnosis, were substituted for the categorical diagnoses. In other words, the output of HAIL was calibrated with respect to the 453 volume scans in the original dataset, and thus was converted from categorical diagnoses into probability forecasts. This procedure makes it possible to validate HAIL's forecasts as probability forecasts (Murphy 1986).

TABLE 1. Calibration of the HAIL expert system.

Diagnosis category*	Number of times given	Number of occurrences of hail within 30 min	Probability of hail, given diagnosis
		Any Hail	
1	251	10	.040
2	60	14	.233
3	75	26	.347
4	34	12	.353
5	33	11	.333
		Severe Hail	
1	251	4	.016
2	60	8	.133
3	75	7	.093
4	34	3	.088
5	33	8	.242

* Description of diagnosis categories: 1) This storm is not significant and not severe. Hail of any size and/or gusty winds are very unlikely. 2) There is a very low probability that this cell may be producing small hail (<¾ in.) and/or moderately strong wind gusts (35–49 kt). 3) This storm is a significant weather producer with small hail (<¾ in.) and/or gusty (35–49 kt) winds. 4) This storm is a significant weather producer wth small hail (<¾ in.) and/or gust (35–49 kt) winds. There is the possibility that it may also be severe with large (≥¾ in.) hail and/or strong (≥50 kt) winds. 5) This storm is severe with large hail (≥¾ in.) and/or strong (≥50 kt) winds.

4. Results

Three types of results are discussed here. First, we describe characteristics of the meteorologists' forecasts. How well do they agree, how consistent are they, and how accurate are they? Then we report on the correspondence between the regression models and the expert-system model and the meteorologists' forecasts. Finally, we compare the accuracy of the meteorologists and the models in order to determine how much of the meteorologists' skill is captured in the models.

a. *The meteorologists' forecasts*

1) Agreement

Correlations among the seven meteorologists' forecasts (A–G) are presented in Table 2. (Correlations

Table 2. Agreement among meteorologists.

Forecast	A	B	C	D	E	F	G
			Any Hail				
A	(.93)*						
B	.91	(.95)					
C	.86	.83	(.89)				
D	.85	.88	.87	(.95)			
E	.90	.88	.91	.84	(.93)		
F	.84	.88	.75	.79	.77	(.92)	
G	.88	.89	.82	.85	.86	.84	(.95)
	Range .75–.91			Median .86			
			Severe Hail				
A	(.97)						
B	.93	(.96)					
C	.87	.88	(.92)				
D	.84	.90	.95	(.95)			
E	.86	.86	.80	.78	(.69)		
F	.88	.92	.82	.86	.78	(.94)	
G	.87	.90	.84	.85	.92	.85	(.93)
	Range .78–.95			Median .86			

* Numbers in parentheses are estimates of consistencies based on 25 repeated trials.

can range from −1.0 to +1.0.) Agreement among meteorologists was moderate to high for both hail and severe hail forecasts. For forecasts of any type of hail, meteorologist F has the lowest level of agreement with other forecasters, but this is not the case for forecasts of severe hail.

2) Consistency

The numbers in parentheses in the diagonal of Table 2 are estimates of the consistency of each meteorologist's forecasts. A meteorologist who made exactly the same forecasts on repeated presentations of the same information would have a consistency of 1.0. Consistency is estimated by correlating the two sets of judgments of 25 repeated volume scans. The forecasters are not perfectly consistent, but their consistency is generally high except for meteorologist E's forecasts of severe hail. His low consistency is due to a few pairs of repeated volume scans for which he gave two quite different probabilities. In one volume scan, his first forecast was 10% and his second was 50%. If this volume scan were eliminated, his consistency would be 0.82.

3) Performance

Skill scores, squared correlation coefficients, conditional biases, and unconditional biases for each forecaster are presented in Table 3. These indices are described in Murphy (1988). The skill score reported in Table 3 is

$$SS = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(\langle x\rangle, x)],$$

where $\mathrm{MSE}(f, x)$ is the mean square error for the forecast (f) relative to the observed event (x) and $\mathrm{MSE}(\langle x\rangle, x)$ is the mean square error for a constant forecast of $\langle x\rangle$ which is the climatological probability of hail in the sample. This measure reflects the accuracy of the forecasts relative to a reference forecast. The maximum skill score is 1.0, and if the MSE for the forecast is equal to the MSE for the climatological forecast, skill is 0.0.

Squared correlations between forecast probabilities and dichotomous variables representing the occurrence of hail and severe hail (0 = no hail, 1 = hail) are also reported in Table 3. The correlation between a probability forecast and a dichotomous verification variable is a point biserial correlation [see Edwards (1976) for a discussion of the properties of this correlation coefficient] and can range from −1.0 to 1.0. This correlation measures the extent to which forecast probabilities are consistently higher when hail occurs than when it does not. The correlation would be 1.00 if 1) the forecast probability were always p_1 when hail occurred, 2) the forecast probability were always p_2 when hail did not occur, and 3) $p_1 > p_2$, regardless of the values of p_1 and p_2. The correlation will be small when variation in the forecasts, given occurrence or nonoccurrence of hail, is large relative to the total variation in forecasts. It is not sensitive to the actual probabilities or to their range; i.e., a forecaster who always gave probabilities between 0.10 and 0.20 could have the same correlation as another forecaster whose probabilities ranged from 0.50 to 1.00. The correlation measures the ability of the forecast to discriminate consistently between occurrence and nonoccurrence of hail. It does not measure "bias," i.e., the extent to which the magnitudes of the forecast probabilities are appropriate for the weather events being forecast. Two kinds

Table 3. Skill scores, correlation, and bias.

Forecaster	Skill score	Squared correlation	Conditional bias	Unconditional bias
		Forecasts of Any Hail		
A	.046	.233	.079	.108
B	−.340	.181	.114	.408
C	.064	.177	.034	.079
D	−.881	.206	.048	1.039
E	.080	.219	.074	.065
F	−1.018	.125	.331	.811
G	−.704	.154	.264	.594
		Forecasts of Severe Hail		
A	.087	.211	.098	.025
B	−.245	.162	.259	.149
C	−.155	.074	.205	.024
D	−.586	.091	.466	.211
E	−.015	.092	.094	.013
F	−.730	.128	.618	.240
G	−.849	.119	.624	.344

of bias identified by Murphy (1988) are reported in Table 3. "Conditional bias" is related to the slope of the regression line relating observed events to forecasts. Conditional bias is zero only when the slope is 1.0. "Unconditional bias" is related to the difference between the mean forecast and the mean event. It is zero only when these two means are equal. Murphy showed that the skill score is equal to the squared correlation coefficient minus the sum of the two bias terms. He pointed out that since the bias terms cannot be negative, the correlation coefficient might be considered a measure of the "potential skill" that might be attained if all conditional and unconditional biases were eliminated.

Most skill scores in Table 3 are negative and the maximum improvement over climatology is only 8.7%. The correlation coefficients, hqwever, indicate that forecasters were able to distinguish between hail- and nonhail-producing storms to some degree. All correlations were positive and significantly different from 0.0 at the 0.01 level of significance. The low skill scores are due to high levels of conditional and unconditional bias. Thus, Table 3 suggests that meteorologists can *potentially* improve over climatology by more than 20%, but they do not achieve that level of improvement because of biases in the forecast.

TABLE 5. Relative weights of cues.

	Cue†					
Forecaster	LDBZ	MDBZ	GRAD	TILT	ROT	ZDR
			Any Hail			
A	.21*	.24*	.13*	.03	.17*	.22*
B	.22*	.27*	.09*	.05	.19*	.17*
C	.17*	.36*	.17*	.08	.12*	.10*
D	.28*	.30*	.10	.14*	.09	.09*
E	.18*	.47*	.07	.04	.10	.14*
F	.30*	.14	.04	.02	.34*	.17*
G	.19*	.42*	.00	.07	.15*	.17*
			Severe Hail			
A	.12	.36*	.21*	.00	.25*	.06
B	.30*	.30*	.08	.01	.22*	.08*
C	.11	.40*	.23*	.09	.07	.10
D	.28*	.28*	.17*	.10*	.13*	.05
E	.14	.68*	.12	.00	.05	.00
F	.26*	.17*	.13*	.00	.33*	.12*
G	.16*	.59*	.00	.04	.15*	.06

* Significant at the .01 level.

† LDBZ reflectivity of core at low level; MDBZ reflectivity of core at midlevel; GRAD strong echo gradient (yes, no); TILT tilt (yes, no); ROT rotation or convergence (m s^{-1}); ZDR favorable ZDR signature (yes, no).

b. Models of the meteorologists' forecasts

1) REGRESSION ANALYSIS

Table 4 presents squared multiple correlations that have been adjusted to correct for overfitting of the regression model due to the number of predictors relative to the number of volume scans. They indicate that the regression models account for 80%–92% of the variance in the meteorologists' forecasts. In other words, these simple weighted-sum models can reproduce the forecasts with a high degree of accuracy and account for nearly all the consistent variation in forecasts. (See Table 2 for proportion of variance that is consistent for each forecaster.)

This result may seem puzzling because the meteorologists invariably reported that their judgment processes involved nonadditive, synergistic aggregation of information. The ability of the regression model to describe meteorologists' information aggregation processes is consistent, however, with the research on human judgment cited in section 3.

TABLE 4. Adjusted squared multiple correlations for regression models of forecasts.

Forecaster	Any hail	Severe hail
A	.90	.84
B	.92	.91
C	.86	.81
D	.89	.86
E	.89	.80
F	.83	.90
G	.87	.91

Regression models can be used to infer how the meteorologists weigh information when they make forecasts. Relative weights of the cues, derived from the regression models, are presented in Table 5 (see the Appendix for derivation of weights). These weights are useful because they can explain, in part, why different meteorologists arrive at different forecasts. In this study, the cues were moderately intercorrelated (Table 6), and, as a result, the weights must be interpreted with caution. The weights that are significantly different from zero (at the 0.01 level of significance) are indicated in the table.

Although the weights differ among meteorologists, they indicate that low- and midlevel reflectivity are generally the most important cues. The notable exception is meteorologist F. For both hail and severe hail, rotation is F's most important cue.

Actual agreement among meteorologists (Table 2) is greater than would be expected based on the differ-

TABLE 6. Cue intercorrelations.

	LDBZ	MDBZ	GRAD	TILT	ROT	ZDR
LDBZ	1.00	.60	.62	.28	.41	.32
MDBZ	.60	1.00	.49	.33	.49	.28
GRAD	.62	.49	1.00	.21	.50	.27
TILT	.28	.33	.21	1.00	.20	.06
ROT	.41	.49	.50	.20	1.00	.19
ZDR	.32	.28	.27	.06	.19	1.00

ences between the weights. This occurs because the cue intercorrelations (Table 6) are all positive. When cues are intercorrelated, different weighting strategies can produce similar forecasts because the cues provide partially redundant information. In this circumstance, agreement among forecasts may be considered "false agreement" (Hammond et al. 1975) because it does not reflect agreement in the underlying forecasting strategy; i.e., there is agreement in fact but not in principle. In the relatively infrequent volume scans when cues diverge, i.e., when some cues indicate hail while other cues indicate no hail, disagreements among meteorologists will emerge. Thus, meteorologists can be expected to disagree most when forecasting is most difficult.

2) The expert system

Correlations between the HAIL expert system and the meteorologists' ranged from 0.70 to 0.85 for forecasts of any hail and from 0.63 to 0.79 for forecasts of severe hail. For all meteorologists, the weighted-sum judgment analysis models reproduced meteorologists' forecasts better than did the HAIL expert system. This includes the forecasts of the meteorologist who developed the rule base for HAIL.

c. Performance of the models

1) Regression models

To what extent do the regression models of the meteorologists capture the accuracy in their forecasts? To answer this question, the regression models described above were applied to the 75 volume scans to produce forecasts. Performance of these models is described in Table 7. Only the correlation coefficients which, as described above, indicate the potential skill of an unbiased forecast, are reported here. In the case of the regression model, unconditional bias of the model is identical to that of the forecaster. Changes in conditional bias reflect changes in the correlation coefficient and in the variance of the forecasts.

Table 7. Performance of forecasts and models of forecasts (correlations).

Forecaster	Original forecasts	Regression models
	Any Hail	
A	.48	.41
E	.47	.45
D	.45	.43
B	.43	.42
C	.42	.45
G	.39	.43
F	.35	.37
	Severe Hail	
A	.46	.37
B	.40	.37
F	.36	.35
G	.34	.37
E	.30	.35
D	.30	.34
C	.27	.34

The models capture most of the (potential) skill in the forecasts for six of the seven meteorologists. Only meteorologist A substantially outperforms the model that is based on his judgments.

The rows of Table 7 have been ordered from highest to lowest correlation of the original forecasts to highlight a pattern in the data. For the least accurate meteorologists, the model outperforms the original forecasts; but for the most accurate meteorologists, the model does worse than the original forecasts. Thus, differences in performance among the models are less than the differences among the original forecasts. This suggests that some (small) component of accuracy (or inaccuracy) may not be captured by the regression models. Whether that component is simply chance (lucky or unlucky forecasts) or a systematic, synergistic process remains to be determined in further research.

The small differences among the correlation coefficients for different regression models in Table 7 also reflect a "flat maximum" effect (Lovie and Lovie 1986; von Winterfeldt and Edwards 1982) due to intercorrelations among the cues. When cues are intercorrelated, it may not matter much how the information provided is integrated into a forecast as long as it is done in a reasonable and consistent fashion. In the hail data used in this study, the cues were intercorrelated (Table 6), the relations between the cues and the probability of hail were all monotonic, and, given the data provided, there was a high degree of uncertainty about whether a storm would produce hail. These are all contributing factors to the flat maximum effect.

For any task with these properties, a weighted-sum model will perform about as well as any other model, and the magnitudes of the weights do not matter much as long as they have the correct sign (Dawes and Corrigan 1974). Researchers have found that the weighted-sum model generally outperforms humans for these kinds of tasks because the model is perfectly consistent whereas the human is not (Goldberg 1969, 1970; Camerer 1981). The model proves superior even though it does not include complex interactions among the cues, or "synergisms," which are important to human experts.

2) Expert system

For forecasts of any hail, the correlation for HAIL is 0.38, slightly above the lowest correlation for a meteorologist. For severe hail forecasts, the correlation for HAIL is 0.41, which is near the level of the best meteorologist and slightly better than his regression model.

5. Discussion

This study illustrated how the subjective component of forecasting can be systematically studied. The design of the experiment made it possible to investigate the following characteristics of the forecasts:

- *Agreement.* Agreement among forecasts was moderately high in this study. Lack of agreement (see Lusk et al. 1988, for example) may indicate that some forecasters are inconsistent or that they are using different forecasting strategies.
- *Consistency.* If the forecasting process is consistent, then identical conditions produce identical forecasts. If the forecasting process is not consistent, then there is a degree of arbitrariness about the forecasts that will reduce their accuracy. In this simple experiment, the forecasts were highly consistent. In general, as the amount of information and the complexity of a task increases, consistency decreases. This fact suggests that forecasts in the field may be less consistent than those in this experiment.
- *Descriptive model.* Statistical regression models provided good descriptions of the forecasts. Furthermore, the regression models were generally as accurate as the original forecasts. In comparison with a complex expert-system model, the regression models provided better approximations to the meteorologists' forecasts and were just as accurate.
- *Parameters of judgment models.* It is useful to describe judgment processes in terms of weight, function form, and organizing principle (Hammond et al. 1975). Weights reflect the relative importance of different items of information. The weights estimated in this study (Table 5) indicated that different meteorologists attached different importance to the cues. Function forms describe the relation between each cue and the forecast. In this study, the reflectivity cues and rotation were related to the forecasts by an exponential function form. The organizing principle governs the way that the various cues are organized into an overall forecast. The organizing principle implicit in the regression models is additive. The expert system employs a nonadditive, synergistic organizing principle. In this study, the additive organizing principle provided the best approximation to the meteorologists' forecasts.

Further research is needed to determine the generality of the results found in this study. In particular, studies involving more realistic forecasting situations are necessary. It must be stressed, however, that our results are consistent with a large body of research and theory in judgment and decision making. It is likely, therefore, that they can be applied to some situations that arise in operational forecasting.

6. Conclusion

The importance of studying the subjective judgment processes involved in weather forecasting is supported by the work of Allen (1982), Allen et al. (1986), and Allan Murphy and his colleagues (e.g., Murphy and Winkler 1971; Murphy and Brown 1984). Our study has shown that research methods used by psychologists to study human judgment processes can be applied to weather forecasting. The experiment suggests that the intuitive processes that weather forecasters use to aggregate information into a forecast can be analyzed and described in quantitative terms.

A number of interesting and important forecasting questions can be addressed using systematic methods borrowed from judgment and decision research. For example, how do novice and experienced forecasters differ with regard to consistency, relative weights, function forms, and organizing principle? What is the effect of advanced workstations on the forecaster's judgment processes? Does additional information reduce the consistency of forecasts, and, if so, how can consistency be increased? Can feedback about judgment parameters be used to improve forecasting skill (Hammond et al. 1975; Hammond 1987b)? How much of the skill of expert forecasters can be captured by computers?

Continued research on cognitive processes in weather forecasting is likely to prove useful in the design of "person-machine" systems for weather forecasting. Design of such systems must be based on realistic views of both machine and human capabilities. Through research in computer science and artificial intelligence, machine capabilities are being expanded. Through the study of human information processing in weather forecasting, we are gaining an understanding of the human judgment process.

Acknowledgments. Special thanks are due the expert meteorologists who participated in this study. C. Bullock, E. Szoke, and L. Murdock provided helpful comments on an earlier version of this paper. We are indebted to K. Hammond, J. McGinley, and S. MacDonald for their support and encouragement. We also wish to thank D. Walker and M. McCoy for providing advice on forecast evaluation and M. Luhring and D. Victor for preparing the manuscript.

This research was supported by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142 and by the following branches of the National Oceanic and Atmospheric Administration: the Artificial Intelligence Program; the Program for Regional Observing and Forecasting Services, Environmental Research Laboratories; and the Office of Meteorology, National Weather Service. The views,

opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

APPENDIX

Calculation of Relative Weights

Regression weights do not indicate the relative importance of the cues because (a) the cues are expressed in different units and (b) there are two weights for each of the continuous cues because the squared terms were included in the regression analysis. The following procedure was used to calculate the relative weights listed in Table 5.

1) For each forecast, regression weights for the model described in section 3c were computed.

2) For forecast j, the continuous cues (low-level reflectivity, midlevel reflectivity, and rotation) were transformed as follows:

$$f_k(X_{ik}) = b_{jk}X_{ik} + b^*_{jk}X^2_{ik}, \qquad \text{for} \quad i = 1\text{–}75 \quad \text{and} \quad k = 1, 2, 5.$$

This transformation combines the two terms for the continuous cues into a single term.

3) A second regression analysis was computed using the three transformed cue variables and the three binary cues to predict the forecast. The regression equation was

$$Y_{ij} = c_j + b_{j1}f_1(X_{i1}) + b_{j2}f_2(X_{i2}) + b_{j3}X_{i3} + b_{j4}X_{i4} + b_{j5}f_5(X_{i5}) + b_{j6}X_{i6} + e_{ij}.$$

This form of the regression equation has only one weight for each cue. It is a simple algebraic transformation of the original regression equation, and the R^2s were identical to those obtained in the original analysis.

4) The regression weights for the standardized form of the regression equation (the beta weights) were summed, and each beta weight was divided by that sum. (The standardized form of the regression equation compensates for differences in units by transforming each variable so that its mean is 0.0 and its variance is 1.0 in the sample.) This calculation gave the relative weights presented in Table 5.

Several methods have been proposed for computing relative weights in judgment analysis. Alternative methods are discussed in Darlington (1968) and Stewart (1988).

REFERENCES

Allen, G., 1982: Probability and judgment in weather forecasting. Preprints, *Ninth Conference on Weather Forecasting and Analysis*, Seattle, Amer. Meteor. Soc., 1–6.

——, V. Ciesielski and W. Bolam, 1986: Evaluation of an expert system to forecast rain in Melbourne. Paper presented at the *First Australian Artificial Intelligence Congress*, Melbourne, 11 pp.

Brown, T. R., 1972: A comparison of judgmental policy equations obtained from human judges under natural and contrived conditions. *Math. Biosci.*, 15, 205–230.

Camerer, C. F., 1981: General conditions for the success of bootstrapping models. *Org. Behav. Human Decis.*, 27, 411–422.

Carroll, B., 1987: Artificial intelligence, Expert systems for clinical diagnosis: Are they worth the effort? *Behav. Sci.*, 32, 274–292.

Darlington, R. B., 1968: Multiple regression in psychological research and practice. *Psych. Bull.*, 69, 161–182.

Dawes, R. M., 1986: Representative thinking in clinical judgment. *Clin. Psych. Rev.*, 6, 425–441.

——, and B. Corrigan, 1974: Linear models in decision making. *Psych. Bull.*, 81, 95–106.

Edwards, A. L., 1976: *An Introduction to Linear Regression and Correlation.* Freeman, 213 pp.

Einhorn, H., and R. M. Hogarth, 1981: Behavioral decision theory: Processes of judgment and choice. *Ann. Rev. Psych.*, 32, 53–88.

Goldberg, L. R., 1969: The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behav. Res.*, 4, 523–536.

——, 1970: Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psych. Bull.*, 73, 422–432.

Hales, J. E., 1987: An examination of the National Weather Service severe local storm warning program and proposed improvements. NOAA Tech. Memo., NWS NSSFC-15, 32 pp.

Hammond, K. R., 1987a: Toward a unified approach to the study of expert judgment. *Expert Judgment and Expert Systems*, J. L. Mumpower, O. Renn, L. D. Phillips and V. R. R. Uppuluri, Eds., Springer-Verlag, 1–16.

——, 1987b: Annotated bibliography on cognitive feedback. University of Colorado, Center for Research on Judgment and Policy, Tech. Rep. 269, 28 pp.

——, T. R. Stewart, B. Brehmer and D. O. Steinman, 1975: Social judgment theory. *Human Judgment and Decision Processes*. M. F. Kaplan and S. Schwartz, Eds., Academic Press, 271–312.

——, G. H. McClelland and J. Mumpower, 1980: *Human Judgment and Decision Making: Theories, Methods, and Procedures.* Praeger, 258 pp.

Haugen, D. A., 1986: The PROFS RT85 forecast exercise. Reprints, *Eleventh Conference on Weather Forecasting and Analysis*, Kansas City, MO, Amer. Meteor. Soc., 335–339.

Hogarth, R. M., 1980: *Judgement and Choice: The Psychology of Decision.* Wiley, 250 pp.

Kirwan, J. R., D. M. Chaput de Saintonge, C. R. B. Joyce and H. L. F. Currey, 1983: Clinical judgment in rheumatoid arthritis. I: Rheumatologists' opinions and the development of "paper patients". *Ann. Rheum. Dis.*, 42, 644–647.

Lovie, A. D., and P. Lovie, 1986: The flat maximum effect and linear scoring models for prediction. *J. Forecasting*, 5, 159–168.

Lusk, C. M., T. R. Stewart and K. R. Hammond, 1988: Judgment and decision making in dynamic tasks: The case of forecasting the microburst. Unpublished manuscript, Center for Research on Judgment and Policy, University of Colorado at Boulder.

Merrem, F. H., and R. H. Brady, 1988: Evaluating an expert system for forecasting. *Proc. Fourth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Anaheim, Amer. Meteor. Soc., 259–261.

Murphy, A. H., 1986: Comparative evaluation of categorical and probabilistic forecasts: Two alternatives to the traditional approach. *Mon. Wea. Rev.*, 114, 245–249.

——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, 116, 2417–2424.

——, and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, 52, 239–247.

——, and B. G. Brown, 1984: A comparative evaluation of objective and subjective weather forecasts in the United States. *J. Forecasting*, 3, 369–393.

Schlatter, T. W., 1985: A day in the life of a modern mesoscale forecaster. *ESA Journal*, 9, 235–256.

Sjoberg, L., 1982: Aided and unaided decision making: Improving intuitive judgment. *J. Forecasting*, 1, 349–363.

Slovic, P., and S. Lichtenstein, 1973: Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Human Judgment and Social Interaction*, L. Rappoport and D. A. Summers, Eds., Holt, Rinehart & Winston, 16–108.

——, B. Fischhoff and S. Lichtenstein, 1977: Behavioral decision theory. *Ann. Rev. Psych.*, 28, 1–39.

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Amer. Meteor. Soc.*, 58, 1036–1044.

Stewart, T. R., 1988: Judgment analysis: Procedures. *Human Judgment: The Social Judgment Theory View*, B. Brehmer and C. R. B. Joyce, Eds., North-Holland, 519 pp.

——, and C. McMillan, 1987: Descriptive and prescriptive models for judgment and decision making: Implications for knowledge engineering. *Expert Judgment and Expert Systems*, J. L. Mumpower, O. Renn, L. D. Phillips, and V. R. R. Uppuluri, Eds., Springer-Verlag, 305–320.

Tennekes, H., 1988: The outlook: Scattered showers. *Bull. Amer. Meteor. Soc.*, 69, 368–372.

von Winterfeldt, D., and W. Edwards, 1982: Costs and payoffs in perceptual research. *Psych. Bull.*, 85, 267–273.

Waterman, D. A., 1986: *A Guide to Expert Systems*. Addison Wesley, 419 pp.

Winston, P. H., 1984: *Artificial Intelligence*. Addison Wesley, 524 pp.

Appendix II

Lusk, C. M., & Hammond, K. R.

1989

Judgment in a dynamic task: Microburst forecasting

(Tech Rep. No. 288)

Boulder, CO: University of Colorado, Center for Research on Judgment and Policy

JUDGMENT IN A DYNAMIC TASK: MICROBURST FORECASTING

Cynthia M. Lusk and
Kenneth R. Hammond

Center for Research on Judgment and Policy

Institute of Cognitive Science

University of Colorado
Boulder, Colorado 80309

Report No. 288

July 1989

This work was supported in part by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142, Work Unit Number 2Q161102B74F. The views, opinions, and findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

Judgment in a Dynamic Task: Microburst Forecasting

The major goals of this research are to (a) study mature professionals engaging in dynamic, and thus representative, task conditions, (b) apply lens model theory to these dynamic conditions, (c) learn how judgments are changed in response to changing conditions, and (d) utilize a hierarchical judgment model to investigate the judgment process from perception of data to final judgment. The results indicate that (a) agreement regarding secondary cue values is modest, not because of differences in perception of primary cue values but because of differences in inferences drawn from them, (b) magnitude of agreement for each secondary cue is related to the proximity of the cue to primary cue data, and (c) agreement in probability judgments is higher when secondary cue values are specified. There was some evidence to suggest an increase in agreement among forecasters' judgments as more information relevant to a judgment was received. Finally, increased information over time resulted in more extreme probability judgments for half the forecasters.

KEYWORDS: lens model, dynamic tasks, experts, forecasting

Virtually all of the research on judgment and decision making has been restricted to studying the behavior of immature subjects in restricted laboratory conditions involving static, unchanging task conditions. But many, if not most, important judgments and decisions are made by mature professionals in response to changing task conditions. The major goals of the research reported here are (1) to remedy these limitations by studying professional experts engaging in a complex, dynamic task conditions representative of their normal working conditions, (2) to ascertain whether the results from lens model theory and research in static tasks generalize to these circumstances, (3) to learn how judgments are changed in response to changing conditions, and (4) to investigate the judgment process from the perception of data to the final judgment through the use of a hierarchical model. The context of the research was severe weather forecasting, specifically, the short-term forecasting (0-30 minutes; Roberts and Wilson, 1989) of microbursts (brief, localized windstorms that are a potentially fatal hazard to aircraft).

The Microburst Forecasting Process

Weather forecasting in general, and microburst forecasting in particular, offers an opportunity to investigate the entire judgment process because it involves (a) the visual perception of data from numerous sources, (b) the assessment of the significance of those data as a determinant of the final judgment, and (c) the aggregation and integration of all the information, including intermediate inferences, to arrive at (d) a final forecast—all with respect to information that is changing over time.

The opportunity for studying the cognitive activity of forecasters coping with a dynamic task under representative conditions immediately raises the question of the generalizability of theory and results from previous lens model studies of judgment in static tasks. Several such studies were carried out with the same forecasters who were studied under dynamic task conditions prior to the present study, one of which is particularly relevant and is described below as Study 1 (Lusk, Stewart and Hammond, 1988). In addition, a hierarchical lens model was constructed to trace out the judgment process from data perception to final judgment.

A Hierarchical Model of the Microburst Forecasting Process

A hierarchical lens model depicting the steps between the storm environment and a judgment about microbursts at a given time is presented in Figure 1. This framework is derived from social judgment theory (Hammond, Stewart, Brehmer, and Steinmann, 1975; Brehmer and Joyce, 1988), which describes the relationship between two systems: the task system in the environment and the cognitive system of the decision maker. The environment of the microburst forecasting task is represented as Phases A, B, and C in Figure 1, which is an adaptation of Brunswik's lens model (Brunswik, 1956; Hammond, et al., 1975; Brehmer and Joyce, 1988.) Phase A represents the physical mechanisms that underlie the weather phenomenon at Phase B. The weather produces objective radar data at Phase C. The cognitive system of the forecaster begins operating at the link between Phases C and D. After reading the perceptual data corresponding to the primary cues at Phase D, the forecaster must infer values of the secondary cues (hypothesized precursors of microbursts) at Phase E and integrate them into a judgment, Phase F, about the likelihood of the occurrence of a microburst. The hierarchical nature of the model implies that error at any phase can be passed on to later phases. Therefore, the upper limit of the accuracy of the final judgment (Phase F) depends to a large extent upon cognitive activities at earlier phases (D and E). (A statistical elaboration of this point in the framework of the lens model is presented in Stewart, 1989; Hammond, et al., 1975; Hammond and Summers, 1972. The concept of limits placed on accuracy by measurement devices is also recognized in meteorology; see, for example, Tribbia and Anthes, 1987.)

Insert Figure 1 about here

Study 1 (Lusk et al., 1988) investigated the link between Phases E and F. Forecasters were presented secondary cue values and asked to make judgments regarding the probability of a microburst. Thus, it provided a "best case scenario" in that it eliminated any error that might occur in the perception of the raw data (primary cues) or secondary cue values (microburst precursors) at Phases D or E. According to the research meteorologists/forecasters who were the subjects in this study, the secondary cues include (a) "descending reflectivity core", (b) "collapsing storm", (c) "organized convergence above cloudbase", (d) "organized convergence/divergence near cloud base", (e) "reflectivity notch", and (f) "rotation." In Study 1, seven forecasters judged the probability of the occurrence of a microburst from a sample of profiles (see Figure 2) representing hypothetical storms.

Insert Figure 2 about here

Analyses of those judgments indicated that the forecasting process could be adequately represented by a linear model. The forecasters, however, believed that they were in fact employing a nonlinear model. The model offered by the forecasters was tested but did not predict the judgments made by forecasters as accurately as the simple linear model, thus reproducing the results often observed in similar studies (see e.g., Dawes, 1982; Einhorn, Kleinmuntz and Kleinmuntz, 1979; see also Pitz and Sachs, 1984; Dawes, Faust, and Meehl, 1989, for a review.) Most important, examination of the weights placed by forecasters on the cues when making judgments revealed that forecasters placed the greatest weight on the cue "descending core," a result with which the forecasters concurred. Only modest agreement was found among the forecasters regarding their microburst probability judgments (mean correlation for the seven forecasters in Study 1 was .74).

Generalization from Static to Dynamic Tasks

Psychological vs. Expert Hypothesis

Because the above results were obtained from a best case scenario in which all forecasters were presented with the same precursor values, no error could enter into the microburst judgment at Phases D and E of the hierarchical model (see Figure 1). Thus, previous work based on lens model theory (Brehmer and Joyce, 1988; Hammond et al., 1975) would lead to the prediction that in the forecaster's typical work setting errors of observation will occur at Phase D and errors of inference will occur at Phase E. In addition, lens model theory predicts that inter-subject disagreement would be lower under these conditions at Phase F. Because this hypothesis is a generalization from previous psychological research we shall call this the Psychological Hypothesis. Alternatively, since experts in a substantive field rarely, if ever, empirically investigate inter- and intra-observer agreement in judgment, and because of the richer information conditions in actual work environments (in this case the actual radar information the forecasters much prefer), experts can be expected to argue that access to increased, more representative information at Phases D and E would result in increased agreement--contrasted with the best case scenario—at Phase F among the forecasters. We call this view the Expert Hypothesis. An additional aim of the present study, therefore, was to discover whether the Psychological or Expert Hypothesis was more nearly correct—under dynamic task conditions representative of the forecasters' normal work environment.

A further generalization from lens model theory is that those precursors (secondary cues) that are more conceptual in nature would produce more disagreement than those precursors that are more readily reducible to point coincidences on the radar screen. Thus, for example, because the precursor "descending core" is difficult to reduce to specific lines or contours on the radar screen, it should produce more disagreement than such precursors as "convergence" of wind flows. Put otherwise, the more subjective secondary cues should evoke more disagreement than the more objectively determined, readily observable secondary cues. The hypotheses below represent a more specific form of these general hypotheses, and in addition, are framed as Psychological Hypotheses which, if disconfirmed, would lend support to our formulation of the Expert Hypothesis.

Specific Hypotheses

Hypothesis 1: Agreement among forecasters regarding probability (forecast) judgments (Phase F) will be lower in the full representation of the task than in the best case scenario study (Lusk et al., 1988, Study 1), because of error introduced at the inferential phases in the judgment hierarchy (Phase E).

Hypothesis 2: Agreement among forecasters regarding secondary cue (precursor) values (Phase E) will be modest, not because the forecasters perceive the radar data differently, but because the forecasters draw different inferences about the secondary cues from the radar data.

Hypothesis 3: The magnitude of agreement among forecasters regarding precursor judgments at Phase E will vary by precursor because of differences in the degree of subjectivity in cue determination.

Effects of Updating Information in Dynamic Task Conditions

The dynamic nature of the task conditions in which forecasters normally operate makes it possible to investigate the effect of the updating and accumulation of information over time on (1) agreement, (2) changes in the forecasts, and (3) confidence.

Expert Hypotheses

Experts would argue that increased information provided by a series of time-dependent observations allows for greater understanding of the events. Therefore, as the amount of information increases over time, their judgments will begin to converge, for example, agreement will increase regarding microburst probabilities. Experts would also argue that the accumulation of information over time (and increased understanding) will lead them to become more extreme in their judgments regarding the probability of a forecast (move closer to either 0 or 1.0). In addition, forecasters should become more confident as a consequence of increased information over time. In short, dynamic task conditions should lead to (a) increased agreement, (b) more extreme judgments and (c) greater confidence.

Psychological Hypotheses

From a psychological point of view, accumulation of information over time is unlikely to lead to increased agreement because, as the results of Study 1 indicated, even in the best case scenario the forecasters did not agree in their precursor judgments and, moreover, were not aware of their differences. Thus, further information would be as likely to drive them apart as bring them together. That is, new information over time is as likely to reinforce the components of the judgment process that produce disagreement as it is likely to reinforce the components that produce agreement. Also, forecasters are not likely to become more extreme in their judgments because, as Study 1 showed, the intra-forecaster judgment policies have such a large error component that new information is likely to have little systematic impact but is likely to be lost in the "noise" of the experts' judgment system. And although the forecasters may become more confident over time, their confidence will be unjustified inasmuch as the agreement in their judgments will not increase. Again, the specific hypotheses are framed as Psychological Hypotheses, and their disconfirmation would lend support to the expert point of view (with the exception of Hypothesis 6, as noted below).

Specific Hypotheses

Hypothesis 4: Agreement among forecasters' precursor and probability judgments will not increase over time because the empirical significance of changing information is not common among the forecasters.

Hypothesis 5: Individual forecaster's probability judgments will not become more extreme over time because no explicit instructions exist for how new information should change judgments.

Hypothesis 6: Forecasters will become more confident in their precursor and forecast judgments. (Note: in this case, the Psychological and Expert Hypotheses are the same but the reasons for them differ.)

Method

In keeping with lens model theory, the method employed was that of representative rather than systematic design (Brunswik, 1952, 1956; Hammond, 1966). That is, an effort was made to represent the actual conditions under which the research meteorologists made their forecasts of microbursts rather than varying one (or n) variable(s) at a time. Each subject's performance was studied separately over six potential microburst cases. Although this sample of cases is admittedly smaller than desirable, many hours of technical work were required to remove each case from the file tapes and observation of the six cases required many hours of the meteorologists' time. Thus, high operational costs limited the size of the sample of cases. The six cases did, however, provide a total of 25 data points for evaluation, as the procedure described below explains.

Procedure

The subjects in this experiment were four research meteorologists from the National Center for Atmospheric Research in Boulder, Colorado. All four meteorologists had participated in Study 1 (Lusk et al., 1988).

The experiment was conducted in two phases. The first phase includes two cases (one microburst and one null case), after which a preliminary assessment of the procedure and results was conducted. Both the psychologists and meteorologists decided that further data would be worth acquiring and the experimental procedure was slightly modified to collect those data. The procedures are detailed below. All procedures were planned and carried out in consultation with a research meteorologist, familiar with the problems and procedures of microburst forecasting, who did not serve as a subject.

During each experimental session, the forecaster was seated in front of a large computer terminal used to present color Doppler radar displays. The experimenter was seated in front of a computer terminal that was used to run the experimental session. At the first session of each phase of the experiment, the forecasters were presented with instructions regarding how the experiment would proceed. The forecasters were then presented with a "volume" of radar data, based on 13 radar scans through the height of the atmosphere. After each volume they made judgments of precursor values and the probability of a microburst. This procedure was repeated until completion of each case.

The Cases

Six cases were used: two in the first phase and four in the second phase. Half of the cases in each phase were null cases and half were microburst cases.

Each case was arranged on a tape in consecutive volumes, each of which scanned through the height of the storm cell. The volume scans were repeated every 2.5 minutes. In the first phase, Case 1 included six volumes. The data for Case 2 spanned eight volumes. However, one volume was skipped due to faulty radar data. In addition, one volume in Case 2 only included the lower seven scan levels. However, judgments were still collected for that short volume. In the second phase all cases included four volumes of data. In both phases, each case terminated before the microburst was evident on the lowest scan or before any obvious or substantial decrease in the intensity or height of the cell in the null cases (i.e., before the outcome became apparent).

The Judgments

The forecasters were asked to make judgments of the six precursor values they had previously indicated to be the cues used in forecasting microbursts: descending core, collapsing storm, convergence/divergence above cloud base, convergence/divergence at or below cloud base, notch, and rotation (Lusk, et al., 1988, Study 1). In addition, forecasters made judgments of the probability of a microburst occurring in the next 5 to 10 minutes.

The judgments regarding precursor values and probability of a microburst were made on the same scales as utilized in Study 1 (see Figure 3). In addition, to the right of each rating scale was a blank for the forecasters to insert their confidence in their precursor judgments (confidence ratings for the probability judgments were not collected in Phase 1, but were collected in Phase 2).

Insert Figure 3 about here

In the first phase, judgments were made after each volume. Therefore, judgments were made six times for Case 1 and seven times for Case 2. In the second phase, judgments were witheld for the first volume. This change was made because (1) many of the precursor judgments were difficult to make with data at only one time period, and (2) to save time so that more cases could be presented. Thus, because there were four volumes for each case in Phase 2, three judgments were made for each of the four cases in the second phase. This resulted in 25 possible data points for each subject (some subjects had 24 data points for some precursor judgments due to the short volume in Case 2).

The Experimental Session

For the first phase, the instructions explained how the experimental sessions would proceed: each case would consist of several volume scans over time of a storm cell that did or did not produce a microburst, starting with the lowest scan at the earliest time. After observing each scan, the forecasters were to tell the experimenter that they were ready for the next level scan. The forecasters were given up to thirty seconds to view each scan. After completion of a volume in this manner, the forecasters filled in the rating sheet. In addition, the instructions stated, in part:

> At the time of the first volume you can assume that a microburst is not presently occurring. Please assume before observing the first scan, that on the basis of prior information (morning soundings, tec.) you have already reached the conclusion that the likelihood of a microburst on this day is .50. Then adjust your probabilities of a microburst after observing the radar data. Each case will terminate prior to evidence of outflow from a microburst or evidence that the storm is obviously dissipating.

Finally, the forecasters were given instructions to "think aloud" and their verbalizations were tape recorded.

The instructions for the second phase explained the changes in the experimental procedure. The forecasters were informed that they would receive sounding data, view only four volumes of data, and make judgments only after the second through fourth volumes. In addition, the instructions explained that the scans would be presented continuously and that they would not need to think aloud.

The forecasters were provided with blank paper for taking notes and felt-tip pens to mark the CRT screen. At the beginning of each case, the forecasters were given the coordinates for the storm cell they were to attend to.

Prior to presentation of each case in the second phase, the forecasters were given the eleven o'clock sounding of the atmosphere for the day from which that case was drawn. The subjects were then asked what the probability of a microburst occurring was, based on the sounding information alone.

In the first phase, half of the forecasters were presented with Case 1 first, and half were presented with Case 2 first. In the second phase, the cases were arranged on a tape in a fixed order. Each forecaster began with a different case, but otherwise the order of presentation was fixed.

Results

Each hypothesis is stated in terms of the psychological point of view. Thus, falsification of the hypothesis lends support to the alternative, the Expert Hypothesis.

Hypotheses Regarding Generalization from Static to Dynamic Tasks

Hypothesis 1: Hypothesis 1 states that agreement among forecasters regarding probability judgments (Phase F) will be lower than in the best case scenario study because of error introduced at Phase D in the judgment hierarchy. The correlations between judgments of each pair of forecasters participating in both the best case scenario study and the present study are presented in Table 1 (for the best case scenario study) and Table 2 (for the present study). The mean correlation from the best case scenario study for the four forecasters participating in both studies is .67 (from Table 1) and the mean correlation from the present study is .49 (from Table 2). (Note that mean correlations presented in this paper were computed by converting raw correlation coefficients to Fisher's Z, computing a mean, then converting the mean back to an r value.) Thus, comparison of the correlations in Table 1 to those in Table 2 indicates that the latter are substantially lower than the former, providing support for Hypothesis 1.

Insert Tables 1 and 2 about here

Hypothesis 2: Hypothesis 2 states that agreement among forecasters regarding secondary cue values will be modest, not because the forecasters perceive the radar data differently, but because the forecasters integrate those data differently. The data used to test this hypothesis were the secondary cue judgments made after each volume. The correlations between the judgments of each pair of forecasters were computed for each precursor and are presented in Table 3.

Insert Table 3 about here

The data in Table 3 clearly indicate a lack of agreement between forecasters regarding the precursor judgments. Although many of the correlations are substantially larger than zero (and are, in fact, statistically significant), they are all substantially lower than acceptable limits for practical use, providing support for Hypothesis 2.

In order to support our assertion that disagreement at Phase E did not occur because forecasters perceive the data differently at Phase D, forecasters' written notes were analyzed. Three of the four forecasters took notes regarding the maximum reflectivity values present at particular scan levels. To compute agreement, each pair of forecasters was considered. An agreement was counted each time both forecasters of the pair recorded a reflectivity value within 5 dBZ of each other. The percentage of agreement was computed by dividing the number of agreements by the number of times both forecasters in a pair recorded some maximum reflectivity value. For Forecasters A and B, the agreement regarding maximum reflectivity values was 96%, 93% for Forecasters A and D, 99% for Forecasters B and D. In short, agreement among forecasters with regard to the purely objective data was very high, lending further support to Hypothesis 2.

Hypothesis 3: Hypothesis 3 states that the magnitude of agreement among forecasters regarding the various precursor judgments will vary by precursor because of differences in the subjectivity involved in the judgment. Examination of Table 3 indicates a higher degree of agreement on some precursors than on others. Each correlation in Table 3 was converted to a Fisher's Z, a mean was computed for each precursor matrix and the mean Z was then converted back into a correlation coefficient. These mean agreement correlations for each precursor are presented in Table 4. As hypothesized, agreement regarding precursor values did vary considerably, with highest agreement for the two convergence precursors, second highest for collapsing storm and notch, lower for rotation and the lowest agreement was for judgments of descending core. It is noteworthy that descending core has the lowest mean agreement and that some of the correlations in Table 3 are actually negative! This result is important because our previous work indicated that forecasters weighted this precursor most heavily in arriving at microburst probability judgments (Lusk et al., 1988, Study 1).

Insert Table 4 about here

Hypotheses Regarding Effects of Updating of Information on Judgments

Hypothesis 4: Hypothesis 4 states that agreement among forecasters' precursor and probability judgments will not increase over time because the empirical significance of new information is not common among the forecasters. To investigate this hypothesis, agreement among the forecasters was computed separately for judgments from the last three volumes of each case. The initial data for these analyses were judgments for each of the last three judgment times (volumes) in each case. Specifically, "Time 1" included judgments from volume 4 for Case 1, volume 6 for Case 2 and volume 2 for the other four cases. Likewise "Time 2" included judgments from volume 5 for Case 1, volume 7 for Case 2 and volume 3 for the other cases. Finally, "Time 3" included data from the last volume of each case. This resulted in a total of 18 data points for the analyses. The means for each time period were computed as in the above analyses.

The means of these values for each precursor and microburst forecast are presented in Table 5. The mean values reported for each time in Table 5 reflect the overall lack of agreement, prevalent even after the final volume (Time 3). However, with the exception of descending core and notch, this measure of agreement does indeed increase over time. Paralleling similar results in overall agreement (Hypothesis 3), the convergence judgments show the most marked increase in agreement over time, while the descending core judgments show the least. Also, the probability of microburst judgments shows an increase over time. These analyses do not provide clear support for Hypothesis 4 because there are some instances of movement toward similar judgments with the accumulation of more evidence. Nevertheless, the agreement at Time 3, while improved from previous times (for some cues), is still lower than is useful in an operational setting (ranging from .18 to .85). In sum, Hypothesis 4 is disconfirmed, but the evidence is weak.

Insert Table 5 about here

Hypothesis 5: Hypothesis 5 states that forecasters' probability judgments will not become more extreme over time because no explicit instructions exist for how new information would change judgments. To test this hypothesis, a separate analysis of variance was performed for each forecaster utilizing the last three judgment times as in the above analyses. It was expected that more information should move forecasters toward a judgment of either 0 (for null cases) or 1 (for microburst cases). Therefore, before conducting the analyses, the probability judgments were converted to absolute values of mid-point deviations. That is, the scale mid-point (.50) was subtracted from each probability judgment and the absolute values were utilized as the data for the analyses. The means of these values for each time period are presented in Table 6. Two different analyses of variance were performed for each forecaster. The first was an omnibus F test and the second was a linear contrast. None of the F ratios for either type of analysis was statistically significant (at the .05 level), although for Forecasters B and C the means are clearly increasing (p-levels for the linear contrast were .12 and .10, respectively).

Insert Table 6 about here

Collapsing across subjects, a significant effect was found for time for the omnibus test (F(2,6) = 8.55, p = .02) and the linear contrast (F(1,6) = 13.64, p = .01; in these analyses time was treated as a repeated measure). These results do not clearly disconfirm Hypothesis 5; two subjects were sensitive to the accumulation of evidence while two were not.

Hypothesis 6: Hypothesis 6 states that forecasters will become more confident in their precursor and probability judgments over time because of increased information, but unjustifiably so, inasmuch as their agreement will not increase. The data for these analyses were confidence ratings for the last three judgment times of each case as in the above analyses. The means of the confidence ratings over time for the cues are presented in Table 7. The top of the table displays the means for each forecaster across time and the bottom includes the means for each cue separately across forecasters. Because Hypothesis 6 asserts an increase in confidence over time, both omnibus tests and linear contrasts were conducted. In addition, the effects of cue type and the interaction of cue with time were examined. Each two-way analysis of variance was conducted for each forecaster separately as well as for all forecasters combined.

Insert Table 7 about here

The omnibus effect of time was significant (F(2,90); p < .05) for each of the forecasters except for Forecaster B. The linear effect of time (a more precise test of the hypothesis) was significant for Forecasters C and D (F(1,90); p < .05) and marginally significant for Forecaster A (p = .11). In the analysis across subjects, the omnibus test of time was marginally significant (F(2,6) = 3.19, p = .11), while the linear contrast was statistically significant (F(1,6) = 6.09, p < .05).

The effect of cue type was significant for Forecaster B and Forecaster D (F(5,90); p < .05) and marginal for Forecaster A (F(5,90); p = .06). The means are presented in Table 8. As Table 8 indicates, there are substantial individual differences in expressions of level of confidence. The effect for cue in the analysis that included all subjects was not significant (F(5,15) = .54, p = .75). The interaction between cue and time was not significant in the individual subjects' analyses, but it was significant when collapsing across subjects (F(10,30) = 2.23, p < .05). As can be seen from the bottom of Table 7, the linear effect of time was not as great for the two convergence cues as for the other cues.

Insert Table 8 about here

Table 9 presents the mean confidence ratings for the probability judgments. No significant effects were found on either the omnibus or linear contrast tests for any of the forecasters separately. However, since confidence judgments for the probability judgments were collected in Phase 2 only, the power of these analyses is relatively low. In fact, with the exception of Forecaster D, the means are in the predicted direction. Moreover, when the data are pooled across subjects the omnibus test indicates a marginal effect of time (F(2,6) = 4.64, p = .06), and the linear contrast indicates a significant effect (F(1,6) = 8.65, p = .03).

Insert Table 9 about here

These analyses indicate support for Hypothesis 6. Across cues, two of the four forecasters' confidence ratings do increase over time. One of the remaining forecasters' ratings are at the ceiling (Forecaster B) and therefore cannot show the effect. The interaction between time and cue type indicates that the effect of time on confidence was greatest for descending core and collapsing storm and least for the two convergence cues. Finally, three of the four forecasters' confidence in their probability judgments did increase over time (though not statistically significant), providing further support for Hypothesis 6.

Summary of Results

Hypotheses Regarding Generalization

The first three Psychological Hypotheses were not disconfirmed. That is: (1) agreement in forecasts was lower under representative task conditions than in the best case scenario because of error introduced at the inferential (secondary cue) level; (2) agreement among the forecasters' judgments of secondary cue values in the hierarchical model was modest, not because of differences in the perception of the radar data but because of differences in inferences drawn from them; and (3) the magnitude of agreement at the secondary cue level varied by secondary cue; the greater the subjectivity the greater the disagreement. Because these hypotheses are derived from lens model theory and research in static tasks, the results constitute a significant generalization to a complex dynamic task representative of that normally encountered by mature professionals.

Hypotheses Regarding Updating of Information

The results regarding the first two of the three Psychological Hypotheses concerning the effects of updated information on judgments in the dynamic task do not yield clear disconfirmation, although the third was more clearly disconfirmed. That is, in regard to Hypothesis 4, the degree of agreement among the forecasters did increase as information was accumulated over time (although the increase was slight for four of the six precursors and for the forecast judgments). In regard to Hypothesis 5, there was some evidence to suggest that the forecasters' probability judgments did become more extreme. Again, however, the evidence was weak; two of four forecasters' judgments did not become more extreme. With regard to Hypothesis 6, there is some evidence that the forecasters do become more confident as information is accumulated over time. In short, the evidence related to the hypotheses concerning judgments of the effects of updated information was far from convincing. The ambiguity of these results, together with the small sample of cases, certainly argues for more research on the impact on judgments of new information received over time.

Verbal Protocols

Examples of the verbalizations are provided in Appendix A. The protocols indicate that during observation of the radar data the forecasters were primarily focusing on both the proximal and inferential levels (Phases D and E) in the hierarchical judgment model (see Figure 1). That is, the verbalizations primarily concern noticing the radar data such as the maximum reflectivity values, convergence or divergence, and making note of the occurrence of features such as a notch at each level scan.

Although these verbalizations are skimpy, if taken at face value they further support the conclusion that the lens model theory does generalize to behavior in dynamic tasks. That is, the verbalizations indicate that the forecasters provide a dichotomous yes or no value regarding inferences about the occurrence of each precursor, then decide exactly what value to circle on the scale. Thus, the cognitive process for making the probability of a microburst judgment was not verbally expressed which suggests that it takes place on an intuitive level. No calculations or applications of a principle or formula were ever observed; in short, no analytical work for organizing the information was evident in the protocols. (See Hammond, Hamm, Grassia, and Pearson, 1987,; Hammond, 1988, for a discussion of intuitive and analytical cognition within the framework of lens model theory.)

Discussion

Theory

The most important theoretical finding is that lens model theory can be generalized, at least in part, to expert judgments made in complex dynamic task conditions. Under representative conditions a higher level of disagreement regarding probability judgments was found than in the best case scenario. Agreement at the primary cue level was found to be very high while agreement at the secondary cue level was only moderate. Moreover, agreement was much higher for some secondary cues than others, indicating differences in proximity of secondary cues to the primary cue level. The notion of differential proximity may lead to a more sophisticated version of the hierarchical model, with secondary cues at Phase E distanced from the primary cues at Phase D according to their respective inter- or intra-observer reliabilities. Future research should investigate the mechanisms underlying the proximity of secondary to primary cues and the extent to which these and other generalizations hold when the limitations mentioned above are reduced.

Methodology

The hierarchical model also helped in guiding the methodology. Data were collected to investigate potential error at each phase in the hierarchy involving judgment processes. The most important aspect of our methodology was the use of representative design in contrast to conventional systematic design of the study, which offered several advantages. By virtue of studying each meteorologist separately over a sample of events (instead of the conventional technique of testing the effect of restricted stimuli presented to a large population of subjects), it was possible to learn that certain results obtained from lens model theory and research in static task conditions generalize to dynamic task conditions involving four mature professionals in circumstances highly representative of their work situation. These generalizations are reflected principally in results concerning agreement among the experts, and thus in the results that favor acceptance of psychological hypotheses over expert hypotheses. These issues could not have been addressed by conventional research design.

Results

The results provided a sharper focus on the reasons for disagreement among experts. The hierarchical model delineates each phase at which human cognitive processes operate and therefore the points at which error can be introduced. The research assessed agreement at each point in the hierarchy and through this procedure it was determined that very little disagreement occurred at the level of perception of the raw data. However, a great deal of disagreement occurred regarding judgments of both precursors and microburst forecasts. Identification of the particular precursors demonstrating high levels of disagreement makes it possible to focus on the variables with the greatest potential for improving judgment. Thus a hierarchical model that separates inferences at an intermediate level from raw data has methodological, theoretical, and practical significance deserving of further work.

Finally, the results regarding the effect of updated information on judgments may have important practical implications, as well as theoretical implications for researchers involved in dynamic tasks. Experts often believe that more data provide a better understanding of a phenomenon, which leads to better predictions. The results of this study suggest this may not be the case, at least in situations where more information is updated information.

Future Directions

Although we believe the present study of dynamic decision making has demonstrated its usefulness, there is much that it does not do. For example, it is based entirely on the study of functional relations between cues and judgments over time; it does not address the topic of pattern recognition. Nor does it investigate the differential role of intuitive and analytical cognition, even though microburst forecasts are based on both analytical, scientific understanding derived from scientific research and intuitive judgments derived from experience. (See Hammond, 1988, for a discussion of the differential role of functional relations and pattern recognition, as well as intuition and analysis in dynamic tasks.) In addition, the difference between the logic of understanding and the logic of prediction (de Montmollin and De Keyser, 1986; see also Brehmer, 1987) is not developed. This topic, hardly touched, is bound to be of importance in circumstances such as weather forecasting in which the decision makers have a scientific basis for understanding that must be combined with an experiential basis for prediction. Nor have we referred to the nature of the "ecological interface", (Rasmussen and Vicente, 1987; see also Schwartz and Howell, 1985) that is, the design of the display of information and its potential for enhancing or reducing the efficacy of dynamic decision making. Also omitted is a discussion of the role of feedback, either cognitive or of the simple outcome form, and its effect on change in judgments (Balzer, Doherty, and O'Connor, in press). And, of course, the small sample of cases rules out a discussion of the accuracy or skill of the forecaster, and the small number of forecasters studied imposes limits on generalization to other forecasters. We mention these issues not only to caution the reader, but to emphasize the complexity that will have to be addressed in future efforts if we are to advance our understanding of judgment and decision making in dynamic tasks.

A further issue concerns the distinction between relative and absolute expertise. In the present case, the meteorologists were indeed experts in the relative sense; they probably have more knowledge and experience than any other meteorologists regarding the microburst phenomena and the vast technology associated with their detection. On the other hand, these meteorologists do not claim to be experts (in fact, resent the use of the term) in the absolute sense; they insist that they do not have full or even satisfactory knowledge about microburst events; rather they describe themselves as in the process of studying them to make predictions of these hazards a practical success. Their situation is analogous to the research physician who understands a specific disease better than anyone else, but who does not pretend to have sufficient knowledge to make highly accurate diagnoses or prognoses.

This distinction is important not only because of the need to be clear about the particular form of expertise under study, but also because these different forms of expertise place different demands on judgment researchers. In the case of absolute expertise, the primary goal of judgment researchers will be to discover how such expertise is applied to various types of problems (e.g., high vs. low uncertainty conditions). In the case of relative expertise, on the other hand, the primary goal of judgment researchers will be to aid the experts to discover what features of the situation are frustrating or enhancing their efforts to improve the status of their expertise. And once discovered, the researcher will want to convey that information to the experts. In the present study of relative expertise, the primary goal was achieved, but it is doubtful that the communication process was successful. The meteorologists showed interest in the hardly disputable facts of disagreement in their use of information, claiming that in the present state of their knowledge such disagreement was not surprising. Nevertheless, in subsequent research they continued (to our surprise) to make subjective judgments about precursors without examining the extent of, or reasons for, their disagreement on precursor judgments in both empirical research and operational tests of forecasting accuracy.

Disagreement of the magnitude observed under representative task conditions raises serious questions about the use of experts—usually a single expert!—as the basis for an expert system, as is customary in work in the field of artificial intelligence (see Adelman, in press, for a detailed discussion). Indeed, one of the forecasters who served as a subject in this study had been used as an expert in the development of an Al expert system for forecasting the occurrence of microbursts. Obviously, had a different forecaster been used, a different system—producing different forecasts—would have been developed.

Appendix

Example Protocols for Study VI

Subject 1: Case 1

S: Okay where are we? The number of the next volume, 4. Ah what's happening? Uh very weak divergent flow at the surface, very weak, only three meters per second. And we've got about 55. It's 55. Very weak. Huh again we're we see at these 55, we get divergence again above. See it really looks like we're getting a little, it's diverging out above cold air, but it's weak. And it gone, oh wow. We get some actually 60 this time, reflectivity. A lot more reflectivity. And actually we're showing a little convergence now. Oh wow it's up to 60 now. But velocity feature not very strong, slight. Still 60, no good velocity feature. I'm not wild about the angle we're getting now. If there were convergence in that core we wouldn't see it well. Now at 55, I'll call it now, it's just only a touch of 60. Slight indication of that notch is at this level, now. This is 15 6 [pause] there's xxxx convergence into that too, hmm. Nice notch now, reflectivity 55. Can't see an obvious velocity feature with it though. Here's where we get the convergence. 45, 45 convergence. Okay we've lost a lot of reflectivity now. And we, now we're actually divergence. It's slipping down into the about 45, maybe 40, at 30 degrees. Oh it's gone only 25 left so we have a real collapsing case here. Boy that was faster wasn't it.

E: yeah

S: I had to move. Just trying to see xxx [silence] The top's coming down. Okay now uh descending reflectivity core, yeah it's still, it's not one of the obvious, the most obvious cases in the world, but it's still descending. I'll put a 7, confidence is only about 50 percent. Collapsing storm, it is collapsing but it's not the most obvious one you ever saw. So I'll put 7, confidence at 60 percent. Organized convergence above cloud base, yes it's still there. It's still, and it's actually descended slightly with time I see. Not much, it's still, it's still primarily in the three to four kilometer zone which is a good zone for it. It's not that strong and organized. I put confidence only at 60, meaning I don't think it's all that significant. Organized, there's still a divergence below cloud base, and I really think that may be significant. Um I'm going to put, I'm circling the one and two, saying, I'm putting 70 on it cause I think the outflow is really divergent above the cold air. It may not make it to the surface very strong. Good reflectivity notch now between 2 and 3 kilometers. I'll put a 9 on it, confidence is, well it's there, 90. Rotation was um not as good, it was weak. Last time I think I had weak. I xxx put down a 6. Um confidence is only 50 percent. Okay now if we're going to have a microburst that's going to occur in this period, I'm not very, I think it's only going to be a very weak outflow though cause the reasons I've given. Last time I gave 25 percent. I'll go with 30 and hope I'm right.

Subject 2: Case 2

S: [Silence] Okay max reflectivity here is 55. Still got weak convergence delta V is 3, okay. [Pause] 55 again, two point two. Weak convergence again. Okay. xxx don't see it this time. 4 and a half

Page 33

degrees, 55. Um still convergent weakly delta V is 3, okay. [Pause] 6.7 xxx 55. [Silence] Um not much going on that's really different, okay. 8.8 is 55. 55 [pause] hmm. A suggestion of xxx divergence on the north edge of cell, delta V is about 3. It's still pretty weak, okay. [Silence] 50 DBZ, 11 degrees. Got that wind change xxx, okay. [Pause] 50 DBZ again. [Silence] Okay. [Silence] Well that's interesting, huh. 50 DBZ, ??erosion?? echo in the back. Notch is still there. It's kind of filling in though, there's mid-line with more echo to the west of the cell than there has been previously. [Silence] Cyclonic, anti-cyclonic couplet there. Um okay [silence] 50 DBZ, this storm really is tilted in height. Sort of see convergence xxxx weak xxxxx [silence] okay. [Silence] okay xxx DBZ [silence] There's some shear areas but nothing really significant. This is 22 degrees, um [pause] okay. [Silence] 45, again we've gotten a couple of shear areas. Cyclonic, anti-cyclonic shear not real couples to speak of [pause] okay. [Silence] xxx xxxx [pause] cyclonic, anti-cyclonic shear okay. [Pause] The cell's falling apart xxx. 35 DBZ. There's convergence ??in the anvil??. [mumbles] 6. [silence]

S: Uh reflectivities are still maintaining themselves pretty well. [Silence] Slightly increasing aloft and then decreasing at the very highest angle. So we don't have a descending core. And it's not collapsing. There's no real convergence above cloud base, except in the xxxx. [Silence] Um [silence] there's not, there's convergence at or below cloud base. xxxx xxxx kilometer, there's that one little spot of divergence ??at one kilometer?? It's really weak though. [Silence] The notch has become weaker. Not as well defined. And there's also xxx flow xxx so I'm going to rotation, no there's some cyclonic shear and that's it. Probability of a microburst within the next 5 to 10 minutes, I'm still going to stick with the 50 percent.

Subject 4: Case 1

S: xxx you look at that point 5 degree velocity and there's nothing there. There is not a microburst outflow. There's some garbage right there, but that's not real. And uh looking all the way up at 2.2 we don't really see any divergence or velocity structure. And we've got the high reflectivity xxx so unless we see some dramatic increases in velocity structure, which we don't really see here at 4.5, it's going to be awful hard to say yes we're going to get something. And uh even at 6.7 we're not seeing any good strong velocity features associated with that core. [Long Silence] slight hint that there may be convergence coming in here that we can't see associated with that notch. And xxx interesting to look at it from a from a radar out here where we could get a better view. Still seeing that notch, but again, as I say, it's not that good of a velocity structure. I did see some sign of convergence xxx. [Silence] Saw some. xxx [silence] xxx rotation xxx some convergence not really that good. [Silence] xxx looks about the same as it was before [silence] okay.

S: xxx other sheet xxx put down thing xxx for can't remember for sure. ??We do have?? some descent of core. The storm has collapsed already. I think there's a slight xxx still kind of collapsing. Uh xxx not really much happening above cloud base xxx. xxx Part of why you think collapsing storm. Slight indication of xxx. We got an indication xxx notch. Well nothing happened last time. Still not seeing it, we've got the high reflectivity down so, not willing to say no chance anymore, but uh got to start backing off a little bit on that probability. I'll be a little less convinced that something's going to happen.

Subject 5: Case 2

S: Okay. It takes it forever. Oh we're going to start with point 5. [Silence] Oh yeah, this guy's racing off to the north, and 55 DBZ core. [pause] And a little convergent shear line still with us way off to the south. Oh that's what happened to the cell. It moved off of its convergent line. Now it's lost its low level support. It's going to crash, okay. [Silence] Oh that's why the core crashed down in such a hurry. [Silence] That's right I did see a sizeable increase in reflectivity. And that's what happened to it. Okay.

E: Is that an okay for me?

S: Yeah, that's an okay okay. [Silence] Oh gosh 60 DBZ. [Silence] No velocity features at all associated with the cell at 4.5 degrees. [Silence] Surprised it hasn't put out an outflow, okay. Wonder why not? [Silence] Oh gee, everything's back down to 55 DBZ now. [Silence] huh. Still no real velocity features. It's really just a flat field. Okay. Notch on the side. [Silence] huh let's see, not much at all going on. Strange, we're up at 8.8 degrees and I don't see much of anything, huh. Okay, go to the next one, if you haven't already. [Silence] 50 to 55, well a little bit of cyclonic shear. Certainly a notch. Okay. [Silence] Oh another cyclonic shear right in the middle of the cell. [Silence] Oh yeah, a little bit of convergence right there, okay. [Silence] Oh hurry up [silence] yep, a little bit of convergence now in the middle of the cell. [Silence] Okay. [Silence] Oh rotation hanging off, way off on the end out

Page 36

in the area of no, not much signal. Uh now we're seeing convergence peppered about here, hither in the thither. Rotation down in the south end where we've always seen it. xxx okay. [Silence] Oh there's a clear rotation near that notch, cyclonic rotation. Okay. [Silence] huh a little bit of divergence right up here. 25.8 degrees, cyclonic shear to the south, probably strong rotation. Huh. [Silence] Is it doing anything? [referring to computer] [Silence] Oh yeah now I see convergence on the western end, right where that notch, okay. [Silence] Oh there's convergence all over the place, 34.8. Uh max reflectivities xxx 40 to 45. Okay.

E: That's it on that one.

S: Okay. Descending reflectivity core, it's obvious. Collapsing storm, probably is, but not real sure yet. Organized convergence or divergence above cloud base, you betcha. Not much convergence at or below cloud base, I didn't seen anything. And I'm pretty sure I didn't see anything. There's a reflectivity notch. There's rotation. I'm a little concerned that I didn't see any divergence at the surface, but what the heck. 90 percent, or is this [silence]

Author Note

The authors would like to thank the following people: Rod Potts for his help in case preparation and data collection; the meteorologists of the National Center for Atmospheric Research for their participation as subjects, Deborah Smith for her help in data analysis and manuscript preparation, Patricia Armstrong for her help in manuscript preparation, and Reid Hastie for comments on an earlier draft.

This research was supported by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142 and by the National Center for Atmospheric Research. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

Reference List

Adelman, L. 'Measurement Issues in Knowledge Engineering', in A. Sage (Ed.), Concise Encyclopedia of Information Processing in Systems and Organizations, Oxford: Pergamon Press, in press.

Balzer, W. K., Doherty, M. E. and O'Connor, R., Jr. 'The Effects of Cognitive Feedback on Performance', Psychological Bulletin, in press.

Brehmer, B. 'System Design and the Psychology of Complex Systems', in J. Rasmussen and P. Zunde (Eds.), Empirical Foundations of Information and Software Science III (pp. 21-33), New York: Plenum, 1987.

Brehmer, B. and Joyce, C. R. B. Human Judgment: The SJT View, Amsterdam: North-Holland, 1988.

Brunswik, E. The Conceptual Framework of Psychology, Chicago: University of Chicago Press, 1952.

Brunswik, E. Perception and the Representative Design of Psychological Experiments (Second Edition), Berkeley: University of California Press, 1956.

Dawes, R. M. 'The Value of Being Explicit When Making Clinical Decisions', in T. A. Wills (Ed.), Basic Processes in Helping Relationships (pp. 37-58), New York: Academic Press, 1982.

Dawes, R. M., Faust, D. and Meehl, P. E. 'Clinical Versus Actuarial Judgment', Science, 243 (1989), 1668-1673.

de Montmollin, M. and De Keyser, V. 'Expert Logic V. Operator Logic', in G. Johannsen, G. Mancini and L. Martensson (Eds.), Analysis, Design and Evaluation of Man-machine Systems (pp. 43-49), Oxford, NY: International Federation for Automatic Control, 1986.

Einhorn, H. J., Kleinmuntz, B. and Kleinmuntz, D. N. 'Linear Regression and Process-tracing Models of Judgment', Psychological Review, 86 (1979), 465-485.

Hammond, K. R. (Ed.). The Psychology of Egon Brunswik, New York: Holt, Rinehart, and Winston, 1966.

Hammond, K. R. 'Judgment and Decision Making in Dynamic Tasks', Information and Decision Technologies, 14 (1988), 3-14.

Hammond, K. R., Hamm, R. M., Grassia, J. and Pearson, T. 'Direct Comparison of the Efficacy of Intuitive and Analytical Cognition in Expert Judgments', IEEE Transactions on Systems, Man, and Cybernetics, SMC-17 (1987), 753-770.

Hammond, K. R., Stewart, T. R., Brehmer, B. and Steinmann, D. 'Social Judgment Theory', in M. F. Kaplan and S. Schwartz (Eds.), Human Judgment and Decision Processes (pp. 271-312), New York: Academic Press, 1975.

Hammond, K. R. and Summers, D. A. 'Cognitive Control', Psychological Review, 79 (1972), 58-67.

Lusk, C. M., Stewart, T. R. and Hammond, K. R. 'Toward the Study of Judgment and Decision Making in Dynamic Tasks: The Case of Forecasting the Microburst' (Technical Report 273), Boulder: University of Colorado, Center for Research on Judgment and Policy, 1988.

Pitz, G. F. and Sachs, N. J. 'Judgment and Decision: Theory and Application', Annual Review of Psychology, 35 (1984), 139-163.

Rasmussen, J. and Vicente, K. J. 'Cognitive Control of Human Activities and Errors: Implications for Ecological Interface Design' (Technical Report Riso-M-2660), Roskilde, Denmark: Riso Library, Riso National Laboratory, 1987.

Roberts, R. D. and Wilson, J. W. 'A Proposed Microburst Nowcasting Procedure using Single Doppler Radar', American Meteorological Society, 28 (1989), 285-303.

Schwartz, D. R. and Howell, W. C. 'Optional Stopping Performance under Graphic and Numeric CRT Formatting', Human Factors, 27 (1985), 433-444.

Stewart, T. R. 'Seven Components of Forecasting Skill', Albany, NY: SUNY at Albany, Center for Policy Research, unpublished manuscript, 1989.

Tribbia, J. J. and Anthes, R. A. 'Scientific Basis of Modern Weather Prediction', Science, 237 (1987), 493-499.

Table 1

Agreement Correlation Coefficients from Best Case Scenario Study (Probability of a Microburst)

	A	B	C
B	.78		
C	.75	.72	
D	.52	.49	.64

$\bar{r}$ (across all forecasters) = .67

Table 2

Agreement Correlation Coefficients (Probability of a Microburst)

	A	B	C
B	.60		
C	.88	.45	
D	.31	.15	.19

$\bar{r}$ (across all forecasts) = .49

Table 3

Agreement Correlation Coefficients for Precursor Judgments

Descending Core

	A	B	C
B	.14		
C	-.06	.12	
D	.10	.35	-.14

Collapsing Storm

	A	B	C
B	.69		
C	.47	.53	
D	.57	.40	.17

Convergence Above Cloud Base

	A	B	C
B	.65		
C	.71	.49	
D	.58	.53	.45

Convergence at/or Below Cloud Base

	A	B	C
B	.54		
C	.43	.76	
D	.77	.59	.45

Notch

	A	B	C
B	.38		
C	.51	.25	
D	.61	.57	.34

Rotation

	A	B	C
B	.06		
C	.12	.39	
D	.51	-.01	.26

Table 4

Mean Precursor Agreement

Precursor	r
Descending Core	.09
Collapsing Storm	.47
Convergence Above	.57
Convergence Below	.59
Notch	.44
Rotation	.22

Table 5

Mean Correlation Coefficients Over Time

	Time 1	Time 2	Time 3
Descending Core	.095	.185	.175
Collapsing Storm	.290	.360	.810
Convergence Above	.435	.680	.745
Convergence Below	.515	.745	.855
Notch	.690	.580	.830
Rotation	-.055	.335	.495
Probability	.405	.540	.600

Table 6

Mean Probability Judgments* Over Time

Forecaster	Time 1	Time 2	Time 3
A	.22	.22	.23
B	.12	.13	.22
C	.16	.18	.26
D	.31	.28	.36
ALL	.20	.20	.26

*These data were converted to absolute values of deviations from .5.

Table 7

Mean Confidence in Cue Judgments

Means for each Forecaster Across Cues

Forecaster	Time 1	Time 2	Time 3
A	.49	.63	.57
B	.95	.93	.95
C	.46	.51	.58
D	.89	.93	.96

Means for each Cue Across Forecasters

	Time1	Time 2	Time 3
Descending Core	.63	.74	.75
Collapsing Storm	.66	.73	.78
Convergence Above	.72	.75	.75
Convergence Below	.73	.79	.77
Notch	.73	.75	.79
Rotation	.71	.72	.76

Table 8

Mean of Precursor Ratings by Forecaster

Precursor	A	B	C	D
Descending Core	.46	.99	.53	.86
Collapsing Storm	.58	.95	.48	.89
Convergence Above	.60	.91	.49	.97
Convergence Below	.63	.86	.59	.97
Notch	.62	.97	.47	.96
Rotation	.49	.96	.56	.92

Table 9

Mean Confidence in Probability Judgments Over Time

Forecaster	Time 1	Time 2	Time 3
A	.40	.50	.53
B	.88	.90	.93
C	.48	.49	.65
D	.84	.80	.86
ALL	.65	.67	.74

Figure 1

Sequence of Phases in Microburst Forecasting

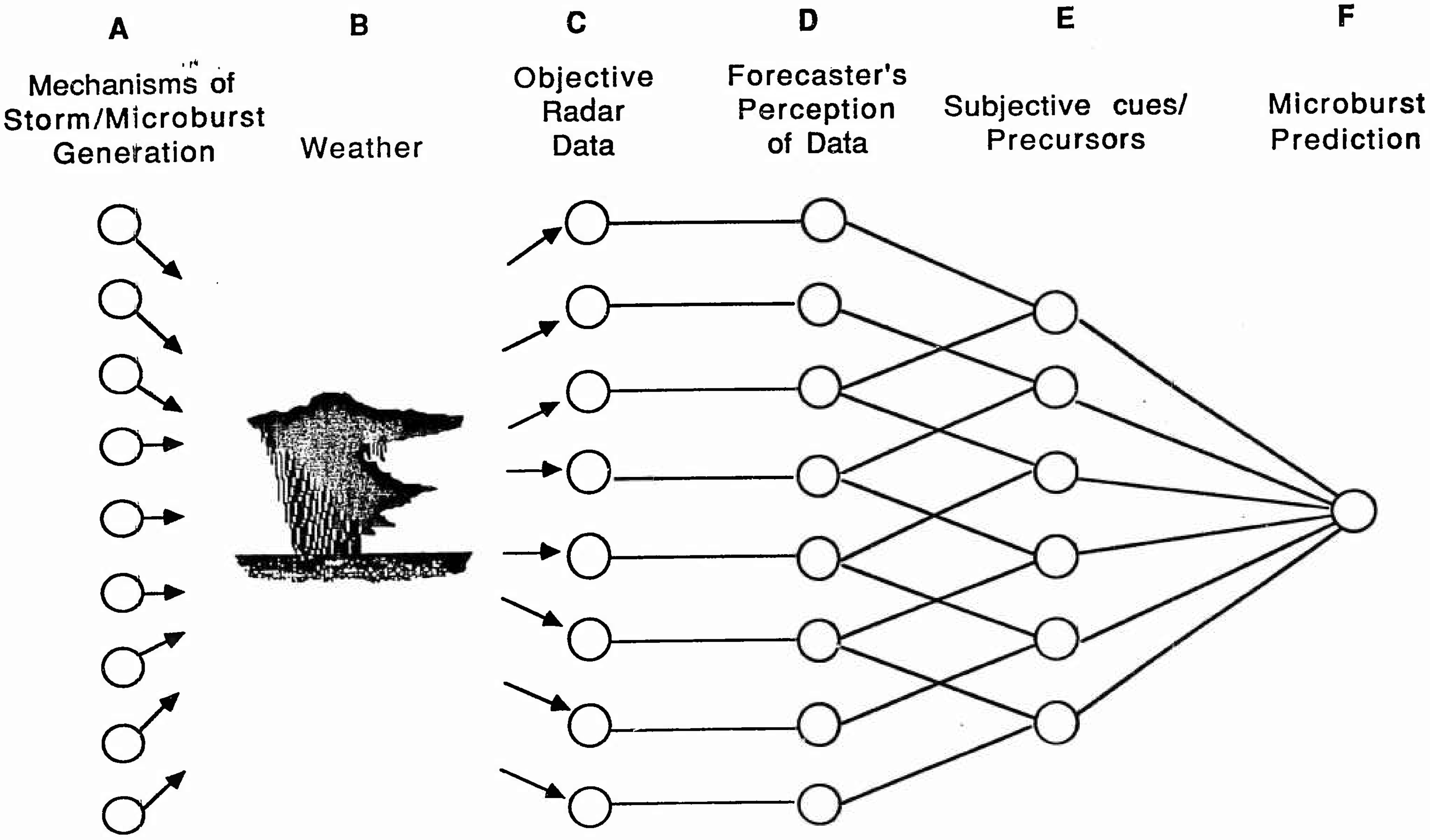
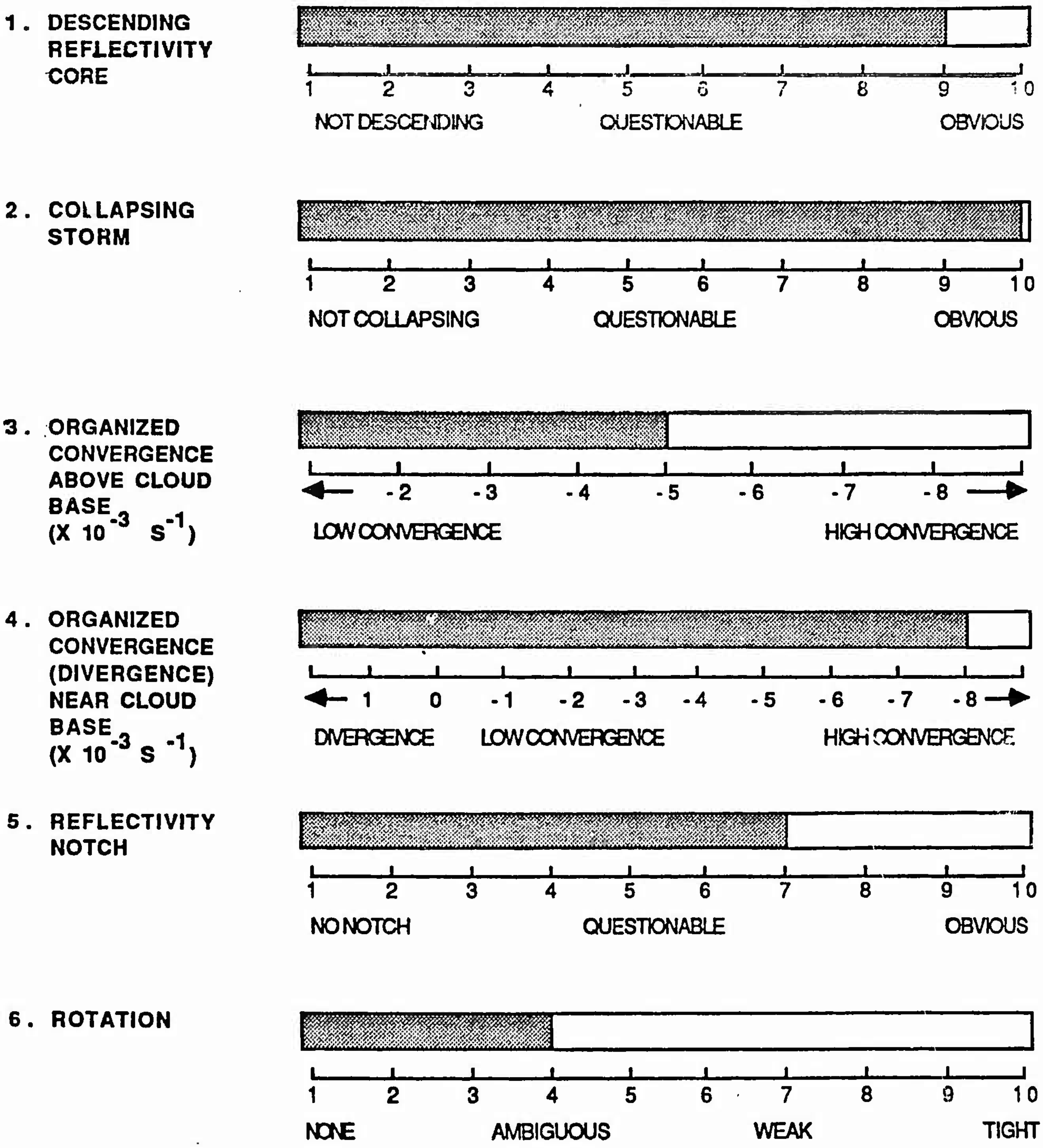

A
Mechanisms of Storm/Microburst Generation
B
Weather
C
Objective Radar Data
D
Forecaster's Perception of Data
E
Subjective cues/ Precursors
F
Microburst Prediction

Figure 2

Example of a Microburst Profile



probability (0.0 - 100) of
microburst within 5 - 10 minutes: ____________

Figure 3

Precursor Judgment Scales

	RATING SHEET	CONFIDENCE (%)
DESCENDING REFLECTIVITY CORE	1 2 3 4 5 6 7 8 9 10 NOT DESCENDING QUESTIONABLE OBVIOUS	________
COLLAPSING STORM	1 2 3 4 5 6 7 8 9 10 NOT COLLAPSING QUESTIONABLE OBVIOUS	________
ORGANIZED CONVERGENCE (DIVERGENCE) ABOVE CLOUD BASE	2 1 0 -1 -2 -3 -4 -5 -6 -7 -8 DIVERGENCE LOW CONVERGENCE $(10ms^{-1}/5km)$ HIGH CONVERGENCE $(40ms^{-1}/5km)$	________
ORGANIZED CONVERGENCE (DIVERGENCE) AT OR BELOW CLOUD BASE	2 1 0 -1 -2 -3 -4 -5 -6 -7 -8 DIVERGENCE LOW CONVERGENCE $(10ms^{-1}/5km)$ HIGH CONVERGENCE $(40ms^{-1}/5km)$	________
REFLECTIVITY NOTCH	1 2 3 4 5 6 7 8 9 10 NO NOTCH QUESTIONABLE OBVIOUS	________
ROTATION	1 2 3 4 5 6 7 8 9 10 NONE AMBIGUOUS WEAK TIGHT	________

probability (0.0 - 1.0) of microburst within 5 - 10 minutes: ________

PLEASE LIST ON THE BACK OF THIS PAGE ANY OTHER FACTORS YOU CONSIDERED WHEN MAKING YOUR MICROBURST JUDGMENT

Appendix III

Lusk, C. M., Stewart, T. R., Hammond, K. R., & Potts, R.

1988

Judgment and decision making in dynamic tasks: The case of forecasting the microburst

(Tech Rep. No. 284)

Boulder: University of Colorado, Center for Research on Judgment and Policy

Judgment and Decision Making in Dynamic Tasks: The Case of Forecasting the Microburst

Cynthia M. Lusk[1], Thomas R. Stewart[2],
Kenneth R. Hammond[1], and Rod Potts[3]

July 1989

[1] Center for Research on Judgment and Policy, University of Colorado at Boulder, Boulder, CO 80309-0344

[2] Center for Research on Judgment and Policy, University of Colorado at Boulder, Boulder, CO 80309-0344 (Current address: Center for Policy Research, The University at Albany, State University of New York, Albany, NY 12222)

[3] Bureau of Meteorology Research Centre, Melbourne, VIC, 3001, Australia

Abstract

Two studies of microburst forecasting were conducted in order to demonstrate the utility of applying theoretical and methodological concepts from judgment and decision making to meteorology. A hierarchical model of the judgment process is outlined in which a precursor identification phase is separated from the prediction phase. In the first study, forecasters were provided with specific, perfectly reliable precursor values and were asked to provide judgments regarding the probability of a microburst. Results indicated that the microburst forecasts were adequately represented by a linear model. Modest agreement was observed among the forecasters' judgments. In the second study forecasters viewed storms under dynamic conditions representative of their usual operational setting. They made judgments regarding precursor values, as well as of the probability of a microburst occurring. The forecasters' agreement regarding microburst predictions was found to be even lower than in the first study. In addition, agreement regarding the (subjectively) most important precursor value was near zero. These results suggest that opportunities to improve forecasting would result from a better understanding of the precursor identification and prediction phases of the forecasting process.

1. Introduction

From the point of view of research and operational meteorologists, the forecasting of any weather phenomenon first requires an understanding of the relevant physical processes which generate a particular weather event (Doswell 1986; Smith et al. 1986). It is generally assumed that the forecaster develops a conceptual model of the phenomenon from an understanding of the physical processes. This conceptual model is then often applied to an operational setting (see, e.g. Mueller et al. 1989; Roberts and Hjelmfeldt 1989).

When involved in the forecast process the operational meteorologist observes, evaluates, and thinks about a stream of weather information, which is continually changing with time. Thus, there are a number of activities which directly involve the cognitive processes of the meteorologist. The data from numerous information sources must be perceived and assimilated by the forecaster. These data must be integrated and their significance for a particular weather event must be assessed. The forecast must then be made, often within strict time limits and with limited information.

Inaccuracies in weather forecasts result because of errors, inconsistencies, or lack of understanding in all of the above (Doswell 1986; Smith et al. 1986). Considerable effort in the past has been placed on improving the basic understanding of the physical processes and the development of conceptual models underlying a particular weather event as well as on providing improved weather information and displays. Yet no research effort has been directed at a better understanding of the cognitive or judgment processes involved. (Although meteorologists do, in fact, recognize the role of human forecasting processes; see, e.g. Doswell 1986; Smith et al. 1986. See Stewart et al. 1989, for an exception.) It is the latter problem to which the present efforts are directed. In particular, the goal of this paper is twofold: 1) to introduce theoretical and methodological concepts from the judgment and decision field, and 2) to apply those concepts to a specific forecasting problem. The research was conducted in the context of severe weather forecasting and in particular was concerned with the forecasting of microbursts--brief, localized wind storms that are a potentially fatal hazard to aircraft.

In the studies reported here, we have investigated cognitive activity at particular stages of a schema that represents both the physical environment of a storm and the perceptual and cognitive activities of the forecaster. The hierarchical model that depicts steps between the environment of a storm and a judgment about microbursts is presented in Figure 1. This framework is derived from social judgment theory (Hammond et al. 1975; Brehmer and Joyce 1988), which describes the relationship between two systems: the task system in the environment and the cognitive system of the decision maker. The environment of the microburst forecasting task is represented as Phases A, B, and C in Figure 1. Phase A represents the physical mechanisms that underlie the weather phenomenon at Phase B. The weather produces objective data from Doppler radar at Phase C. The cognitive system of the forecaster begins operating at Phase D. After reading the radar data (a perceptual task) at Phase D, the forecaster must extract the cues that are hypothesized precursors of microbursts at Phase E and integrate them into a judgment about the occurrence of a microburst at Phase F. The decomposition of the final microburst judgment incorporates and separates perceptual and conceptual cognitive activities. Perceptual activities are represented at Phase D. Forecasters' conceptual understanding of how those data combine to indicate precursors and how the precursors combine to arrive at a microburst prediction are represented in Phases E and F.

Insert Figure 1 about here

The hierarchical nature of the model implies that error at any phase can be passed on to later phases. Therefore, the quality of the final judgment depends on perceptions and judgments at each prior phase. Errors are not only likely to be cumulative but are apt to have nonlinear consequences as well. Anthes' (1986) argument that "any error, no matter how small, will eventually grow and contaminate even a perfect model's forecast" (p. 637), should, of course, be applied to forecasters' perceptions and judgments as well as physical instruments (see Tribbia and Anthes 1987).

Two studies were conducted investigating the different phases of the microburst prediction task. In Study 1, judgment analysis was used to investigate microburst prediction at Phase F. Following the procedures of social judgment theory, we first identified the cues (precursors) that forecasters use to identify microbursts. We then generated a sample of cases representing hypothetical storms. For each case, forecasters judged the probability of a microburst. From analyses of those judgments we determined how the forecasters integrated the cues into a judgment. In addition, we assessed intra- and inter-forecaster consistency in judgments, and individual differences in forecasters' integration of the cues.

In Study 2 we investigated the microburst prediction task in a setting representative of the real-time situation in which forecasters normally operate. Forecasters observed Doppler radar scans of storms over time (some of which produced microbursts and some of which did not) and made judgments regarding precursor values and the probability of a microburst. This study assessed the overall degree of agreement among the forecasters in Phases D, E, and F of Figure 1 in order to determine at which phase in the judgment process disagreement may be occurring. In addition, the effects of updated information over time on agreement was assessed.

Each study is described in turn below. At the end of each study a brief summary of the results and implications are presented. Following the second study is a general summary and discussion of the results and their implications.

2. Study 1: Judgment analysis of microburst nowcasts

We began our research program by investigating the conceptual models forecasters used to make microburst nowcasts when provided with precursor data. This allowed us to determine the degree to which forecasters' judgments agree when they are provided with the same data, and the degree to which disagreement was due to a different conceptual model, and/or due to inconsistency in the application of conceptual model. Finally, the linear model from our judgment analysis was compared to a model provided by the forecasters during a discussion.

The method and analyses reported follow the procedures of social judgment theory (Hammond et al. 1975; Stewart 1988). Another example of applications of these procedures to meteorology may be found in Stewart et al. (1989).

a. Procedure

1) PROBLEM STRUCTURING

Problem structuring, which includes defining the judgment of interest, describing the forecast scenario, identifying the most important cues, defining the cue ranges, and describing relations among the cues, was the focus of an initial meeting with the forecasters. A proposed structure, based on Roberts and Wilson (1987) and on previous discussions with one of the forecasters, was presented. The three forecasters present discussed the proposed structure and suggested a few changes to the scenario and an extension of the range of one cue. They agreed on the problem structure described below.

The judgment. The judgment of interest was defined as the probability (0-100%) that a microburst will be produced by the storm under observation within 5-10 minutes.

The scenario. The judgment scenario describes the conditions leading up to the forecast. It was constructed by fixing the values of certain variables that are expected to influence forecasting strategy. (The effect of the scenario variables on the forecasting strategy is an empirical question that can be addressed by varying scenarios.) The present scenario was described as follows:

1) The morning sounding was favorable for microbursts.

2) There has been a pattern of moderate storms in the vicinity.

3) Microbursts have been observed with other storms earlier in the day.

4) The temperature is still near the convective temperature.

5) The event under observation is a mature storm that is isolated, but possibly multicellular.

6) The level of reflectivity of the event is moderate.

The cues (precursors). The cues included in the storm cases were: 1) descending reflectivity core, 2) collapsing storm top, 3) organized convergence above cloud base, 4) organized convergence/divergence near cloud base, 5) reflectivity notch, and 6) rotation.

The ranges for these cues are presented in Figure 2, an example of a case that the forecasters judged. Abstract scales were used for cues 1, 2, 5, and 6 because physical measures for these features were not available at the time this study was conducted (see Roberts and Wilson 1989, for a discussion of these radar features).

Insert Figure 2 about here

Relations among the cues. The forecasters agreed that no combination of values was physically impossible but that collapsing storms without descending cores were uninteresting when microbursts are being forecast. They also said that divergence near the cloud base would be possible but rare.

Generation of hypothetical cases. Because real microburst cases were not available in sufficient numbers at the time this study was conducted, hypothetical microburst cases were generated. Each case consisted of a different mix of values for the six cues. The properties of the cases conformed to the problem structure previously described.

The POLICY-PC program (Executive Decision Services 1986) was used to generate hypothetical cases. This program generates random-integer cue values with specified ranges. Of 50 cases initially generated, eight were eliminated because they indicated collapsing storms without descending cores (i.e., the value of collapsing storm was more than 3 points higher than the value for descending core).

Eight of the original 50 cases had values of +2 or +1 for the cue convergence near cloud base. These positive values indicate divergence rather than convergence. Because the forecasters had indicated that this would rarely occur, half these cases were selected at random and dropped. Twelve new cases were generated randomly to bring the number of cases back to 50. The resulting intercorrelations among the cues were low; the highest was between collapsing storm and descending core (.31).

The same procedure was used to generate 25 new cases for the second session.

2) COLLECTION OF JUDGMENTS

Each of five forecasters from the National Center for Atmospheric Research (NCAR) in Boulder, Colorado judged 50 cases in individual sessions lasting from 20 to 45 minutes. Approximately one week later the same forecasters judged another set of 50 cases. The second set of 50 consisted of 25 new cases followed by 25 repeated cases from the first session. The 25 new cases were included for cross-validation, that is, so that the model derived from the first session could be tested on a new sample of cases. The 25 repeated cases were the even-numbered cases from the first session presented in a random order. These cases were included to assess the reliability of the forecasters' judgments.

Two forecasters from Lincoln Laboratories (MIT) judged the same cases, but the interval between Session I and Session II for them was a few minutes instead of a week.

b. Results

1) AGREEMENT

Correlations among the seven forecasters' judgments are presented in Table 1. They range from .45 to .90, indicating moderate agreement among forecasters. The correlations indicate that Forecaster E differs from the other forecasters.

Insert Table 1 about here
----

2) CONSISTENCY

Consistency[1] indicates the extent to which a forecaster makes similar judgments when the same information is presented on different occasions. It was measured by correlating the pairs of judgments made on the 25 repeated cases. The consistencies are reported in the last row of Table 1.

The consistencies are moderate to high. Forecaster E has the lowest consistency, which explains in part why his forecasts do not agree with those of the other forecasters. The two forecasters from Lincoln Laboratories (F & G) have the highest consistencies, probably because of a memory effect; their judgments were repeated within a few minutes.

3) JUDGMENT ANALYSIS: REGRESSION MODELS OF JUDGMENTS

Judgment analysis is based on a pervasive finding in research on judgment and decision making: In a variety of fields of expertise, simple algebraic models can reproduce the judgments of experts (Slovic and Lichtenstein 1973; Hammond et al. 1987; Dawes et al. 1989; Brehmer and Joyce 1988). Often a simple linear model predicts the judgments of experts as well as or better than more complex models (Dawes and Corrigan 1974). Judgment analysis (Hammond et al. 1975; Stewart 1988) uses multiple regression analysis to model the judgments of experts.

Models of the following form were statistically fit to the forecasts made by each forecaster:

$$Y_{ij} = c_j + b_{j1}X_{i1} + b_{j2}X_{i2} + b_{j3}X_{i3} + b_{j4}X_{i4} + b_{j5}X_{i5} + b_{j6}X_{i6} + e_{ij}$$

where

Y_{ij} the forecast made by forecaster j based on case i,
c_j a constant for forecaster j,
b_{jk} the weight for cue k,
X_{ik} the value of cue k on case i, and
e_{ij} the residual for forecaster j on case i.

The parameters (c_j and the b_{jk}s) of the model were determined so that the sum of the squared differences between the predictions of the model and the actual forecasts were a minimum; that is, for forecaster j, the sum of the $(e_{ij})^2$ over all the cases is minimized.

The correspondence between the statistical model and the actual forecasts is given by the multiple correlation (R) which can range from 0 to 1, with 1 indicating perfect fit. The squared multiple correlation (R^2) indicates the proportion of variance in the forecasts that is accounted for by the model.

Table 2 shows that the regression models account for 68 to 91% of the variance in the forecasts. In other words, these simple linear models can reproduce the forecasts with a fairly high degree of accuracy and account for most of the consistent variation in forecasts.

Insert Table 2 about here
----

4) RELATIVE WEIGHTS

Relative weights derived from the regression models of each forecaster are presented in Figure 3. These weights, which are based on the standardized regression weights (beta weights) adjusted to sum to 100, indicate the relative importance of each cue to each forecaster. (See the Appendix for details on the derivation of these weights). Six of seven forecasters placed the greatest weight on descending core. The weights for Forecaster E differ substantially from the others. This forecaster placed little weight on the cue descending core and had the largest weight for the cue notch. This pattern of weights explains the differences, apparent in Table 1, between Forecaster E and the other forecasters.

Insert Figure 3 about here
----

5) AGREEMENT BETWEEN COMPONENTS OF JUDGMENTS

The regression model of each forecaster can be used to decompose each forecast into two parts: the linear component, which is the part that is captured by the linear model, and the nonlinear component, the part that is not. Correlations between each of these components of judgment across forecasters are presented in Table 3. The correlations among the linear components of the forecasts (labeled $\underline{G}$ by Hursch et al. 1964) are very high, but the correlations between the nonlinear components of the forecasts (labeled $\underline{C}$ by Tucker 1964) are quite low.

Insert Table 3 about here

The G coefficients in Table 3 measure the agreement that forecasters would achieve if they applied the relative weights described in Figure 3 with perfect consistency. The differences between the values of G and the agreement correlations reported in Table 1 indicate the amount of disagreement due to lack of fit of the linear model to the forecasts.

The C coefficients measure agreement in the nonlinear part of the forecasts. The low values of C indicate that if forecasters are using nonlinear processes to organize the cues into a microburst forecast, the results of those processes differ across forecasters. Until further research is conducted, plausible interpretation of the low C coefficients is that most of the nonlinear component of the forecasts is unreliable, or "error," variance.

6) A NONLINEAR MODEL

When the results of this study were presented to the NCAR forecasters, they insisted that the linear model was not an adequate representation of the way they forecast microbursts. The most important nonlinearity that the forecasters described involved the use of cutoffs on descending core and collapsing storm. They indicated, both in discussion and in writing, that they used a two-stage process in forecasting. If descending core and collapsing storm were low, then the probability of a microburst would be low, regardless of the other cues. On the other hand, high values of descending core and collapsing storm would indicate a downdraft, and the forecasters would look at the other cues to determine the strength of the downdraft.

To test the ability of this nonadditive model to explain the forecasters' judgments, the sample of 75 cases was divided into nine subgroups:

	Collapsing Storm		
Descending Core	Not (1-3)	Questionable (4-7)	Obvious (8-10)
Not (1-3)	A LOW PROBABILITY	B LOW PROBABILITY	C LOW PROBABILITY
Questionable (4-7)	D LOW PROBABILITY	E	F
Obvious (8-10)	G	H	I

(This table was developed by the forecasters.)

Next, for each forecaster, the mean judgment for all cases falling in Cells A, B, C, or D (a total of 18 cases) was calculated. The cell mean was considered the predicted judgment for every case falling in that cell. Then a linear regression equation for the remaining cases, those falling in Cells E, F, G, H, or I was computed. For each forecaster the predicted scores from linear regression were combined with the means to create a variable that includes predicted scores for all 75 cases. The predicted judgments based on this model were correlated with the actual judgments.

A comparison of the correlations between this nonlinear model and the multiple correlations presented in Table 2 showed that the linear model was superior for six of the seven forecasters. For Forecaster F, the nonlinear model and the linear model were equally accurate. Thus, the simple linear model reproduced the forecasters' judgments better than did the nonlinear process that they suggested.

7) SUMMARY

The results of Study 1 show that, when the cues used in forecasting microbursts are specified for the forecasters (rather than perceived), agreement among forecasters was moderate. Further, the forecasting process is adequately described by a simple linear model. Weights derived from that model clarify the relative importance of the cues which, in turn, explain similarities and differences among forecasters. Finally, the simple linear model reproduced the forecasters' judgments as well as or better than did a more complicated linear model that they suggested.

3. Study 2: Judgments of precursors and microbursts in a displaced real-time setting

Study 1 demonstrated a moderate degree of agreement among microburst forecasters and that a linear combination of precursor values will represent microburst forecasts (Phase F of the hierarchical judgment model). Our next step was to investigate judgments regarding the precursor values (Phase E) that are combined to yield the final forecast. In doing so we designed a dynamic situation that was representative of that in which the forecasters typically operate.

a. Procedure

The subjects in this experiment were four of the five NCAR microburst forecasters who participated in Study 1.

The experiment was conducted in two phases. In Phase 1 the forecasters each viewed one microburst and one null case. The procedures were then revised (as described below) in order to increase the amount of data that could be collected in a shorter period of time.

1) OVERVIEW

During each experimental session the forecaster was seated in front of a large computer terminal used to present color Doppler radar displays. The experimenter was seated in front of another computer terminal that was used to run the experimental session. At the first session of each phase of the experiment, the forecasters were presented with instructions regarding how the experiment would proceed. The forecasters were presented with a volume of radar data, after which they made judgments of precursor values and the probability of a microburst. The presentation of data and making of judgments were repeated until completion of each case.

2) THE CASES

Six cases were used to generate the data in this study: two in the first phase and four in the second phase. Half of the cases in each phase were null cases and half were microburst cases.

Each case consisted of a set of radar volume scans (or volumes) of reflectivity and Doppler velocity data, presented chronologically. The volumes each comprised two and one-half minutes of real time data. Each consisted of 13 scans, starting with either the .5 or 1.1 degree elevation scan and terminating with either the 34.8 or the 39.9 degree scan. In the first phase, Case 1 included six volumes. The data for Case 2 spanned eight volumes. However, one volume was skipped due to faulty data. In addition, one volume in Case 2 only included the lower seven scans. However, judgments were still collected for that short volume. In the second phase all cases included four volumes of data. Each case terminated before the microburst was evident on the lowest scan or before any obvious or substantial decrease in the intensity or height of the cell in the null cases.

3) THE JUDGMENTS

The forecasters were asked to make judgments of the six precursor values they had indicated to be the cues in Study 1: descending core, collapsing storm, convergence above cloud base, convergence/divergence at or below cloud base, notch, and rotation. In addition, forecasters made judgments of the probability of a microburst occurring in the next 5 to 10 minutes.

The judgments regarding precursor values and probability of a microburst were made on the same scales as in Study 1. In addition, to the right of each rating scale was a blank for the forecasters to insert their confidence in their precursor judgments. In Phase 1, forecasters' instructions regarding confidence judgments included the following: "Your confidence may be expressed as the probability you believe that your precursor judgment is correct. A zero probability would indicate that you are certain you are not correct and a probability of 100% would indicate you are certain you are correct. A confidence value of 50% indicates your precursor judgment is as likely to be incorrect as correct." In the second phase, the instructions stated: "We would like to clarify what those confidence judgments mean. Your confidence may be expressed on a scale from zero to 100. A zero rating would mean that you have no confidence at all in your judgment, a rating of 100 would mean that you are completely confident, and a rating of 50 would mean that you are half-way in between. A rating of 75 (or 25) would of course indicate greater (or lesser) confidence than the midpoint of 50." The rating sheet is shown in Figure 4.

Insert Figure 4 about here

In the first phase, judgments were made after each volume. Therefore, judgments were made six times for Case 1 and seven times for Case 2. In the second phase judgments were made after all but the first volume. Thus, three judgments were made for each of the four cases in the second phase.

4) THE EXPERIMENTAL SESSION

At the beginning of the first session in each phase, the forecasters were provided with written instructions which explained that each case would consist of several volume scans, over time, of a cell that did or did not produce a microburst, starting with the lowest scan at the earliest time. When they finished observing each scan, the forecasters were instructed to tell the experimenter that they were ready for the next level scan. The forecasters were given up to thirty seconds to view each scan. After completion of a volume in this manner, the forecasters filled in the rating sheet. In addition, the instructions stated, in part:

> At the time of the first volume you can assume that a microburst is not presently occurring. Please assume before observing the first scan, that on the basis of prior information (morning soundings, etc.) you have already reached the conclusion that the likelihood of a microburst on this day is .50. Then adjust your probabilities of a microburst after observing the radar data. Each case will terminate prior to evidence of outflow from a microburst or evidence that the storm is obviously dissipating.

Finally, the forecasters were given instructions to think aloud and their verbalizations were recorded.

The instructions for the second phase explained the changes in the experimental procedure. The forecasters were informed that they would receive the noon sounding data, view only four volumes of data, and make judgments as in the first phase after the second through fourth volumes. In addition, the instructions explained that the scans within each volume would be presented continuously and that they did not need to think aloud.

The forecasters were provided with blank paper for taking notes and felt tip pens to mark the screen. The date for each case was masked on the computer screen. At the beginning of each case, the forecasters were told the coordinates where the cell they were to attend to was presently located.

In the first phase, half of the forecasters were presented with Case 1 first, and half were presented with Case 2 first. In the second phase, the cases were arranged on a tape in a fixed order. Each forecaster began with a different case, but otherwise the order of presentation was fixed.

b. Results

1) OVERALL AGREEMENT AMONG FORECASTERS

Analyses were conducted to determine the degree of agreement between forecasters' judgments of precursor values and agreement between forecasters' judgments of the probability of a microburst. The data used in these analyses were the judgments made after each volume. Thus, 25 data points are possible for each subject (some analyses have a slightly lower number of data points in instances where forecasters did not provide ratings). The correlations between the judgments of each pair of forecasters were computed for each precursor and are presented in Table 4. Similarly, the correlations between judgments of the probability of a microburst were computed and are presented in Table 5.

Insert Tables 4 and 5 about here.

Tables 4 and 5 clearly indicate a lack of agreement between forecasters regarding both the precursor and probability judgments. Although many of the correlations are substantially larger than zero (and are, in fact, statistically significant), they are all substantially smaller than 1.0 or any other level of acceptable agreement.

Comparison of the level of agreement for the different precursors in Table 4 indicates a higher degree of agreement on some precursors than on others. Particularly noteworthy are the low and even negative (!) correlations for judgments of descending core. This result is particularly important because this precursor is the one which forecasters weighted most heavily in arriving at microburst probability judgments (as indicated in Study 1).

Agreement regarding precursor values was highest for the two convergence precursors, second highest for collapsing storm and notch, and lowest for rotation and descending core. The different levels of agreement between precursors are probably due to the different levels of abstraction or stages necessary to make judgments of the precursor values. For example, the two convergence precursors are probably the precursor values most directly obtained (from the radar velocities). In contrast, the descending core judgment requires that the forecaster combine information about maximum reflectivity values over times and heights.

The above results regarding agreement concern judgments at Phase E in our hierarchical model. Given the considerable disagreement regarding precursor values, it is important to determine the extent of agreement regarding perception of the data one step previous in the judgment process at Phase D in order to ascertain whether such disagreement is cumulative upward in the judgment hierarchy. The velocity (in meters per second) and reflectivity (in dBZ's) data from the Doppler radar are both presented as colored images. Three of the four forecasters took extensive notes regarding the dBZ values and these notes were utilized to assess the agreement regarding forecasters' translation from colors to numerical data. The number of times each pair of forecasters agreed and disagreed on the dBz values in their notes was counted (agreement was defined as values within 5 dBZ's of each other). An agreement score for each pair of forecasters was calculated by dividing the number of agreements by the sum of the agreements and disagreements. For Forecasters A and B this score was 96%, 93% for Forecasters A and E, and 99% for B and E. This result indicates that 1) forecasters are in agreement regarding the raw reflectivity values and 2) disagreement occurs when combining these values into precursor judgments.

2) AGREEMENT OVER TIME

Did agreement increase over time? The answer to this question is important because it indicates whether increasing information over time does or does not lead to converging judgments. For each precursor and the microburst probability judgments the correlations were computed between each pair of forecasters for each of the last three volumes separately. The initial data for these analyses were judgments for each of the last three judgment times (volumes) in each case. Specifically, "Time 1" included judgments from volume 4 for Case 1, volume 6 for Case 2 and volume

2 for the other four cases. Likewise "Time 2" included judgments from volume 5 for Case 1, volume 7 for Case 2 and volume 3 for the other cases. Finally, "Time 3" included data from the last volume of each case. This resulted in three correlation matrices for three times. Each correlation was then converted to a Fisher's _z_ [2] and the mean _z_ of each correlation matrix was computed and then converted back to a value for _r_. These mean _r_ values are presented in Table 6.

Insert Table 6 about here.

As can be seen from Table 6, the degree of increased agreement over time varies by precursor. Agreement clearly increases for the precursors collapsing storm, rotation, and the two convergence precursors. But for descending core (the most highly weighted precursor), notch, and the probability of a microburst it is less clear that agreement increases over time.

3) CONFIDENCE IN JUDGMENTS

The mean confidence ratings across all six cases are presented in Table 7 separately for each precursor and probability judgment. (Note that the confidence ratings for the probability judgments were collected only for the last four cases.) It is clear from Table 7 that forecasters were generally at least 50% confident in their judgments, indicating at least some degree of confidence in their ratings. In fact, Forecasters B and E were very confident in their judgments. The importance of these findings and particularly the change in confidence over time are discussed elsewhere

(Lusk and Hammond 1989). In the present context, the most important point is that although the forecasters were making markedly different forecasts, each expressed rather high confidence in the accuracy of their forecasts.

Insert Table 7 about here.
----

c. Summary

Although the conclusions drawn from this study are based on only six cases (25 data points) and must therefore be treated with caution, it is clear that there was a pervasive lack of agreement among the forecasters' judgments of precursor values. It is important to note that the level of measurement at any level in the judgment process (see Figure 1) sets the upper level for accuracy at the final stage of microburst prediction, as is the case for any measuring instrument human or otherwise; (see, e.g. Tribbia and Anthes 1987). Equally important is that these results apparently came as a surprise to the forecasters themselves. Some of the forecasters were clearly very confident in their judgments. However, there is no evidence to suggest that the forecasters had ever made any attempt to ascertain whether their judgments coincided. Indeed, there is no indication that such studies of forecasters' agreement have ever been carried out within the meteorological profession. In short, either false assumptions about inter-forecaster agreement are pervasive in meteorology, or meteorologists are simply indifferent to the significance of inter-forecaster disagreement.

4. General discussion

Our goal in this research was to study cognitive processes underlying microburst forecasting and to discover how research in judgment and decision making could help forecasters provide better forecasts. Thus, we devised a conceptual framework (see Figure 1) of the phases involved in arriving at a judgment of the probability that a storm will produce a microburst. The forecasters' perception of the radar data (Phase D), extraction of the cues (Phase E), and judgment of a microburst (Phase F) were studied. Study 1 provided a "best case scenario" for the forecasters. That is, if it is true that the best available precursors of microbursts are those identified by the forecasters and used in the study, then the forecasters were making judgments on the basis of error-free information because the forecasters did not have to determine the cue values perceptually. From Study 1 we learned that a simple linear model is a good descriptor of forecasters' judgments. In addition, the results of Study 1 indicated that when provided with a best case scenario the forecasters disagreed with one another regarding their microburst judgments and demonstrated inconsistency in their own judgments.

Study 2 indicated an even greater degree of disagreement regarding microburst judgments than did Study 1. This is contrary to what many people outside the judgment and decision making field might expect (but consistent with what those inside the field would expect), especially those who assert that profile cases (as used in Study 1, see Figure 2) "rob" subjects of information they typically use when making judgments. Study 2 was designed in a manner to be representative of the real time situation in which the forecasters normally operate. But their performance was worse than with the profile cases (Study 1). Fortunately, the design of the research (specifically the collection of cue judgments) made available the data appropriate for determining the basis of the disagreement. Utilizing the hierarchical model, the degree of disagreement was traced back through successive nodes in the judgment hierarchy. Considerable disagreement was found regarding precursor judgments at Phase E, with the amount of disagreement varying by precursor. Moving back one phase further in the judgment process to Phase D considerable agreement was found regarding forecasters' perceptions of reflectivity values. For example, when the objective radar data (Phase C) is perceived by the forecasters (Phase D), they perceive the same reflectivity values. However, when those values are integrated over time and height to generate a judgment of descending core, considerable disagreement appeared.

After reviewing these clearly undesirable results, we recommended that the forecasters increase agreement by constructing clear operational definitions of each precursor, a step apparently not taken earlier. The procedure for producing definitions should include first, scientific knowledge of physical mechanisms, and second, a framework outlining the phases from those mechanisms to the judgment of a precursor value similar to the one we have been utilizing. That is, once an explicit, public theoretical definition of the physical processes has been agreed upon, a model should then be developed that describes how the physical mechanisms are manifested on the radar displays, and how those mechanisms, once perceived by a forecaster, can provide data for each microburst precursor.

Then at each phase of the model, empirical tes's of agreement should be employed (as in our research). Finally, a formal training program could be established which utilizes the model developed during the definition process in training exercises that track performance and provide feedback. The "replay" capacity of the modern workstation brings these steps within practical means.

In addition to training exercises, a formal research program should be established to study the cognitive processes underlying meteorological judgments. Although meteorologists recognize the importance of the process by which forecasts are generated, they have not made use of, indeed, show no evidence of being aware of, the roughly forty years of research on judgment and decision making. For example, Doswell (1986) offers a model of the human forecasting process, and in doing so makes many assertions about that process. Yet he offers no empirical support for them (except for one reference to Allen's 1981 pioneering work). Researchers in the field of cognitive science would find these assertions to be as amateur as a layperson's weather forecasts. On the other hand, Smith et al. (1986) acknowledge the role of "a growing science based on decision theory (that) seems likely to help forecasters arrive at more objective decisions" (p. 43), but fails to direct the reader to the literature where research on "decision theory" might be found (e.g., Arkes and Hammond 1988; Baron 1988; Hogarth 1980; Kahneman et al. 1982).

The purpose of this research process was twofold: First, it was intended to assist cognitive psychologists to increase their understanding of judgment and decision making processes in dynamic (i.e., changing) task conditions, and second, to assist meteorologists to increase their understanding of the cognitive aspects of meteorological judgments with the aim of improving those judgments. It is our view that these goals were achieved to a limited but not trivial extent. Results pertinent to psychological research on dynamic tasks are discussed elsewhere (Lusk and Hammond 1989), but it is clear that specific information that would not have been otherwise obtained was brought to bear on a specific forecasting problem. These results pointed directly to steps that could be taken to improve the process. Perhaps the most significant contribution of this research to meteorology is the demonstration of the utility of applying theoretical and methodological concepts of judgment and decision making to the forecasting process.

REFERENCES

Allen, G., 1981: Aiding the weather forecaster: Comments and suggestions from a decision analytic perspective. *Aust. Meteor. Mag.*, 29, 25-29.

Anthes, R. A., 1986: The general question of predictability. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., American Meteorological Society, 636-656.

Arkes, H. R. and K. R. Hammond, (Eds.), 1986: *Judgment and Decision Making*. Cambridge University Press, 812 pp.

Baron, J., 1988: *Thinking and Deciding*. Cambridge University Press, 516 pp.

Brehmer, B. and C. R. B. Joyce, 1988: *Human Judgment: The SJT View*. Elsevier Science Publishers B. V., 520 pp.

Cohen, J., and P. Cohen, 1983: *Applied Multiple/Regression Analysis for the Behavioral Sciences* (2nd Ed.). Lawrence Erlbaum Associates, 545 pp.

Dawes, R. M., and B. Corrigan, 1974: Linear models in decision making. *Psych. Bull.*, 81, 95-106.

Dawes, R. M., D. Faust and P. E. Meehl, 1989: Clinical versus actuarial judgment. *Science*, 243, 1668-1673.

Doswell, C. A. III, 1986: Short-range forecasting. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., American Meteorological Society, 689-719.

Executive Decision Services, 1986: POLICY PC Reference Manual, Version 2.0. Albany, New York.

Hammond, K. R., R. M. Hamm, J. Grassia and T. Pearson, 1987: Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Trans. on Sys., Man, and Cyb.*, SMC-17(5), 753-770.

Hammond, K. R., T. R. Stewart, B. Brehmer, and D. Steinmann, 1975: Social judgment theory. *Human Judgment and Decision Processes: Formal and Mathematical Approaches*, M. F. Kaplan and S. Schwartz, Eds., Academic Press, 271-312.

Hogarth, R. M., 1980: *Judgment and Choice: The Psychology of Decision*. Wiley, 250 pp.

Hursch, C. J., K. R. Hammond, and J. L. Hursch, 1964: Some methodological considerations in multiple-cue probability studies. *Psych. Rev.*, 71, 42-60.

Judd, C. M., and G. H. McClelland, 1988: *Data Analysis: A Model Comparison Approach*. Harcourt Brace Javonovich, 528 pp.

Kahneman, D., P. Slovic, and A. Tversky, 1982: *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 555 pp.

Lusk, C. M., and K. R. Hammond, 1989: Judgment in a dynamic task: Nowcasting the microburst. University of Colorado, Center for Research on Judgment and Policy, Tech. Rep. 288, 53 pp.

Mueller, C. K., J. W. Wilson, and B. Heckman, 1989: Evaluation of the TDWR aviation nowcasting experiment. Preprints, *Third International Conference on the Aviation Weather System*, Anaheim, Amer. Meteor. Soc., 212-216.

Roberts, R. D., and M. R. Hjelmfeldt, 1989: Evaluation of microburst nowcasting during TDWR 1987. Preprints, *Third International Conference on the Aviation Weather System*, Anaheim, Amer. Meteor. Soc., 206-211.

Roberts, R. D., and J. W. Wilson, 1987: Nowcasting microbursts using Doppler radar in a forecaster-computer environment. *Proc. Symposium Mesoscale Analysis and Forecasting*, Vancouver, European Space Agency SP-282, 43-48.

Roberts, R. D. and J. W. Wilson, 1989: A proposed microburst nowcasting procedure using single-Doppler radar. *J. of App. Meteor.*, 28, 285-303.

Slovic, P., and S. Lichtenstein, 1973: Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Human Judgment and Social Interaction*, L. Rappoport and D. A. Summers, Eds., Holt, Rinehart & Winston, 16-108.

Smith, D. L., F. L. Zuckerberg, J. T. Schaefer, and G. E. Rasch, 1986: Forecast problems: The meteorological and operational factors. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., Amer. Meteor. Soc., 36-49.

Stewart, T. R., 1988: Judgment analysis: Procedures. Human Judgment: The SJT View. B. Brehmer and C. R. B. Joyce, Eds., Elsevier Science Publishers B. V., 41-74.

Stewart, T. R., W. R. Moninger, J. Grassia, R. H. Brady, and F. H. Merrem, 1989: Analysis of expert judgment and skill in a hail forecasting experiment. Wea. and Forecasting, 4, 24-34.

Tribbia, J. J., and R. A. Anthes, 1987: Scientific basis of modern weather prediction. Science, 237(4814), 493-499.

Tucker, L. R., 1964: A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. Psych. Rev., 71, 528-530.

APPENDIX

The raw regression coefficients and their standard errors for each forecaster are presented in Table A. In their raw form, it is difficult to directly compare the regression coefficients to ascertain different weightings of cues because the values of the coefficients are in the original scale units and the cues were measured on different scales. Therefore we transformed the regression coefficients into relative weights so that comparisons can be made between cues and between forecasters. Computation of the relative weights proceeded as follows. First the standardized regression coefficients (the beta weights) were computed. The standardized form of the regression equation compensates for differences in units by transforming each variable so that its mean is 0.0 and its variance is 1.0 in the sample. Then we computed relative weights. For each forecaster, the beta weights were summed and each beta weight was divided by that sum. Finally, each relative weight was multiplied by 100, which makes the interpretation of the relative weights much clearer than either the raw or beta weights.

Insert Table A about here

Footnotes

[1] The term "consistency" in this paper refers to what psychologists mean by "reliability". Consistency is used here to avoid confusion of reliability with the term "calibration".

[2] When correlations are used as dependent variables it is recommended (Cohen & Cohen, 1983; Judd & McClelland, 1988) that they first be transformed to Fisher's _z_.

AUTHOR NOTE

This research was supported by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142 and by the National Center for Atmospheric Research. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of the Army or NCAR position, policy, or decision, unless so designated by other official documentation.

Table 1

Study 1: Agreement and Consistency Correlation Coefficients

	A	B	C	D	E	F	G
B	.78						
C	.82	.81					
D	.75	.72	.76				
E	.52	.49	.45	.64			
F	.80	.90	.84	.75	.45		
G	.85	.80	.81	.71	.45	.81	
Consistency	.81	.92	.79	.89	.76	.95	.98

Table 2

Study 1: Multiple Correlations

Forecaster	R	R^2
A	.89	.79
B	.92	.84
C	.89	.80
D	.86	.73
E	.83	.68
F	.95	.91
G	.88	.77

Table 3

Study 1: Agreement
Linear Additive Component (G)

	A	B	C	D	E	F
B	.93					
C	.96	.96				
D	.93	.87	.92			
E	.71	.59	.64	.85		
F	.93	.98	.94	.85	.53	
G	1.00	.93	.97	.93	.69	.94

Study 1: Agreement
Nonlinear or Nonadditive Component (C)

	A	B	C	D	E	F
B	.09					
C	.27	.12				
D	.19	.19	.23			
E	.02	.19	-.09	.14		
F	.05	.38	.29	.37	.21	
G	.33	.29	.25	.04	-.18	.18

Table 4

Study 2: Agreement Correlation Coefficients for Judgments of Precursors

Descending Core

	A	B	D
B	.14		
D	-.06	.12	
E	.10	.35	-.14

Collapsing Storm

	A	B	D
B	.69		
D	.47	.53	
E	.57	.40	.17

Convergence Above Cloud Base

	A	B	D
B	.65		
D	.71	.49	
E	.58	.53	.45

Convergence at/or Below Cloud Base

	A	B	D
B	.54		
D	.43	.76	
E	.77	.59	.45

Notch

	A	B	D
B	.38		
D	.51	.25	
E	.61	.57	.34

Rotation

	A	B	D
B	.06		
D	.12	.39	
E	.51	-.01	.26

Table 5

Study 2: Agreement Correlation Coefficients for Judgments of Probability of a Microburst

	A	B	D
A	.60		
B	.88	.45	
D	.31	.15	.19

Table 6

Study 2: Mean Agreement Correlation Coefficients Over Time

	Time 1	Time 2	Time 3
Descending core	.095	.185	.170
Collapsing Storm	.290	.360	.810
Convergence Above	.435	.680	.745
Comvergence Below	.515	.745	.855
Notch	.690	.580	.830
Rotation	-.055	.340	.495
Probability	.405	.540	.605

Table 7

Study 2: Mean Confidence Ratings

	A	B	D	E	ALL
Descending core	.43	.95	.54	.79	.68
Collapsing Storm	.50	.92	.50	.79	.68
Convergence Above	.57	.90	.50	.95	.73
Convergence Below	.62	.89	.60	.95	.77
Notch	.60	.95	.50	.94	.75
Rotation	.51	.97	.56	.92	.74
Mean	.54	.93	.53	.89	
Probability	.48	.90	.54	.83	.69

Table A

Unstandardized Regression Coefficients (Standard Error in Parentheses)

	A		B		C		D		E		F		G	
Descending Core	7.38	(.64)	9.49	(.59)	5.09	(.38)	4.96	(.67)	1.14	(.57)	9.71	(.43)	6.51	(.58)
Collapsing Storm	1.74	(.60)	1.42	(.56)	0.72	(.36)	3.39	(.63)	2.76	(.54)	1.00	(.41)	1.28	(.55)
Convergence Above	3.61	(.63)	1.34	(.58)	0.78	(.38)	1.98	(.67)	3.22	(.57)	1.59	(.43)	2.80	(.58)
Convergence Near	1.34	(.50)	1.26	(.46)	1.51	(.30)	2.36	(.52)	2.41	(.44)	0.49	(.33)	1.26	(.45)
Notch	-.39	(.53)	1.02	(.49)	-.29	(.32)	0.75	(.56)	1.83	(.48)	0.98	(.36)	-.23	(.49)
Rotation	1.89	(.61)	0.46	(.56)	1.09	(.37)	2.42	(.64)	0.38	(.55)	2.25	(.41)	2.26	(.56)
Constant	-51.44	(8.06)	-26.61		5.05		-46.24		4.34		-49.13		-36.01	
				(7.46)		(4.85)		(8.43)		(7.21)		(5.49)		(7.35)

FIGURE CAPTIONS

Figure 1. Sequence of phases in microburst forecasting.

Figure 2. Example of a microburst case.

Figure 3. Study 1: Relative weights.

Figure 4. Study 2: Precursor judgment scales.

Figure 1

Sequence of Phases in Microburst Forecasting

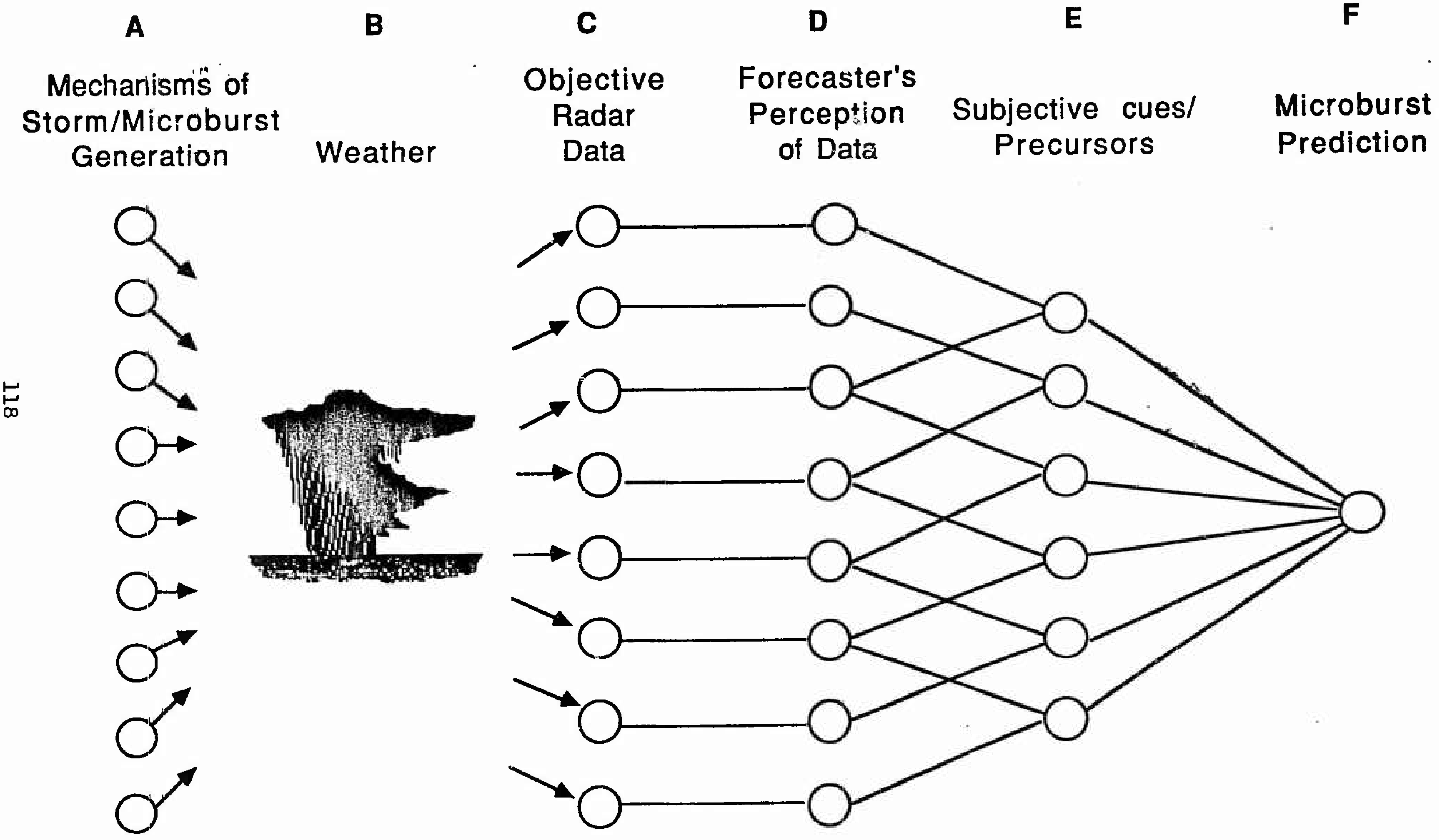
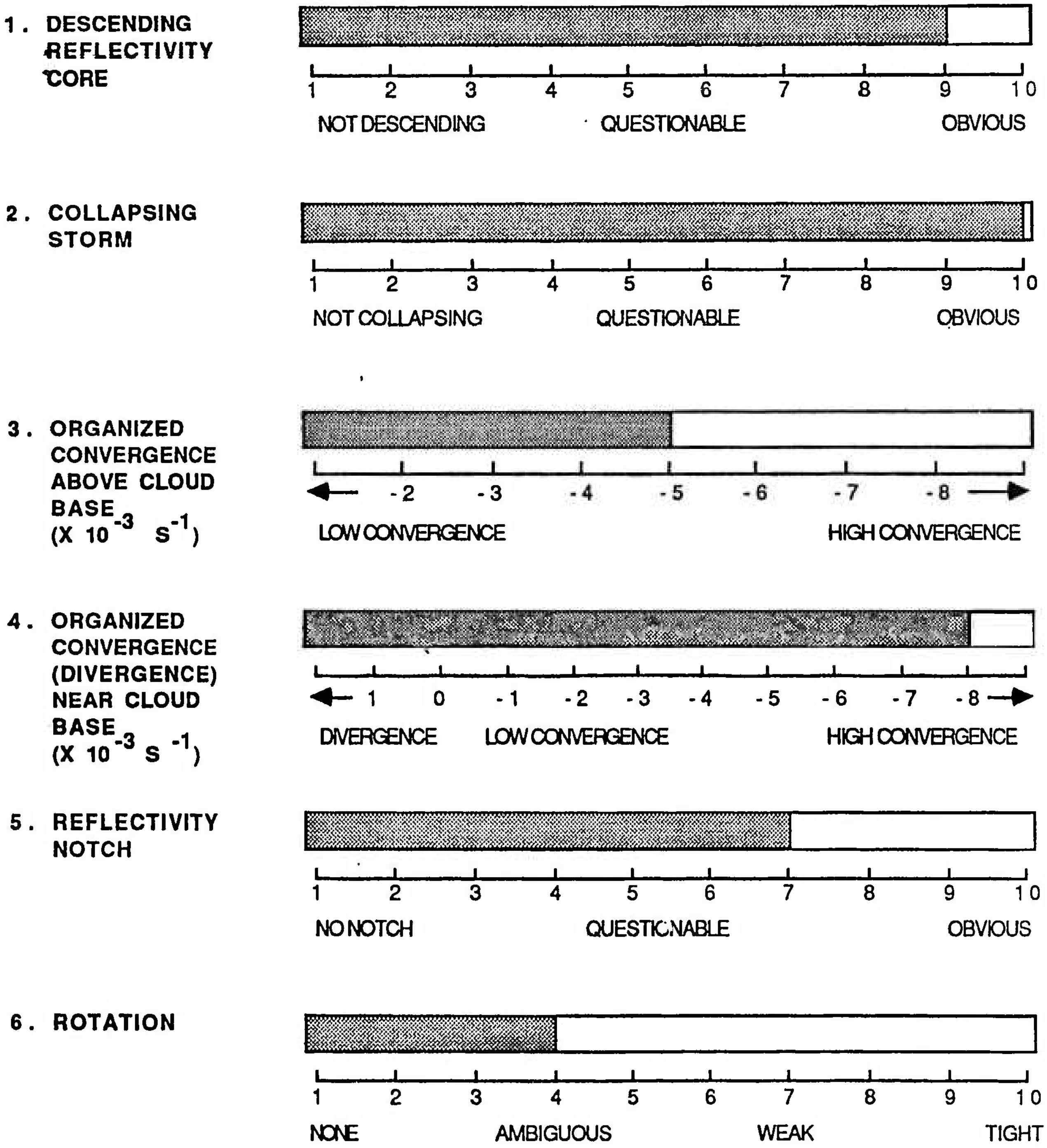

A
Mechanisms of Storm/Microburst Generation
B
Weather
C
Objective Radar Data
D
Forecaster's Perception of Data
E
Subjective cues/ Precursors
F
Microburst Prediction

Figure 2

Study 1: Example of a Microburst Case



probability (0.0 - 100) of
microburst within 5 - 10 minutes: ____________

Figure 3

Study 1: Relative Weights

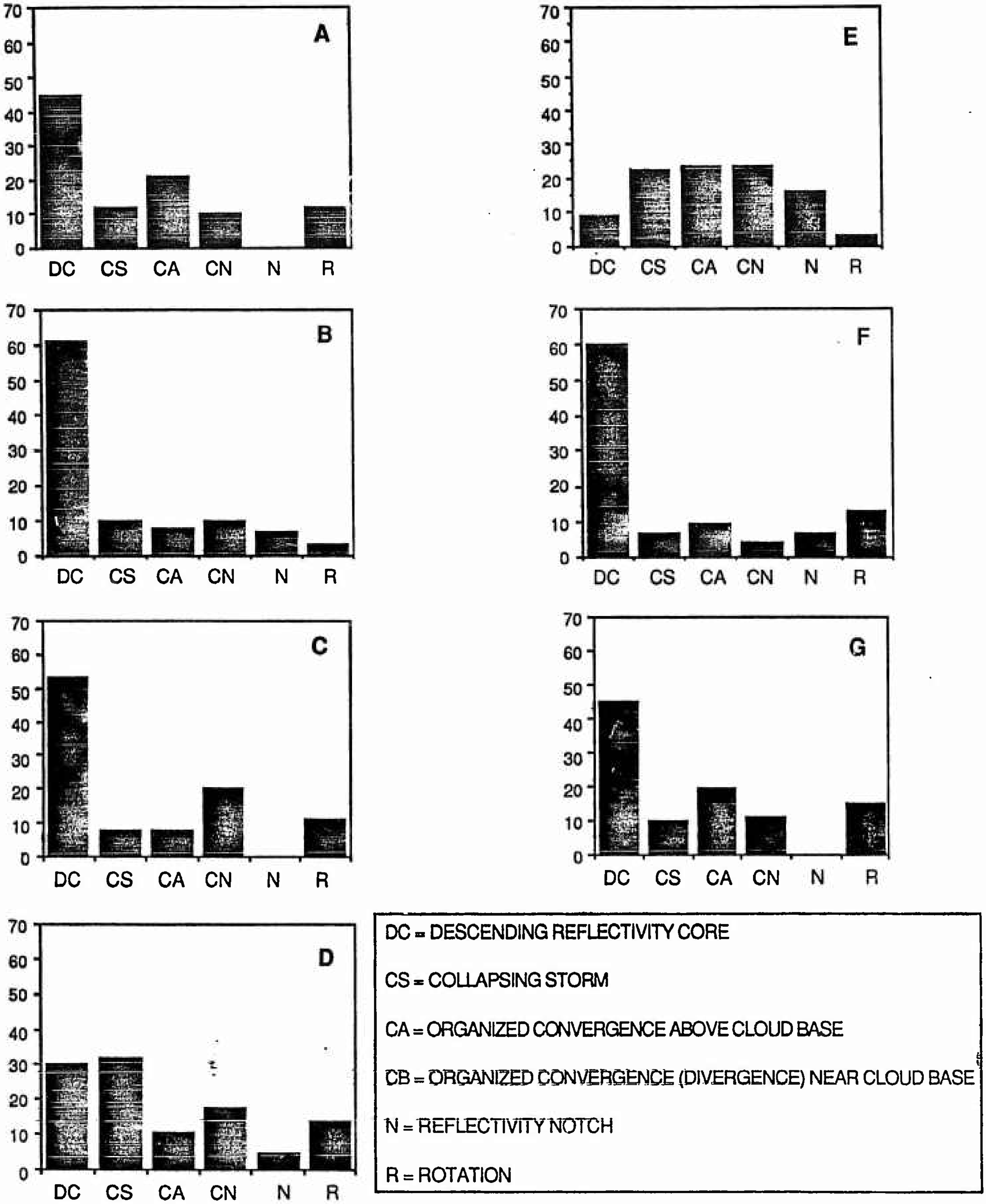


Figure 4

Study 2: Precursor Judgment Scales

RATING SHEET

CONFIDENCE (%)

DESCENDING REFLECTIVITY CORE

1 2 3 4 5 6 7 8 9 10

NOT DESCENDING — QUESTIONABLE — OBVIOUS

____

COLLAPSING STORM

1 2 3 4 5 6 7 8 9 10

NOT COLLAPSING — QUESTIONABLE — OBVIOUS

____

ORGANIZED CONVERGENCE (DIVERGENCE) ABOVE CLOUD BASE

2 1 0 -1 -2 -3 -4 -5 -6 -7 -8

DIVERGENCE — LOW CONVERGENCE ($10ms^{-1}$ /5km) — HIGH CONVERGENCE ($40ms^{-1}$ /5km)

____

ORGANIZED CONVERGENCE (DIVERGENCE) AT OR BELOW CLOUD BASE

2 1 0 -1 -2 -3 -4 -5 -6 -7 -8

DIVERGENCE — LOW CONVERGENCE ($10ms^{-1}$ /5km) — HIGH CONVERGENCE ($40ms^{-1}$ /5km)

____

REFLECTIVITY NOTCH

1 2 3 4 5 6 7 8 9 10

NO NOTCH — QUESTIONABLE — OBVIOUS

____

ROTATION

1 2 3 4 5 6 7 8 9 10

NONE — AMBIGUOUS — WEAK — TIGHT

____

probability (0.0 - 1.0) of
microburst within 5 - 10 minutes: ____________

PLEASE LIST ON THE BACK OF THIS PAGE ANY OTHER FACTORS
YOU CONSIDERED WHEN MAKING YOUR MICROBURST JUDGMENT

Appendix IV

Lusk, C. M., Mross, E. F., & Hammond, K. R

1989

Judgment and decision making under stress: A preliminary study of convection forecasts

(Tech Rep. No. 287)

Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

JUDGMENT AND DECISION MAKING UNDER STRESS: A PRELIMINARY STUDY OF CONVECTION FORECASTS

Cynthia M. Lusk, Ernest F. Mross, and Kenneth R. Hammond

Center for Research on Judgment and Policy

Institute of Cognitive Science

University of Colorado
Boulder, Colorado 80309

Report No. 287

June 1989

Working paper
Please do not cite or quote without
permission from authors

This work was supported in part by the Office of Basic Research, Army Research Institute, Contract MDA903-86-C-0142, Work Unit Number 2Q161102B74F. The views, opinions, and findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

The analyses reported here are based upon a subset of data collected during the Terminal Doppler Weather Radar (TDWR) experiment in the Denver, Colorado area during the summer of 1988 by meteorologists at the National Center for Atmospheric Research.

The Forecasts

A full description of the forecasting environment is presented in Mueller, Wilson, and Heckman (1989) and only a brief description will be provided here. A team of two to three forecasters made consensus forecasts each hour between noon and 7 p.m. The forecasts were probability forecasts. A separate forecast was made for a 10 km. radius circle centered on Stapleton International Airport and the same size circle for the Kiowa gateway. At each forecast time, forecasts were made for convection at the 30 dBZ level and the 50 dBZ level. In addition, at each forecast time for each dBZ level four forecasts were made: one for the first 15 minutes after the hour, one for the 15-30 minute time period, one for the 30-45 minute time period and one for the 45-60 minute time period. The data analyzed in our study include only the 30 dBZ forecasts and verification data for the two locations combined. Thus for each forecast period (0-15 min., 15-30 min., etc.) the full sample in these analyses includes 521 forecasts spanning 36 days.

Operationalization of Stress: Construction of Low and High Stress Samples

A common operationalization of stress in the experimental literature is time pressure (Ben Zur & Breznitz, 1981; Payne, Bettman

& Johnson, 1988; Rothstein, 1986; Schwartz & Howell, 1985). We reasoned that on days when the environment was active forecasters had more data they must attend to, resulting in less time to produce forecasts. Therefore we identified low and high activity days as our operationalization of stress. Specifically, for each forecast time we determined the number of forecasts with a probability of 20 or above (0 forecasts if all the forecasts for both Stapleton and Kiowa were below 20 percent; 1 forecast if either Stapleton and Kiowa, but not both, had one forecast of 20 percent or above; 2 forecasts if both Stapleton and Kiowa had at least one forecast of 20 percent or above). The number of forecasts of 20 percent or above were then summed over all the forecasts for that day. This sum represents the amount of stress occurring on a given day. The days were then divided into low and high stress days according to a median split of the stress variable sum (and excluding all days with a sum of 0). This resulted in 194 forecasts spanning 13 days in each sample.

Results

Three different measures of performance were computed, each yielding different information: (a) contingency table skill indices, (b) the skill score and its decomposition into linear and different bias components, and (c) signal detection theory analyses. Each type of analysis was first applied to the sample as a whole, then to the low and high stress samples separately. The results for each type of measure are summarized in Tables 1 through 6 and are discussed briefly below.

Contingency table skill measures. Construction of the contingency tables and definitions of the skill indices are provided in Mueller et al. (1989) and Donaldson, Dyer, and Kraus (1975). The data included in Mueller et al. (1989) include only Stapleton data, while those reported here include the forecasts from both Stapleton and Kiowa. The results for the full sample are presented in Table 1. Table 1 reveals a general trend in the data: Forecasters do very well on forecasts that are more immediate, with skill declining as the time period for which they are forecasting becomes more distant. Table 2 presents the results for the low and high stress days separately. The skill indices indicate that forecasters tend to perform better on more active high stress days than on less active low stress days.

Decomposition of skill scores. One means of investigating sources of errors in judgment is to decompose the skill indicated in those judgments into (a) the linear relationship between predictions and observed values, (b) unconditional bias, and (c) conditional bias, as suggested by Murphy (1988) and Stewart (1989). The skill score that is decomposed is the Brier score or the mean-square-error between the forecasts and observations. The linear relationship is measured by r-squared, where r is the correlation between forecasts and observations. The conditional bias is a measure of nonsystematic bias in the forecasts. It is related to the slope of the regression line (it is 0 if the slope is 1). The unconditional bias reflects the systematic bias in the forecasts, and is related to the intercept of the regression line (see Murphy, 1988; Stewart, 1989 for further descriptions of these measures).

The results of this decomposition are presented in Tables 3 and 4. In the sample as a whole (Table 3), the skill scores and the squared correlations parallel the results of the above contingency table skill indices (better performance for forecasts closer to the forecast time). The conditional bias and unconditional bias measures are very low, indicating that in the full sample these are not sources of bias in forecasting. We were not surprised to find these measures low because the forecasts were generated by consensus and through this procedure forecasters may remove the biases that could be exhibited by an individual forecaster.

However, we believed that decomposition of skill into the above components might yield insight into the nature of the performance decrement on low stress days indicated in the above contingency table analyses (Table 2). As Table 4 indicates, the skill scores and correlation coefficients indicate better performance under high compared to low stress. In addition, for the low stress days the skill scores were in many cases *negative*. These negative skill scores are due in part to large conditional bias measures. The conditional bias score reflects the extent to which variability in the forecasts is larger than it should be, given the correlation. Thus, a major cause of decrement in skill under low stress conditions may be due to conditional bias.

Signal detection analyses. Signal detection theory (SDT) yields two measures of interest: d' and β. D' represents the degree to which two mutually exclusive events (e.g., weather events and nonevents) can be discriminated from each other; β represents the decision criterion (or "response bias") which is applied to this decision (i.e., the subjective criterion for saying, "Yes, there's something out there" vs. "No, there is nothing out there"). It is important to note that *d'* is *completely independent* from β. That is, *d'* is a measure of the ability to perform a task (e.g., distinguishing convection from nonconvection) which is independent from the criterion (β) used to say "Yes" in the task. That independence is lacking between the skill indices (FAR and POD) reported above (Table 1 and 2). Thus, these SDT measures should further clarify differences in the low and high stress samples.

Table 5 presents *D(A)* and β for the full sample (*D(A)* was used as a measure of *d'* because it has the fewest assumptions regarding distributions from which the samples were drawn). As with the previous measures (Tables 1 and 3), diminishing discriminability is indicated as the forecast period becomes more removed.

With regard to the effects of stress, the SDT measures allow examination of two hypotheses about the effects of stress on forecasters. One hypothesis states that forecasters "pay more attention to what's going on" on high stress days. If this is true, then it follows that forecasters should be better at predicting meteorological events under high stress. As with the other measures, there is support for this hypothesis. The *D(A)*'s in the high stress condition are larger than their respective low stress counterparts for each forecast period. Thus it appears that forecasters are better at discriminating convection from non-convection on high stress days.

Mueller (informal communication) has suggested that different forecasting processes occur on low and high stress days. This suggestion can be addressed with reference to β. That is, on high stress (activity) days forecasters tend to predict less activity than is occurring, while on low stress (activity) days forecasters tend to predict more activity than is occurring. Thus on high stress days forecasters are *less* likely to say "yes" and on low stress days forecasters are *more* likely to say "yes." This difference would be reflected in a lower criterion for saying yes (i.e., a lower β) on high stress than low stress days. This hypothesis receives mixed support. In two comparisons (the 0-15 and 30-45 forecast periods) β is lower in the high stress samples, indicating that the forecasters are more likely to say "yes," whereas in the other two comparisons β is higher in the high-stress cells, indicating that forecasters are not more inclined to respond "yes" on high stress days.

Summary and Conclusion

In sum, all of the results presented indicate a decrement in performance on low stress (activity) days compared to high stress (activity) days. The bias measures (Table 4) indicate that the decrement may be due, in part, to larger judgmental biases occurring during low stress days. In addition, there is some evidence (Table 6) that forecasters use a higher criterion (β) under low stress than high stress conditions. More research is necessary to clarify and expand these findings. Although the present data indicate forecasters may introduce bias into their judgments or a different decision criterion may be operating on low stress days, the processes accounting for the differences are unknown.

References

Ben Zur, M., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica, 47,* 89-104.

Donaldson, R. J., Dyer, R. M., & Kraus, M. J. (1975). An objective evaluator of technique for predicting severe weather events. In *Preprints of the 9th Conference on Severe Local Storms* (pp. 321-326). Boston: AMS.

Mueller, C. K., Wilson, J. W., & Heckman, B. (1989). Evaluation of the TDWR aviation nowcasting experiment. In *Preprints of the 3rd International Conference on the Aviation Weather System.*

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationship to the correlation coefficient. *Monthly Weather Review, 116,* 2417-2424.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534-552.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes, 37,* 83-92.

Schwartz, D. R., & Howell, W. C. (1985). Optional stopping performance under graphic and numeric CRT formatting. *Human Factors, 27,* 433-444.

Stewart, Thomas R. (1989). *Seven components of forecasting skill.* Unpublished manuscript.

Table 1

Contingency Table Skill Scores

Forecast	POD	FAR	CSI
0-15 minutes	.82	.18	.69
15-30 minutes	.69	.31	.52
30-45 minutes	.64	.38	.46
45-60 minutes	.55	.40	.40

Table 2

Contingency Table Skill Scores
Low & High Stress

Low Stress

Forecast	POD	FAR	CSI
0-15 minutes	.57	.20	.50
15-30 minutes	.56	.58	.31
30-45 minutes	.55	.65	.27
45-60 minutes	.50	.64	.26

High Stress

Forecast	POD	FAR	CSI
0-15 minutes	.86	.18	.73
15-30 minutes	.71	.23	.59
30-45 minutes	.67	.26	.54
45-60 minutes	.56	.30	.45

Table 3

Murphy's Decomposition of Skill Scores

Forecast	Skill Score	r^2	Conditional Bias	Unconditional Bias
0-15 minutes	.66041	.66048	.00006	.00000
15-30 minutes	.51953	.51970	.00005	.00012
30-45 minutes	.35808	.36252	.00424	.00020
45-60 minutes	.36591	.36869	.00163	.00115

Table 4

Murphy's Decomposition of Skill Scores Low & High Stress

Low Stress

Forecast	Skill Score	r^2	Conditional Bias	Unconditional Bias
0-15 minutes	.33173	.37405	.03254	.00979
15-30 minutes	.16149	.26832	.08713	.01970
30-45 minutes	-.04792	.13935	.15647	.03080
45-60 minutes	-.09634	.12048	.16534	.05148

High Stress

Forecast	Skill Score	r^2	Conditional Bias	Unconditional Bias
0-15 minutes	.69380	.69672	.00117	.00175
15-30 minutes	.54701	.55801	.00235	.00865
30-45 minutes	.40243	.40883	.00062	.00578
45-60 minutes	.42032	.42863	.00417	.00414

Table 5

SDT Analysis

Forecast	D(A)	β
0-15 minutes	2.67	2.52
15-30 minutes	2.53	4.61
30-45 minutes	2.05	3.01
45-60 minutes	2.05	4.04

Table 6

SDT Analysis
Low & High Stress

	Low Stress		High Stress	
Forecast	D(A)	β	D(A)	β
0-15 minutes	1.69	5.75	2.67	1.67
15-30 minutes	1.99	3.05	2.34	3.70
30-45 minutes	1.36	2.73	1.86	2.15
45-60 minutes	1.19	2.47	1.97	3.22

Appendix V

A Sample of Annotations from the *Display Bibliography*

(Please do not cite or quote)

Alpha-Numeric versus Graphic Displays Empirical Papers

Bemis, S. V., Leeds, J. L., & Winer, E. A.
(1988).
Operator performance as a function of type of display: Conventional versus perspective. Human Factors, 30, 163-169.

Domain: Hostile aircraft detection and selection of closest interceptor (closest friendly aircraft to the hostile craft)
Within Subjects: Male naval staff operational personnel

Procedure

Independent Variable: Display format of aircraft altitude information (conventional display in which subjects "hooked" a symbol to obtain the altitude information which was then presented numerically or perspective display in which vertical lines represented the altitude of the aircraft).

Dependent Variables: Response time to detect threats and select interceptors, number of false alarms and number of omissions for threat detection and interceptor selection.

Task: Thirty symbols (representing aircraft) were presented either conventionally or perspectively and the subjects' task was to detect hostile aircraft and to select the closest interceptor. "Instructions described the three modes of operation: pick [obtaining the altitude information], detect [detecting threats], and intercept [selecting an interceptor]" (p. 166). Subjects participated in a practice scenario followed by two 30-minute test scenarios, one for each of the two display formats.

Conclusions

"This experiment revealed a significant reduction in errors of detection and interception with the use of the perspective display. As expected, response time for selecting interceptors was greatly reduced in the perspective display condition. No difference existed in response time for threat detection between the two display conditions. The regression analyses showed that no correlation existed between (1) response time and errors for the combined tasks of detecting threats and selecting interceptors for both display conditions (perspective and conventional) and (2) response time and errors for each task (detect only and intercept only) for both display conditions" (pp. 168-169).

Benbasat, I., & Dexter, A. S.
(1985).
An experimental evaluation of graphical and color-enhanced information presentation. Management Science, 31, 1348-1364.

Domain: Managerial decision task allocating promotional budgets across three territories
Between Subjects: Undergraduate and graduate students in marketing courses

Procedure

Independent Variables: Display format of allocation reports (table or Cartesian line graph), display color of allocation reports (mono or multicolor), and individual differences (field-dependence or field-independence).

Dependent Variables: Profit performance, decision making time, and user perceptions (ratings of display attributes).

Task: The task in this experiment "involves the allocation of a fixed promotional budget across three territories with the objective of maximizing the resultant total profit" (p. 1349). Subjects were presented with a "history report that showed for each decision the promotional allocations made to each territory, profit by territory, and total profit. . . . the subjects proceeded to make their own allocations to the three territories in an attempt to maximize total profit over the ten decision periods" (pp. 1349-1350).

Conclusions

"In terms of the color effects, it was found that subjects with multi-color reports outperformed those with mono-color ones. Multi-color reports were also found [from user ratings] to be more understandable. . . . The direction of the results indicates that the decision quality and time performance improvement effected by color appear to be higher for graphical than for tabular reports, although no statistically significant interaction effects were observed between color and information system [display format]. . . . Color-coding especially improved the decision quality of field-dependent subjects. Field dependent subjects with mono-color reports had the worst profit performance. . . . As expected, no differences were found between tabular and graphical reports in terms of the quality of decision making [profit performance]" (pp. 1361-1362).

Benbasat, I., & Dexter, A. S.
(1986).
An investigation of the effectiveness of color and graphical information presentation under varying time constraints. MIS Quarterly, 10, 59-83.

Domain: Managerial decision task allocating a promotional budget across three territories
Between Subjects: 58 MBA students, five undergraduates and two doctoral students

Procedure

Independent Variables: Display format of allocation reports (tables, Cartesian line graphs, or a combined report containing both tables and graphs), display color of allocation reports (mono or multicolor), and time constraint within which to utilize the reports (low, 15 minutes, or high, 5 minutes).

Dependent Variables: Profit performance, time performance (amount of time taken during the time constraint) and user perceptions of display information (ratings of display attributes).

Task: The "task involves the allocation of a fixed promotional budget across three territories with the objective of maximizing the resultant total profit. . . . The subjects were shown, on a CRT screen [for either 5 minutes or 15 minutes], a report that represented profit at levels of promotion from $0 to the promotional budget limit for each territory. . . . The subjects were then asked to allocate the entire budget, or any portion thereof, to the three territories. . . . They were told that their performance would be evaluated based on the amount of profit they obtained from this one allocation decision" (pp. 60-61).

Conclusions

"The subjects in the tabular group [in the 15 minute time constraint] took 44% more time (15 vs. 10.5 minutes) than the graphical group subjects. . . . There were no time or performance differences associated with information presentation in the 5 minute group. . . . Given a reasonable amount of time to solve the problem (15 minute): 1. Combined group subjects had a higher profit performance than the graphical group subjects, but the same as the tabular group. 2. Combined and graphical group subjects completed the task faster [in the 15 minute condition] than the tabular group subjects. 3. Combined and graphical reports were preferred over tabular reports" (pp. 77-78). In addition, the authors found no significant differences for time or profit performance between the three display formats during the 5 minute time constraint. Color coding did have a significant effect on profit performance. Subjects using multicolor reports in the 5 minute time constraint performed significantly better than those subjects using monocolor reports.

Benbasat, I., Dexter, A. S., & Todd, P.
(1986).
The influence of color and graphical information presentation in a managerial decision simulation. Human-Computer Interaction, 2, 65-92.

Domain: Managerial decision task allocating promotional budgets across three territories
Between Subjects: Undergraduate and graduate students in marketing courses

Procedure

Independent Variables: Display format of allocation reports (tables or Cartesian line graphs) and display color of allocation reports (mono or multicolor).

Dependent Variables: Profit performance, number of reports requested, number of simulations taken before entering the optimal scheme and stopping, report use time (average time taken in trials where reports were requested), and ratings of report attributes.

Task: The "task involves the allocation of all or part of a promotional budget across three territories with the objective of maximizing the resultant total profit. . . . Subjects could simulate up to 30 different allocations with the objective of finding the optimal allocation of the promotional budget across the three territories. . . . After each allocation, the subject was automatically given feedback [which displayed subject's allocations and the total profit for that allocation]" (p. 70). At this point, subjects could proceed directly onward, they could request additional feedback and then proceed to another allocation, or they could enter the allocation scheme that they considered optimal and stop.

Conclusions

"As expected, color and information presentation differences did not influence profit performance. . . . Subjects with tabular reports requested more history reports over all trials than did the subjects with graphical reports. Color was not significant as a main effect. However, there was an interaction effect--monocolor graphical reports were requested the least. . . . Information presentation format significantly influenced the number of trials; as expected tabular group subjects simulated fewer allocation schemes. Color coding also reduced the number of trials needed to complete the task. . . . Subjects with multicolor reports spent significantly longer in report periods [report use time] than subjects with monocolor reports. . . . [No main effect due to display was found for report use time.] There were no significant differences in subjects' responses to the questionnaire items dealing with report attributes" (pp. 82-85).

Benbasat, I., & Schroeder, R. G.
(1977).
An experimental investigation of some MIS design variables.
MIS Quarterly, 1(1), 37-50.

Domain: Inventory/production management
Between Subjects: Students enrolled in operations management course

Procedure

Independent Variables: Form of report (table or graph), decision aid (available or not), exception reports (available or not), number of available reports (necessary or overload), decision making style (high analytic--formal planned analysis and low analytic--spontaneous trial and error with emphasis on feedback), and knowledge of functional area (low or high).

Dependent Variables: Cost performance, time performance, number of reports requested.

Task: "The experimental game was divided into ten decision points. At each decision point the subject requested information and . . . made three decisions: setting an order point, setting an order quantity, and setting the daily production figures for the next twenty periods. . . . The objective of each subject/manager when making the decisions was to minimize the total cost to the firm" (p. 42).

Conclusions

"Of the six main effects there were two which significantly affected cost, decision aids and graphical displays. Those subjects who had decision aids available had a mean cost that was 10 percent lower than the cost of those who did not have decision aids available. The graphical displays also reduced the average cost, but by only 6 percent. There was also an interaction effect between decision aids and form of presentation which indicated that subjects who had neither aids nor graphical reports had the poorest cost performance. . . . Although decision aids improved cost performance, subjects using them on the average took significantly longer (about 60%) to make their decisions. Observations of the subjects indicated that the extra time was due to deciding on input parameters for the models, analyzing model outputs, and combining the outputs of the aids with the information received from the historical reporting systems. . . . The number of reports requested by a subject was influenced by several variables and their interactions. Subjects having graphical reports used less reports than the ones having listed tabular reports. The extra reports used by the listed tabular group subjects came mainly from the 'overload' set of reports. Subjects with a high inventory knowledge used less reports than the ones with a low inventory knowledge. The extra reports used by the low inventory subjects again came from the 'overload' set of reports. The interaction between functional area knowledge and decision making style indicated that the low analytics with a low functional area knowledge used the most reports. . . . The high analytics who were using decision aids used less reports than the low analytics who were using decision aids. But for subjects with no decision aids the situation was reversed with the high analytics using more reports than low analytics. This finding would indicate the reliance of the high analytics on model usage, and that the models have more information value for them than the low analytics who have to rely more on the reporting system even when they have the aids. Subjects with the overload report set requested more reports than the subjects with the necessary report set, but did not outperform them. This finding would tend to indicate that the more reports that are made available, the more a subject requests. This result lends further evidence to the assertion that the subjects did not really know what reports were needed, but relied heavily instead on what was made available to them" (pp. 43, 46).

Carter, L. F.
(1948).
The relative effectiveness of presenting numerical data by the use of tables and graphs. Washington, DC: U.S. Department of Commerce.

Domain: Locating and interpolating values in tables or graphs with linear or curvilinear relationships between variables
Within Subjects: Military personnel

Procedure

Independent Variable: Display format (tables or Cartesian line graph), relationship between displayed variables (linear or curvilinear), and data set (one or family [multiple displays using different values for a constant]).

Dependent Variables: Number of problems completed (speed) and number of errors made (accuracy).

Task: "Four tables and four graphs representing the numerical relationships expressed by the equations y=cx and y=x2 were constructed. . . . These tables and graphs were designed to determine whether tabular presentation or graphical presentation is more effective when the user is required to: a. Enter the table on graph with tabulate arguments (" The term 'tabulated argument' is used to indicate that the value employed in entering the table is one of the printed values appearing in the first column of the table") (p. 1) . . . b. Enter . . . with one non-tabulated argument . . . c. Enter . . . with two non-tabulated arguments" (p. 1). In addition, the material was designed to determine if one set of data or a family of data are more effective. Subjects were required to answer a number of problems that involved using the presented information.

Conclusions

"When the values used as arguments are tabulated in the tables it is more efficient, in terms of both the speed and accuracy with which an individual can determine required information, to present both linear and curvilinear data, and single sets and families of data in tabular form rather than graphic form. . . . When the values used as arguments are non-tabulated values; i.e., when the required information must be obtained from the table by interpolation, it is more efficient in terms of the speed and just as efficient in terms of the accuracy with which an individual can determine required information, to present both linear and curvilinear, and single sets and families of data in graphic form rather than in tabular form" (p. 2).

Dickson, G. W., DeSanctis, G., & McBride, D. J.
(1986).
Understanding the effectiveness of computer graphics for decision support: A cumulative experimental approach. Communications of the ACM, 29, 40-47.

Experiment One
Domain: Bank loan decision making
Between Subjects: Undergraduate business students

Procedure

Independent Variable: Display format of loan information (tables or bar graphs).

Dependent Variables: Interpretation accuracy ("scores for items requiring the subject to identify values, compare values, or observe trends were summed") (p. 42), decision quality ("items related to loan qualification, loan amount, and loan riskiness were summed") (p. 42), task difficulty (as rated by subjects), and report readability (as rated by subjects).

Task: "The subjects were presented with a short case which described the situation of a small business in need of a loan. Subjects were told they were to play the role of a bank loan officer. They were asked to read financial statements of the small business . . . determine if the firm qualified for a loan, determine the maximum amount of the loan, and rate the riskiness of the loan" (p. 41). The subjects were asked to use a step-by-step procedure to determine loan qualification, loan amount, and riskiness. Finally, the authors' describe the task as a familiar task to the subjects (low task complexity) with a relatively easy task structure (the step-by-step procedure given to the subjects in order to arrive at a decision).

Conclusions

"Interpretation accuracy and decision quality scores were not significantly different for the two [display] groups. Subjects receiving graphical reports tended to rate the task as more difficult, but differences in ratings between the two treatment groups were not significant. With regard to report readability, the graphical group found the reports to be more difficult to read than the tabular group, and this difference was significant" (p. 42).

Experiment Two
Domain: Forecasting product demand
Between Subjects: Undergraduate business students

Procedure

Independent Variables: Display format of product demand information (tables or "line plots"). Note: The authors do not illustrate or define "line plot".

Dependent Variables: Interpretation accuracy (sum of interpretive question scores), decision quality ("forecast accuracy in each of three periods for each of three products") (p. 43) and difficulty rating of the task.

Task: "Subjects were first presented with a short case describing a chemical manufacturer in need of assistance in forecasting demand for three of its products. The subjects were then given demand histories for each of the three products. . . . After reading the three reports, the subjects were asked five interpretive questions. . . . and then asked to provide specific estimates of demand for each of the three products for three months into the future [nine total forecasts were made]" (p. 42). This task though still somewhat familiar to subjects, was less familiar than the task in the first experiment (a higher level of task complexity than the task in the first experiment) and it had a more difficult task structure (no specific guidelines were given to subjects in order to help them arrive at a decision) than the task in the first experiment.

Conclusions

"The average interpretation accuracy scores for both the tabular and graphic groups were reasonably low (lower scores indicate a better performance). Scores for the graphic group were lower . . . than those for the tabular group. . . . However, the difference between interpretation scores for the two groups was not significant. . . . [For decision quality] in eight of the nine forecasts, subjects presented with graphs [significantly] outperformed those working with tables. . . . Consistent with their better performance, the graphical group perceived the task to be [significantly] easier than the tabular group" (pp. 42-43).

Experiment Three
Domain: Understanding of a business graphics report
Between Subjects: Undergraduate business students

Procedure

Independent Variables: Display format of the report (tabular or graphical), subset (complete presentation in which all information is presented at one time or subset presentation in which subjects were given half of the information and those questions relating to that half and then given the second half of the information and those questions relating to the second half), and visual ("recall" in which subjects looked at the reports for a time period and then answered the questions without access to the reports or "lookup" in which subjects had access to the reports when answering the questions). Note: The authors do not illustrate or define the type of graphical presentation used in this experiment.

Dependent Variables: "Getting the message" measure ("the number of correct and incorrect statements appropriately identified in a list of 32 statements") (p. 44) and "traditional message" measure, ("the total number of correct answers to a set of 17 questions") (p. 44).

Task: "Subjects were first presented with a short case describing a producer of computer graphics software that had contracted with a research organization to do a survey of users of computer graphics. The subjects were told that they were involved in an experiment to evaluate the quality of the research firm's final report. They were told that they would receive a report on current usage of graphics technology in business and would be asked a series of questions about what the report was trying to convey" (p. 43). This task was an unfamiliar task to the subjects (high level of task complexity) and it had a difficult task structure (no guidelines were given to subjects in order to help them arrive at a decision).

Conclusions

"First, on the traditional measure, the lookup groups did very well (scoring almost perfectly), while the recall groups did poorly [a significant difference]. On the other hand, there was little difference in performance on the [getting the] message measure. About the only clear pattern to emerge is that the groups getting the information in 'chunks' (the subset groups) nearly always 'got the message' better than their counterparts working with the entire set of reports [a significant difference]. . . . None of the two-way interaction effects are significant [for getting the message measure]. [For the traditional message measure] there are two significant two-way interactions, format by subset and visual by subset. These results suggest that it made a difference to the recall group (but not the lookup group) whether or not they had the complete set or two subsets of reports. The recall group performed better with subsets. Similarly, performance on the traditional message measure showed a relationship between the format employed and whether or not the subjects had a complete set of reports or a subset. In the case of having the complete set of reports (and many questions to answer), the graphical group did significantly better than the tabular group. However, in the case of those receiving the material in two parts, the opposite result occurred" (pp. 44-45). No sigificant differences were found between the two display formats for either the "getting the message" measure or the "traditional" measure.

Overall Conclusions

"This program of cumulative experiments indicates that generalized claims of superiority of graphic presentation are unsupported, at least for decision-related activities. In fact, the experiments suggest that the effectiveness of the data display format is largely a function of the characteristics of the task at hand" (p. 40).

Feliciano, G. D., Powers, R. D., & Kearl, B. E.
(1963).
The presentation of statistical information. Audio Visual Communication Review, 11(3), 32-39.

Domain: Agricultural information
Within Subjects: High school students, agricultural college students, homemakers

Procedure

Independent Variable: Form of information (1. long detailed table [including information not required for test questions], 2. short, simple table, 3. horizontal bar graph, 4. four- to six-paragraph text, 5. numbers 1 and 4 combined, 6. numbers 2 and 4 combined, 7. numbers 3 and 4 combined.

Dependent Variable: Scores from seven questions requiring interpretation of presented information.

Task: Four experiments were conducted wherein different subjects, form of information presentation, and order of presentation were used. Subjects were presented with information about agriculture and answered interpretive questions.

Conclusions

"1. The horizontal bar graphs that were used consistently produced significantly better scores than did the long tables, short tables, or text by itself. 2. Short tables resulted in better scores than did long tables when the test groups were homemakers clubs made up of women who had had little or no recent formal training and/or experience with statistical presentation methods. No significant differences were obtained between these two methods in test groups of high school and college students whose training or experience with statistical tables was more recent. 3. Both the short tables and the long tables resulted in significantly better scores than did textual presentations by themselves. 4. Using horizontal grouped bar graphs to reinforce text gave significantly higher scores than did the use of short tables or long tables for this purpose. 5. No significant difference in scores was obtained when text was reinforced by short tables as against long tables. Both kinds of reinforcement were more effective than text alone. 6. Text reinforced with horizontal grouped bar graphs and text reinforced with short tables were both significantly better than the horizontal grouped bar graphs by themselves. 7. Horizontal grouped bar graphs, even without textual reinforcements, resulted in better scores than the long tables with textual reinforcement. 8. Short tables with textual reinforcement gave better scores than the short tables without textual reinforcement" (p. 37).

Garceau, S., Oral, M., & Rahn, R. J.
(1988).
The influence of data-presentation mode on strategic decision-making performance. Computers and Operations Research, 15(5), 479-488.

Domain: Determining necessary actions to place a given company in a better competitive position.
Mixed: Managers and specialists in economic planning in Quebec civil service.

Procedure

Independent Variables: Presentation mode (table or bar chart), decision-process phase (intelligence--finding relevant data, and design--establishing relationships among the data), and cognitive style (logical or intuitive).

Dependent Variables: Quality of solution, time required to reach solution, confidence.

Task: "participants completed a questionnaire [about academic training, age, etc. and assessed cognitive style], read a document presenting the case and tried to solve a given problem using a micro-computer-based system giving direct access to the data of the problem" (p. 482).

Conclusions

"The results of the experiment can be summarized briefly as: tabular presentation leads to better solutions, faster, in the Intelligence phase; no significant difference between the two modes was evident in the Design phase; when cognitive style (rational vs intuitive) was taken into account, the effect of the mode of presentation was significant for the whole set of observed variables in both phases" (p. 480). "The logical style participants using tables produced very good results in a relatively long time but with a high level of confidence in their solutions. The intuitive style participants using graphics produced equally good solutions but in a shorter time than the logical-tabular group. As well they had a lower level of confidence in their solutions" (p. 487).

Kerkar, S. P., & Howell, W. C.
(1984).
The effect of information display format on multiple-cue judgment.
(Tech. Rep. No. 84-2). Houston: Rice University, Department of Psychology.

Experiment One

Domain: Evaluation of applicant profiles for a secretarial job
Within Subjects: Undergraduate students

Procedure

Independent Variables: Display format (numerical or bar graph) and decision task (rating task or choice task). (NOTE: Both variables were counterbalanced producing eight experimental conditions; also, statistical tests were conducted only on display format for each decision task separately; in the analyses for the rating task, cue was treated as an independent variable; in the analyses for the choice task, consistency of policy (consistent [actual numerical choices evaluated with reference to a numerical policy] and inconsistent [actual graphical choices evaluated against numerical policy]) was treated as an independent variable.

Dependent Variable: Utilizing subjects' ratings and choices in separate regression analyses, weights for each cue and consistency measures were derived for each subject as dependent variables for judgment task, "accuracy" scores which compared predicted choices from a subjects' policy to actual choices in the choice task.

Task: "The basic tasks required subjects either to rate the suitability of applicants for the job of secretary [rating task] or to decide whether they should be hired [choice task]. . . . More specifically, subjects were presented with profiles of information about hypothetical applicants which were comprised of four [uncorrelated] dimensions: intelligence, motivation, skill, and experience. Each profile was represented in one of two ways: as a set of numerical scores (numerical format) or as a set of bar graphs (graphic format). (p. 7).

Conclusions

With respect to the rating task, "for every subject, a separate policy equation was obtained for the numerical and graphic displays by regressing each type of judgment on the four cues. . . . The main effect of format was not significant . . . suggesting that the average weights for all cues combined were comparable for the two formats. Obviously this is less meaningful than the cue x format interaction (which compares weighting policies for the formats); this interaction was highly significant. . . . The main effect of cues was also significant. . . . Clearly, therefore, subjects weighted the four cues differently [in the two formats]. . . . the specific weights attached to each cue were more uniform in the graphic than in the numerical display" (pp. 10-11). No differences due to format were found for the traditional measure of consistency, R2. However, the authors partitioned R2 into the sum of squares for the predictions and sum of squares for the errors (which were transformed) and found the mean error sum of squares were not different while the mean error differences were statistically different. "What this finding suggests is that the graphic format produced considerably more precision in judgment than did the numerical format, a conclusion that is reinforced by the fact that variability in raw criterion judgments was also significantly lower for the graphic displays. . . . Since there was no external criterion available to define choice accuracy, the subjects' own numerical and graphic rating policies were used as criteria. That is, 'policy captured' weights were applied to the cue values for each pair of choice profiles to determine which profile should be chosen if the individual was consistent with his/her own policy. These predicted choices were then compared to actual choices under the two formats to obtain 'accuracy' measures. Since there were two policies (numerical and graphic) for each set of values, it was also possible to compare decision 'accuracy' for consistent criteria (e.g., actual numerical choices evaluated with reference to a numerical policy) with those for inconsistent criteria (e.g., actual graphic choices evaluated against a numerical policy). . . . Neither the effect of format nor the interaction between format x consistency of policy was significant. . . . This suggests that despite the differences in subjects' rating policies under the two formats, they predicted choices with similar levels of accuracy. There was, however, a significant effect of consistency. . . . Although the absolute differences were extremely small, a consistent policy predicted slightly better than an inconsistent one. This implies that subjects' rating and choice behavior were more similar when information was displayed in identical than in different formats. Thus, while numerical and graphic cues were processed differently, the same display mode induced similar kinds of processing for both rating and choice tasks" (pp. 13-14).

Experiment Two

Domain: Evaluation of applicant profiles for a secretarial job
Within Subjects: Undergraduate students

Procedure

Independent Variable: Display format (numerical or bar graph). NOTE: As above, cue was treated as an independent variable in the analysis.

Dependent Variables: Cue weights and consistency from regression analysis.

Task: The task in this experiment was the same as that in the first experiment, except that the choice task was eliminated.

Conclusions

"These results replicate the primary finding of Experiment 1--format again produced a differential weighting of cues. . . . numerical policies were less consistent than graphic ones. . . . lower consistency of numerical judgments resulted largely from greater error in these judgments than in graphic ones" (p. 18-19).

Experiment Three

Domain: Judging teaching effectiveness
Mixed: Undergraduate students

Procedure

Independent Variables: Display format (numerical or bar graph), number of cues presented (six cues or four cues), and order of the cue presentation.

Dependent Variables: Same as in the second experiment.

Task: Subjects participated in a rating task of job applicants for a position as an instructor. Subjects "were presented [with the cues] successively, at a 2 second rate. . . . After all four or six cues were presented. . . . the subject proceeded to write down his/her rating" (pp. 22-23). This procedure was completed until completion of 200 profiles. Subjects had a brief rest period between numerical and graph formats.

Conclusions

"Looking first at the four-cue ANOVA, the means [of the weights] for the numerical and graphic formats were .64 and .52 respectively, a difference that was significant. . . . Thus, cues tended to be weighted more heavily on average under the numerical than the graphic display. However, as predicted, the cue [order] x format interaction did not approach significance. . . . The six-cue ANOVA also failed to reveal a significant cue [order] x format interaction. . . . However, the main effect of format found in the four-cue condition was absent here. . . . In sum, there was no evidence for a differential weighting of cues presented sequentially under the two formats--the cue x format interaction found consistently in the first two experiments was eliminated in this one. This supports our hypothesis of holistic processing of graphic cues. However, there were some processing differences as a function of format; the numerical format produced larger overall cue weights than the graphic format in the four-cue condition" (pp. 24-25).

Overall Conclusions

"The most important finding was that subjects weighted the same cues differently when displayed numerically than they did when displayed in graphic form. That is, their judgments and choices suggested that they attached consistently more (or less) importance to particular items of information under one format than under the other. . . . These differences disappeared, however, under conditions of sequential cue presentation (Experiment 3), a situation designed to minimize the holistic processing tendency believed to occur with the graphic format" (pp. 26-27).

Lucas, H. C.
(1981).
An experimental investigation of the use of computer-based graphics in decision making. Management Science, 27, 757-768.

Domain: Product demand (importing cases of whiskey).
Between Subjects: Participants in a graduate school summer business executive program who held middle or upper management positions.

Procedure

Independent Variables: Five treatment groups were used: "1. Tabular output of data on a hard copy terminal . . . 2. Tabular output on a CRT . . . 3. Graphical output only of the probability distribution . . . 4. Graphical and tabular output of the probability distribution . . . 5. Graphical output of simulation costs data . . . 6. Graphical and tabular simulation of cost data" (pp. 760-761). In addition, decision making style was assessed and subjects were classified as heuristic or analytic decision makers.

Dependent Variables: Two measures of performance: "The 'best simulation' variable is the lowest cost of the eight simulation runs using the five years of historical data. The 'next year simulation' variable is the average of running data from the same distribution ten times for one year in the future using the subject's final order quantities" (p. 761); measures of understanding of inventory theory and probability, ratings of display usefulness, and enjoyment.

Task: "The firm in the simulation imports whiskey to the U.S. To simplify the problem, the company has to place a yearly order in December for four deliveries which are made at the beginning of each quarter. . . . the subjects had the following data available on the computer. 1. A display of the last five years of demand by quarter. 2. A display of frequency and cumulative relative frequency of demand by quarter for the last ten years. 3. A simulation capability allowing ordering decisions to be tested using the last five years of data (subjects were able to exercise this option up to eight times during the experiment). 4. The ability to test the final ordering decision on data for the following year drawn from the same distribution as the historical data" (p. 760). Subjects were asked questions about the data they had received and the usefulness of different displays.

Conclusions

CRT Tabular versus Hard Copy Output

"The group using a hard copy terminal performed significantly better than the group using the CRT without graphical output. However, the CRT group found the data from the simulation output more useful. The hard copy group had a higher test score on questions relating to inventory theory as well. Comparing the hard copy group to all other subjects who used the CRT shows that the hard copy group had superior performance and better test scores on inventory understanding" (p. 763).

Graphical versus Tabular (both on CRT)

"The CRT tabular group found the simulation output more useful than the two graphics treatments for demand frequencies and had significantly lower scores on the test of inventory understanding. For the two groups receiving graphs of the simulation output, one group had a higher score on each of the tests [CRT graphics only group and CRT graphics and tabular group both outperformed the CRT tabular only group] and one graphics group [CRT graphics only] reported significantly more enjoyment from the exercise than the tabular CRT group" (p. 763).

Graphics and Tabular versus Tabular Only (both on CRT)

"The only significant differences among the treatments are for the usefulness of demand frequency distributions and simulation output graphs. For the group receiving the graph of probabilities and tables of probabilities as well, the scores were higher on the inventory test than for the group that received the graph alone. For the graph of the simulation results the group receiving both graphics and tabular information reported greater usefulness for the simulation output compared to the group that received graphics alone" (pp. 764-765).

Heuristic versus Analytic Decisionmakers

"No differences between tabular presentation on the CRT and graphics were found without controlling for decision style. . . . the best simulation value was better under graphics and that most of superiority was in the heuristic group, that is, heuristic decisionmakers in the graphics treatment group had the best (lowest cost) simulation results. There is almost no difference between treatments for the analytics. . . . It is interesting to note that the results for the test of inventory understanding exhibit significant interaction, but in the other direction from performance. Taken as a whole, the graphics groups have had the highest scores. However, here the differences are greatest for analytics. . . . For reported usefulness of the frequency distributions and simulation output, heuristics scored highest, though the results are statistically significant only for the first of these variables" (pp. 765-766).

Overall Conclusions

"The results of the experiment provide limited support for the use of graphics presentation in an information system. Decision or cognitive style also appears to be an important variable influencing the performance of an individual and the reaction to an information system" (p. 757).

Lucas, H. C., & Nielsen, N. R.
(1980).
The impact of the mode of information presentation on learning and performance. Management Science, 26, 982-993.

Domain: Logistics management game in which products were to be shipped via various means.
Between Subjects: MBA students, industrial engineers (IEs), and senior executives (SEPs).

Procedure

Independent Variables: Cognitive style (analytic or heuristic), treatment groups used during two time periods. (1. MBA, teletype hard copy terminal (TTY) and four basic reports at Time 1, TTY and four basic reports at Time 2; 2. MBA, TTY and four basic reports at Time 1, TTY and all reports at Time 2; 3. MBA, CRT and four basic reports at Time 1, CRT graphics at Time 2; 4. IE, TTY and four basic reports at Time 1, TTY and all reports at Time 2; 5. IE, CRT and four basic reports at Time 1, CRT graphics at Time 2; 6. SEP, TTY and four basic reports at Time 1, TTY and all reports at Time 2; 7. SEP, CRT and four basic reports at Time 1, CRT and all reports at Time 2; and 8. SEP, CRT and four basic reports at Time 1, CRT graphics at Time 2.)

Dependent Variables: Profit, rate of profit increase, learning "measured by the difference in the slope or rate of profit increase between the two sessions" (p. 987).

Task: "All firms (players) have a constant-cost production facility located on the West Coast of the country and compete in a constant-price market on the East Coast. A variety of different modes of transport are available to move the product" (p. 984). Thus, the players must make decisions about which of the various modes of shipping to use and the volumes of product to be shipped via the chosen modes. Subjects participated in two sessions with 24 decision making periods per session.

Conclusions

"Hypothesis 1 (additional information will result in greater learning and better performance) is not strongly supported by the experimental results. . . . Grouping the data . . . by experimental treatment showed no significant differences. . . . Hypothesis 2 (superiority of CRT terminals) receives some support from the data. The combination of all groups using CRT's had higher first and second-session profits and had a higher rate of profit increase during the second session than the combination of all groups using teletypes. . . . Hypothesis 3 (graphics presentations will result in greater learning and performance) received very little support from the experiment. When the data . . . are grouped by experimental treatment and when all graphics treatments are combined, there is no clear superiority for graphics. However, . . . the superiority of graphics is statistically significant for the IEs for first session profits and for rate-of-profit increase and for executives for second session profits and for rate of profit increase. . . . Hypothesis 4 (influence of background variables) is supported for 'learning'. For each comparison, there is less learning for MBAs than for the other two groups, while IEs exhibit the most learning between the two sessions. All three groups differ on work and military experience, and cognitive style" (pp. 989-990).

Lusk, E. J., & Kersnick, M.
(1979).
The effect of cognitive style and report format on task performance: The MIS design consequences. *Management Science, 25*(8), 787-798.

Domain: Annual income of different professionals
Between Subjects: Undergraduate students

Procedure

Independent Variables: Report format (A: raw data in tabular form, B: tabular percentages, C: raw data in a frequency histogram, D: raw data accumulative frequencies in tabular form, E: percentages in cumulative frequency graph) and cognitive style (low or high analytic).

Dependent Variables: Performance and complexity ratings

Task: Subjects were given one of the five reports and asked to answer twenty questions

Conclusions

"The results of the experiment were (1) the perceived complexity rankings (lowest to highest) for both the high and low analytics were Report A, Report B, Report C, Report D, and Report E, (2) individuals classified as high analytic outperformed the individuals classified as low analytic on each of the five reports, and (3) for both the high and low analytics task performance decreased as perceived complexity increased" (p. 787).

Remus, W.
(1984).
An empirical investigation of the impact of graphical and tabular data presentation on decision making. *Management Science*, *30*, 533-542.

Domain: Production scheduling
Between Subjects: Undergraduate business students

Procedure

Independent Variables: Display format of production scheduling information (Cartesian graph or table).

Dependent Variables: Four performance measures: "First, as the subjects made decisions the actual costs were calculated. . . . using the paint plant's quadratic cost function. Second . . . the individual decision makers were modeled with regression rules. . . . the resulting decision rules were then used to make the workforce and production decisions. . . . Third . . . these betas [estimated from the above regression analysis] were averaged to form composite rules [collective policy of subjects] for each treatment. . . . Lastly, the production and workforce decisions were calculated using the Holt, Modigliani, and Muth optimal rules; the costs were again found using the quadratic cost function" (p. 537).

Task: "The subjects were first given a presentation on production scheduling including several examples of how to use the graphical and tabular aids to make better decisions. . . . [Before making their first decision] subjects received first the next three period's sales forecasts. Based on these forecasts, the inventory position, current workforce size and worker productivity, the subjects scheduled the production volume and decided how many workers to employ. . . . The subjects then received the actual sales costs, the new inventory level, and the average cost thus far. This cycle was repeated for each of the 24 periods" (pp. 535-536).

Conclusions

"In the first 12 periods (the learning phase) there was no significant advantage for either type of display [regarding actual costs]. This was also true in the last 12 periods (the stable decision making phase). When the regression rules were used, the resulting costs provide a comparison of production schedulers who are consistent in applying their managerial judgment. . . . neither the tabular nor the graphical displays resulted in lower costs. The regression rule costs however, where [sic] significantly lower . . . than the actual costs. . . . The results from the composite regression rule analysis . . . were significantly lower . . . than the regression rule or actual costs. In both the learning and the stable decision making phases, the tabular display costs were significantly lower than the graphical display costs. . . . In both the learning and stable decision making phases, the optimal costs were significantly lower [than the actual, regression, and composite rule costs]. . . . Tabular displays generally yielded costs which were lower, but not significantly lower, than graphical displays. . . . These results do not unequivocally point to the better decision aid for the individual decision maker since the erratic components of decision making disguise the benefits of the tabular aids. Thus individual decision makers may choose either aid but should focus their attention on consistent decision making. Only when that occurs can the benefits of tabular aids be significant" (pp. 538-540).

Schwartz, D. R., & Howell, W. C.
(1985).
Stopping performance under graphic and numeric CRT formatting. Human Factors, 27, 433-444.

Experiment One

Domain: Simulated hurricane-tracking scenarios
Mixed: Undergraduate students

Procedure

Independent Variables: Display format (latitude by longitude grid or a table containing latitude and longitude information), decision aid (providing subjects the probability that the hurricane would hit the city or no decision aid), and trial block (one, two, or three).

Dependent Variables: Information sampling (point at which the terminal decision was made), decision accuracy ("with reference to a normative model that may or may not be a good description of the subject's intuitive model") (p. 438), and latency (response time).

Task: "Each subject [for one of the two display formats and for one of the two decision aid conditions] was required to monitor a series of simulated hurricanes in their advance toward a heavily populated target area. . . . subjects were required to make one of three responses after each advance: stay (i.e., minimize the potential losses associated with decision delay as in committing to intensified protective measures rather than retreat), evacuate (i.e., order total abandonment of the city), or wait (i.e., postpone any terminal action). . . . the subject's task was to decide when to stop gathering information [terminal decision] and, at that point, which action to take based upon (1) the storm's current location, and (2) the subjective expectation of costs associated with the various options in that situation" (p. 435).

Conclusions

"No significant main effects or interactions attributable to the format difference were obtained on any of the measures. . . . Interestingly, sampling tended to increase significantly rather than decrease over trial blocks, whereas performance accuracy and latency both improved significantly. . . . the presence of a decision aid [significantly] raised the mean accuracy of all decisions from 75.25% to 79.50%. . . . Of course these mean values also reflect the disproportionate influence of early, easy wait decisions; by contrast, accuracy for the critical last four positions averaged 55.00% and 62.70% for unaided and aided conditions repsectively. . . . Although mean latency for the aided decisions in these [last four] positions averaged 122 ms longer than for the unaided ones, and the difference was principally vested in the numeric display mode, neither effect was statistically significant" (pp. 438-439).

Experiment Two

Domain: Simulated hurricane-tracking scenarios
Mixed: Undergraduate students

Procedure

Independent Variables: Display format (same as in the first experiment), time stress (rate at which update information is presented, [300, 700, or 1630]), and trial blocks (one, two, and three). (Note: an additional independent variable was the counterbalancing of display format).

Dependent Variables: Information sampling (point when the terminal decision was made), decision accuracy, and latency (response time).

Task: The task in this experiment was identical to the task in the first experiment with the exception of the addition of the time stress variable, the elimination of the decision aid and the addition of trials.

Conclusions

"Looking first at the accuracy measures, it is apparent that the graphic format produced consistently superior performance, particularly under the more stressful pacing conditions and in the critical later stages of each storm. . . . In the analysis of overall accuracy . . . where storm position (early versus late) was included as a separate variable, it is clear that the effect was limited primarily to the later decisions. . . . the graphic superiority was limited to the more stressful conditions. . . . the overall tendency was to oversample . . . nearly a fourfold increase over that for Experiment 1. . . . Thus, the mere existence of time pressure for making individual decisions, and the consequent reduction in opportunity to process the available information, appears to have caused subjects to seek more information. . . . the numerical format promotes an increase in sampling over trial blocks, but that this format effect is a relatively small modulation in a very large tendency for time stress to promote oversampling. Since overall accuracy was about 7% lower than for the comparable condition in Experiment 1, this dramatic increase in oversampling does not appear to have been very productive. . . . the time stress manipulation produced a systematic decrease in accuracy" (pp. 440-442).

Overall Conclusions

"Display formation had a significant effect when time pressure was involved: subjects reached earlier and better terminal [stay or evacuate] decision under the graphic than the numerical format (Experiment 2). The differences reduced to nonsignificance under self-pacing (Experiment 1), although significant improvements were obtained by use of a simple aiding device (calculation of worst-case probabilities). Results are generally consistent with Hammond's cognitive consistency [sic] theory" (p. 433).

Stock, D., & Watson, C. J.
(1984).
Human judgment accuracy, multidimensional graphics, and humans versus models. Journal of Accounting Research, 22, 192-206.

Domain: Financial setting involving judgments of corporate bond rating changes
Between Subjects: Accounting undergraduate and graduate students and faculty

Procedure

Independent Variables: Presentation of financial information (face display in which features corresponded to information or tables) and expertise (elementary accounting class, intermediate I accounting class, intermediate II accounting class, or expert subjects who were accounting doctoral candidates or faculty members).

Dependent Variables: Classification accuracy

Task: All groups of subjects were presented with "a sequence of six [schematic] faces or a table of financial ratios for each firm for the six years 1969 through 1974" (p. 196). Subjects were asked to detect a change in the bond rating and to classify the type of change into one of three change categories: downgraded, upgraded, or no change. Forty-two firms in all were used; however, the schematic faces groups only received 21 of the firms due to the time limit. In addition, the expert group also were presented with the multidimensional graphic displays plus estimates from a decision model.

Conclusions

"Participants given information in the form of multidimensional faces were able to classify bonds into rating change categories more accurately than those given tables of financial ratios. Furthermore, this relationship carried over from individuals with minimal accounting training to those having higher levels of training" (p. 201).

Tullis, T. S.
(1981).
An evaluation of alphanumeric, graphic, and color information displays. Human Factors, 23, 541-550.

Domain: Telephone line testing system
Within Subjects: Experienced Bell System employees

Procedure

Independent Variables: Display format (narrative which used words and phases structured table, black-and-white graphic which included a schematic of the telephone system, and color graphic), session (one which included training, exercises, testing on each of the four formats and two which included testing on the four formats) and presentation sequence (using a Latin-square design).

Dependent Variables: Accuracy, response time, number of exercises in the training (first) session to achieve 80% accuracy, and preference.

Task: "The framework for the present study was a computer-based system for the telecommunications industry. This system, entitled Mechanized Loop Testing, provides a means for diagnosing problems on telephone lines. Upon request from a person seated at a CRT, the system accesses the telephone line, measures a wide variety of electrical characteristics, performs some interpretation of those characteristics, and displays the results on the CRT. The person can then make a decision about the nature of the problem on the line and the action needed to correct it" (p. 542). After training, subjects were presented with two sets of 37 displays of test results for each of the four formats. They were asked questions "which ranged from simple identification to complex integration and decision making" (p. 543).

Conclusions

"The most pervasive finding is the superiority of the two graphic formats over the narrative format: response times were significantly shorter, fewer training exercises were required to achieve the accuracy criterion, and subjective ratings of overall quality were significantly better. The only results that do not reflect this superiority are the accuracy data, which, as explained earlier, indicate a ceiling effect. Another consistent finding is the lack of a significant difference between color and black-and-white graphic formats. Subjects' response times to these formats did not significantly differ in either session [regarding response times, number of training exercises and subjective assessments]" (pp. 547-548).

Vicino, F. L., & Ringel, S.
(1966).
Decision making with updated graphic vs. alpha-numeric information.
(Tech. Research Note 178). Washington, D.C.: Army Personnel Research
Office.

Domain: Battlefield scenario
Between Subjects: Military personnel

Procedure

Independent Variables: Display format of battlefield information (tables or graphs in which the background was a line map) and rate of updating information (every slide update yielding 14 slides total or every second slide update yielding 7 slides total).

Dependent Variables: Decision score (points were given based upon when the final decision was made and whether it was correct or incorrect; earlier final decisions resulted in more points being awarded provided that the decision was correct; incorrect decisions were not awarded any points), decision speed, and confidence.

Task: "A series of slides depicting three enemy sectors was presented to 37 subjects. Successive slides showed the enemy forces in one of the three sectors forming for attack at a faster rate and with more appropriate disposition of forces than the forces in the other two sectors. After each slide was presented, each subject was asked to make a decision as to which of the enemy forces was preparing to attack and to indicate how confident he was about the decision. At each stage, the subject had the option of declaring his decision final" (p. ii).

Conclusions

"No differences in quality or timeliness of decision or in confidence that a decision was correct were found between alpha-numeric and graphic presentation. No differences were found in results with the two rates of updating. Subjects showed greater shifts in level of confidence from slide to slide in the 7-slide updating than in the 14-slide updating. This difference held for both modes of presentation. On the average, subjects whose final decision was correct had made the correct response approximately three-fourths of the way to their final decision" (p. ii).

Wainer, H., & Reiser, M.
(1976).
Assessing the efficacy of visual displays. Proceedings of the American Statistical Association, Social Statistical Section, 1, Part 1, 89-92.

Domain: Interpretation of crime statistics
Within Subjects: Undergraduate students

Procedure

Independent Variables: Display format (tabular, bar chart, Cartesian rectangular display which has an X axis with white victim on the left side and black victim on the right and a Y axis with white offender above the axis and black offender below the axis, or a "floating four-fold contingency display") and occasion (first trial and second trial).

Dependent Variables: Response time and response speed (1/time). Subjects also ordered the displays from easiest to most difficult to use.

Task: "An assertive statement was presented to the subject, and then followed by one of the displays. Each statement took the form: 'In the crime of armed robbery (rape, aggravated assault), white (black) criminals victimized whites (blacks) more often than they victimized blacks (whites)'. The subject's task was to decide whether the statement was true or false, based on the information in the display" (p. 90).

Conclusions

Due to outliers the mid-means were used in the analysis of response time "which still shows the Cartesian rectangles as the display of choice . . . followed closely by the bar charts. Next we have the FCD [floating four-fold contingency display], while the table of numbers brings up the rear. Although the order of displays for the second trial is not the same as in the first trial, the display occasion effect was not significant. . . . The FCD was judged hardest to use, but this judgment may reflect unfamiliarity more than difficulty of use. . . . Although the effect of display is significant when reaction time is the dependent variable [see above for order], unfortunately, it is not significant when the dependent variable is transformed to speed" (p. 91).

Watson, C. J., & Driver, R. W.
(1983).
The influence of computer graphics on the recall of information.
MIS Quarterly, 7(1), 45-53.

Domain: Number of University of Utah M.D. graduates residing in different U.S. locations
Between Subjects: Business students

Procedure

Independent Variable: Mode of information presentation (3-dimensional map graph or table)

Dependent Variable: Immediate and delayed recall

Task: Subjects were presented with information regarding M.D. graduates' residence in either tabular or graphical form, "instructed to study the information for one minute, after which they would be asked to respond to questions about the information. At the end of the study period the stimuli were taken from the subjects. A list of the top six states, in terms of the percentage of physicians located in the state, was then presented to the subjects as a recall cue. The list was not rank ordered. Subjects were then asked to rank the six states in order, from highest relative frequency to lowest, based on the information they had previously studied. No mention was made that another session would be held at a later date" (p. 49). Subjects were asked to rank order the same list at a second session four weeks later.

Conclusions

Rank order correlations between each subject's ranks and actual ranks were computed. Tests of differences between the mean correlations for the two presentation modes were not significant for either session. "This investigation does not lend support to the notion that computer plots of three dimensional graphics as a mode for conveying information in IS, will result in increased recall, both immediate and delayed, of that information, when compared to the more traditional tabular mode of presentation. Computer plots of three dimensional maps were not superior to the tabular presentation of data with respect to the degree of recall of information" (p. 51).

Wickens, C. D., & Scott, B. D.
(1983).
A comparison of verbal and graphical information presentation in a complex information integration decision task. (Tech. Rep. No. EPL-83-1/ONR-83-1). Urbana-Champaign: University of Illinois, Engineering-Psychology Research Laboratory.

Experiment One

Domain: Battlefield scenario
Within Subjects: Undergraduate students

Procedure

Independent Variables: Display format (verbal data in which information values are listed or spatial-graphical presentation), problem size (six, eight, or ten cues), trial variability (low variability, problems in which diagnosticity [the relevance of a piece of information] and reliability [the credibility of the source of the information] are postively correlated or high variability [cues with high diagnosticity and low reliability or low diagnosticity and high reliability]), presentation time (fast, cues presented for three seconds or slow, cues presented for five seconds), and weighted difference (5-10%, 15-20%, or 25-30%), which "was computed by dividing the absolute difference of support presented for the two different hypotheses by the total support presented for both hypotheses" (p. 17).

Dependent Variables: Accuracy and confidence

Task: "The subject was designated as a commander, responsible for defending an area through which an attack from a fictitious threat force was eminent [sic]. It was the subject's duty to analyze the available intelligence [using an adding-mutiplying model] and decide which avenue of approach, north or south, the threat force would take. . . . The information for each hypothesis in each problem was presented sequentially from several sources of information or cues. Each source conveyed information for one of the two possible hypotheses. The worth of each source was determined by two dimensions [reliability and diagnosticity]. . . . Subjects were instructed to evaluate the information presented and decide which hypothesis concerning future threat force actions was most likely to occur" (p. 8).

Conclusions

"An analysis of variance . . . revealed significant main effects on decision accuracy for three [display format, trial variability, and presentation time] of the five variables studied. No interaction effects were found to be statistically significant. . . . As predicted, the spatial code format yielded an [significant] improvement in decision accuracy over the verbal format. The effect of trial variability on decision accuracy was also statistically significant. . . . Decision accuracy was best in the low variability condition. . . . Finally, the main effect of time was statistically significant. . . . Decision accuracy was greater in the fast presentation condition (3 seconds) than in the slow presentation condition (5 seconds). . . . [For the confidence variable] the main effects of code and problem size were not significant. . . . Significant effects were found for evidence, time and trial variability, and for the code x time and the problem size x time interactions. A very large effect of weighted difference was found. . . . faster speed generated reliably higher confidence. . . . Increasing problem size from 8-10 increases confidence at the fast rate but diminishes it at the slow rate. . . . The verbal display shows a greater increase in confidence with faster speed than does the spatial. . . . Confidence ratings were higher in low variability trials than in high variability trials" (pp. 19-24).

Experiment Two

Domain: Battlefield scenario
Within Subjects: Undergraduate students

Procedure

Independent Variables: All variables were the same as in the first experiment with the exception of problem size, which was held constant at eight cues, presentation rate, which was held constant at the slow (5 seconds) rate, and variability, in which only high variability cues were used.

Dependent Variables: Accuracy and confidence

Task: The task was the same as the task in the first experiment.

Conclusions

"[For accuracy] the main effect of weighted difference was statistically significant. . . . It is evident that decision accuracy was poorest in the trials of low weighted difference. . . . [For the confidence variable] the main effect of code and the code x weighted difference interaction were not significant. . . . The main effect on weighted difference was very large. . . . This demonstrates that, as in Experiment 1, subjects are becoming increasingly confident as a greater difference in evidence between the competing hypotheses exists" (pp. 30-31).

Overall Conclusions

"The results of Experiment 1 indicate that decision accuracy is enhanced with the integrated spatial display format. . . . The consistent effects of weighted difference on confidence in both experiments demonstrate the ability of the subjects to extract more evidence and therefore increase their confidence as more diagnostic evidence is presented. . . . Examining the integration of the individual dimensions of reliability and diagnosticity in a finer grain revealed two further effects. Experiment 1 demonstrated that a negative correlation between these variables (producing for the spatial display an increase in shape variability) reduced both the accuracy and confidence of prediction. . . . Individual cue values were also examined in Experiment 2 whose data suggested that subjects tended to over-value low levels of reliability, thereby reducing both their accuracy and confidence, relative to the values observed in Experiment 1. . . . The problem size or number of cues within a trial did not have an effect on confidence" (pp. 31-34).

Wright, F. W.
(1987).

A note on the usefulness of graphical displays for decision making. Unpublished manuscript, University of California at Irvine, Graduate School of Management, Irvine.

Domain: Judgments about the relationship between two variables
Mixed: First year MBA students

Procedure

Independent Variables: Display format of the data pair values (scatterplots or tables) and correlation level between the data pair values (no correlation, moderate correlation, and high correlation).

Dependent Variables: Judgment accuracy and confidence rating

Task: "The subjects in the graphical presentation condition saw a . . . scatterplot for each of 3 sets of 60 data points with the same scale values for both the X and Y variables. . . . Subjects in the tabular presentation condition were presented with 60 unordered data pairs on one page. . . . All subjects were told the underlying population means . . . and standard deviations . . . for both variables. . . . The subjects were asked to provide a correlation judgment . . . and an indication of their confidence in their correlation judgment . . . for each of the three correlation conditions" (pp. 8-9).

Conclusions

For judgment accuracy "highly significant main effects [are indicated] for the two presentation modes . . . and for the correlation levels [low correlations resulted in fewest errors, followed by moderate and high correlations respectively]. . . . For the subjects using the graphical displays, noticeably smaller judgment errors are indicated. . . . The presentation mode x correlation level interaction is statistically significant . . . in contrast to the mean (and median) judgments in the graphical condition, the mean judgments in the tabular mode tend to 'flatten out' as the correlation increases. . . . [With respect to the confidence measure] the subjects indicated significantly more confidence in their judgments given the graphical presentation. . . . The main effect for confidence across the three correlation levels is also significant. . . . The presentation mode x correlation level interaction is not significant" (pp. 9-11).

Graphical Displays Empirical Papers

Jarvenpaa, S. L.
(1989).
The effect of task demands and graphical format on information processing strategies. Management Science, 35, 285-303.

Domain: Choosing a restaurant site from multiple possibilities
Mixed: Second year MBA graduate students

Procedure

Independent Variables: Task demands (one of four choice rules subjects were instructed to use: linear, conjunctive, majority of confirming dimension, or elimination-by-aspect) and graphical display format ("(1) An attribute bar chart, arranged by attribute, provided a congruent organization of data for experimental instructions eliciting majority of confirming dimensions and elimination-by-aspect strategies, both of which use attribute-based processing. (2) An alternative bar chart, organized by alternatives, provided a congruent organization for experimental instructions eliciting linear and conjunctive strategies, both of which use alternative-based processing. (3) A grouped bar chart (i.e., matrix arrangement) organized attribute information by alternatives") (pp. 289-290).

Dependent Variables: Acquistion direction (processing of information by attributes or alternatives during acquistion stage) measured by coding the verbal reports, evaluation direction (processing of information by attributes or alternatives during evaluation stage) measured by coding the verbal reports, decision time, and decision quality ("degree of correspondence between the participant's response and the response specified by the rule") (p. 292).

Task: "The four experimental tasks [one for each of the four task demands, which was manipulated within subjects] used in the experiment involved a choice problem. The research participants were asked to select a restaurant site from a set of six alternatives [presented in one of the three graphical display formats] where each alternative was described on seven attributes" (p. 288). In addition, subjects were asked to think-aloud while performing the tasks.

Conclusions

For acquisition direction, "alternative bar charts elicited alternative processing and attribute bar charts elicited attribute processing, but grouped bar charts, contrary to expectations, tended to elicit attribute rather than alternative processing. . . . Thus, support was found for the hypothesis that acquistion direction is a function of the graphical format, not a function of the congruence between the graphical format and the task. . . . [For evaluation direction] the effects of the graphical format were contingent on the task demands. Specifically, attribute bar charts elicited more attribute driven evaluation direction in the MCD [majority of confirming dimensions] and EBA [elimination-by-aspect] tasks (i.e., tasks eliciting attribute processing) than the LNR [linear] task (i.e., a task eliciting alternative processing), but not more than in the CNJ [conjunctive] task (i.e., a task eliciting alternative processing). The alternative bar charts, on the other hand, elicited more alternative-driven evaluation direction in the LNR and CNJ tasks than in the MCD and EBA tasks. Grouped bar charts, by contrast, elicited more attribute-driven evaluation direction in the MCD and EBA tasks than in the LNR and CNJ tasks. . . . the alternative bar charts in the LNR and CNJ tasks (i.e., tasks eliciting alternative processing) required less processing time than the bar attribute charts. . . . In the MCD and EBA tasks (i.e., tasks eliciting attribute processing) attribute bar charts did not appear to provide time savings over the alternative bar charts and were in fact at the disadvantage compared to the grouped charts. . . . However, only a main effect for task demands . . . was found for decision quality. The LNR and CNJ tasks resulted in better performance than the EBA and the CNJ tasks. . . . Participants using grouped bar charts in the LNR task . . . took less time than those using attribute bar charts. . . . The participants using alternative bar charts took less time in the MCD task . . . than those using attribute bar charts. . . . The participants using grouped bar charts also took significantly less time in the EBA task than those using attribute bar charts. . . . Grouped bar charts hence appeared to be a better presentation format in the EBA task than the attribute bar charts. In summary, the results provide some support that participants adapted to incongruent situations by varying their decision time. No effect of congruence was found for decision accuracy" (pp. 294-297).

MacGregor, D., & Slovic, P.
(1986).
Graphic representation of judgmental information. Human-Computer Interaction, 2, 179-200.

Experiment One

Domain: Multiple-cue judgments of marathon completion time
Between Subjects: University students

Procedure

Independent Variables: Graphic display format of cues (bar graph display, deviation display [bar graph presenting each cue as deviation form its mean], spoke display [cues presented on orthogonal radials], or face display [each cue was assigned to a particular facial feature; the "assignment of cues to facial features was done to associate the most predictive cue . . . with one of the more salient facial features" (p. 184)].

Dependent Variables: Lens model statisics including achievement index, matching index, consistency index, and utilization coefficients.

Task: "The task used in this study was designed to require subjects to integrate multiple items of information, having differing importance, into a judgment for which a criterion (true value) existed. Specifically, the task selected required individuals to make estimates of the time (in hours and minutes) it took each of 40 runners to complete a marathon. Each runner was decribed in terms of a set of four information cues" (p. 181).

Conclusions

"The overall effect of subjects' ability to utilize the set of cues effectively is summarized . . . by the achievement index, ra, the correlation between the subjects' judgments of runners' marathon completion times, and the actual completion times. Mean values of ra across the four display conditions differed significantly. . . . The mean ra across subjects was lowest for the deviation display (.39) and highest for the face display (.63). For the spoke display, ra was only slightly greater than that for the deviation display, whereas ra for the bar graph display occupied an intermediate position. . . . The extent to which the system of cues characterizing the environment is reflected in subjects' production of a resonse is indicated by the matching index, G. . . . Both the deviation and spoke displays had the lowest values for the matching index, whereas the bar graph and face displays had the highest [face display was the highest]. . . . Response consistency, Rs, is measured by the degree to which an individual's judgments are predictable from a linear model. . . . Rs values for the four display types ranged from a low of .63 for the bar graph display to a high of .83 for the face display. . . . Utilization coefficients indicate the correlations of subjects' judgments of marathon completion times with each of the information cues. . . . Across display types, however, utilization of the cues generally varied widely" (pp. 187-192).

Experiment Two

Domain: Multiple-cue judgments of marathon completion time
Between Subjects: University students

Procedure

Independent Variable: The authors compared results from the face display used in the first experiment with a new face display. In this experiment, the more predictive cues were assigned to "less salient facial features . . . a tendency opposite that in Study 1" (p. 194).

Dependent Variables: Same lens model statisics as in the first study.

Task: The task was identical to that in the first experiment. However, only the face display was used.

Conclusions

"Overall performance as indicated by the achievement index, ra, was considerably lower for the Face 2 display [face display used in the second study] (.40) than for the original, Face 1, display (.63). Indeed, in terms of achievement, subjects receiving the Face 2 display did only about about as well as those receiving the deviation and spoke displays in the first study. Agreement between the predictive system in the environment and subjects' response production was also poorer for the Face 2 than the Face 1 display. . . . Consistency of responses (Rs) was also lower for the Face 2 display. A partial explanation for the degraded performance of the face display under an alternative assignment of cues can be seen by examining the cue-utilization coefficients. . . . Although utilization coefficients for the Face 1 display approach those of cue validities, the utilization coefficients for the Face 2 display are lower on average and are distributed more flatly. Moreover, cue utilizations for Face 2 are similar to those for the two poorest performing displays from Study 1. . . . Apparently subjects exposed to the Face 2 display were less able to utilize the information portrayed than were individuals receiving the Face 1 display" (p. 195).

Overall Conclusions

"Taken together, these studies demonstrate the strong effects of graphic display formats on the quality of judgmental performance. . . . the lens model analysis strongly suggested that some display formats [i.e., face display used in the second experiment, the deviation display used in the first experiment] can lead to confusion in the use of information cues. Conversely, some graphic formats [i.e., face display used in the first experiment] may be better than others at aiding the user in developing a consistent scheme for relating display features to the requirements of the judgment task" (p. 196).

Schutz, H. G.
(1961).
An evaluation of formats for graphic trend displays--Experiment II.
Human Factors, 3, 99-107.

Domain: Analysis of trend displays
Within Subjects: Professional male Battelle employees

Procedure

Independent Variables: Display format (Cartesian line graph, vertical bar graph, or horizontal bar graph) number of points per display (6, 12, or 18) amount of missing data per display (none, one-sixth, or one-third), subjects (ten randomly selected subjects) and replication (first or second).

Dependent Variables: Response time and accuracy (0 points were given if the subject's response was both wrong in direction and probability, 1 point was given if the subject's response was either correct direction, wrong probability or wrong direction, correct probability, and 2 points were given if the subject's response was correct direction and correct probability).

Task: Subjects were first given a general orientation which included the procedure to be followed. They were asked to memorize some rules which would help them in analyzing the trend displays. Subjects were then presented with the various displays in which a set of six points had been chosen as the set of points about which the subject would make a judgment. Subjects made judgments about the direction the trend level was moving (upward, downward, or staying the same) as well as finding the trend with the highest probability of occurrence.

Conclusions

For the time score analysis, "mean squares were computed for only the main effects and all two- and three-way interactions for the replication variable. . . . All main effects were significant" (p. 105). Line graph resulted in the fastest time score followed by vertical bar graph and horizontal bar graph respectively. Six points resulted in the fastest time score followed by 12 points and 18 points respectively. Zero missing data resulted in the fastest time score followed by one-sixth missing data and one-third missing data respectively. "Two of the two-way interactions, subjects by replications and format by missing data, were significant. . . . It is apparent that the three formats do not differ significantly from one another at the one-third level of missing data, whereas at the other two levels of missing data, the formats are significantly different [line graph with zero missing data resulted in the fastest time, followed by vertical bar graph with no missing data, line graph with one-sixth missing data, horizontal bar graph with zero missing data, vertical bar graph with one-sixth missing data and horizontal bar graph with one-sixth missing data]. The significant subject by replication interaction signifies that some subjects improved in performance for the second replication and others did not. . . . because the accuracy scores did not lend themselves to ordinary analysis-of-variance techniques, and, further, because the correlation (r=-0.90) between time and accuracy was so high, the accuracy data were not extensively analyzed. A Friedman two-way, non-parametric analysis of variance was conducted on the accuracy scores for the three main effects of format, number-of-points, and missing data. . . . The almost identical means for the replications did not require an analysis. This analysis, however, resulted in significant differences among the levels for all the remaining variables" (pp. 105-106). Line graph resulted in the most accuracy followed by vertical bar graph and horizontal bar graph respectively. Six points resulted in the most accuracy followed by 12 points and 18 points respectively. Zero missing data resulted in the most accuracy followed by one-third missing and one-sixth missing data respectively. "No significant [interaction] effects were found. . . . The significant effects of format on both speed of performance and accuracy of performance indicate clearly that the line graph resulted in superior performance compared with the vertical-bar or horizontal-bar type. . . . The interaction between format and missing data indicates that at the higher level of missing data (one-third missing), the three formats do not differ significantly. . . . The number of points variable can be viewed as a type of irrelevant data situation. The subjects must search for a particular set of four to six points in a context of points which are of no value. The degradation of performance that results from an increase in irrelevant points indicates that for situations comparable to these experimental conditions it would be better to display only the minimum number of points that will be needed in looking for a trend" (pp. 106-107).

Simkin, D., & Hastie, R.
(1987).
An information-processing analysis of graph perception. Journal of the American Statistical Association, 82, 454-465.

Experiment One

Domain: Discrimination, comparison and proportion judgments
Within Subjects: Undergraduate students

Procedure

Independent Variables: Elementary code (position code which was a simple bar chart, length code which was a divided bar chart, and proportion code which was a pie chart) and trial blocks (1-10).

Dependent Variables: Reaction time and accuracy

Task: "Subjects were seated at the computer terminal and the task was explained to them. [Following 10 practice trials] the subjects then worked uninterrupted through the 90 experimental trials . . . They began each trial by pressing the space bar with their thumb, causing the graph [the paired bar or pie charts] to be displayed. As quickly as possible, they were to indicate if the division on the left [bar or pie chart] or right [bar or pie chart] was smaller . . . (discrimination judgment). . . . After the discrimination response, the text signaled them to enter a number from 1 to 100 that was their judgment of the percentage the smaller division was of the larger (the comparison judgment). . . . The procedure for the proportion judgment was quite similar. . . . The task was to judge the percentage that the division represented of the whole bar or pie" (p. 455). In the comparison judgment, all graphs were displayed for one second. In the proportion judgment, all graphs were displayed for 0.5 seconds.

Conclusions

"Elementary code was a significant factor in the analysis of the discrimination reaction time. . . . The reaction time for the discrimination judgment of the position code was significantly faster than that of the other two codes. . . . (subjects made few errors [on the discrimination task]). . . . Elementary code was a significant factor in the comparison absolute-error analysis. . . . Position yielded the most accurate comparison judgments, followed by length then angle, which was least accurate. The comparison reaction-time data show the same ordering, but the difference between length and angle is not significant. Not surprisingly, elementary code accounted for a significant amount of variance in both the proportion absolute-error analysis . . . and the proportion reaction-time analysis. . . . There was a reversal in the ordering of the codes by accuracy for the proportion judgment from the comparison judgment. Length led to significantly fewer accurate judgments than the other two codes, which did not differ from each other. Although the angle judgments were the most accurate, they also took the most time to make" (p. 456).

Experiment Two

Domain: Discrimination, comparison and proportion judgments
Within Subjects: Undergraduate students

Procedure

Independent Variables: Same as in the first experiment.

Dependent Variables: Same as in the first experiment.

Task: Same as in first experiment except that the graphs were left on the screen until the subjects made their estimates.

Conclusions

"The analysis again revealed elementary code to be a significant factor for all dependent measures [discrimination reaction time, comparison judgment errors, comparison judgment reaction time, proportion judgment errors, and proportion judgment reaction time]. . . . The direction of the differences replicated Experiment 1 completely. Experiment 2 replicated the elementary code by judgment-task interaction from Experiment 1. Note also that the subjects took more time in Experiment 2 than in Experiment 1 to make their proportion and comparison judgments, resulting in more accurate responses" (p. 457).

Overall Conclusions

Based upon the results from the two experiments, the authors formulated "an information-processing theoretical analysis of our graph-perception findings" (p. 457) which "points to anchoring as the key process for proportion and comparison judgments. When making a proportion-of-the-whole judgment, the more accurate anchoring possible with position and angle codes accounts for their superiority over the length code. Although processing angles is more difficult than processing linear aspects, this judgment for the pie chart is a special case in which the anchors at the perceptually salient angles of 0o, 90o, and 180o. When making a comparison judgment, the position code is superior to the other two codes. Length again suffers from less accurate anchoring. Angles provide the least accurate estimates because of the inferior anchoring when these angles are not longer at perceptually salient angles. . . . We have tried to make three major points in our analysis. First, people have schemata for graphs that include slots for the conceptual message of the graph. Second, we demonstrated that elementary code and judgment task interact to determine performance. Third, we proposed elementary processes of anchoring, scanning, projection, superimposition, and detection operators to explain these interactions" (p. 465).

Verhagen, L. H. J. M.
(1981).
Experiments with bar graph process supervision displays on VDUs.
Applied Ergonomics, 12, 39-45.

Note: It is unclear whether variables were manipulated within or between subjects.

Experiment One

Domain: Detection of over and underflow alarms
Within Subjects: Undergraduate students

Procedure

Independent Variables: Display format (bar graph and stroke-type graph in which only the end of the bar graph is presented) and number of deviating variables per slide (two and nine).

Dependent Variables: Search time and proportion of errors

Task: Subjects were presented with the displays and asked to detect overflow and underflow alarm variables. Subjects counted the number of deviating variables (underflow and overflow alarm variables) and called out the number once they had finished counting. Twenty-four variables were presented on a display at one time.

Conclusions

"The overall mean time needed for counting the number of deviating variables was 7.5 s for bar-type and 6.1 s for stroke-type presentation. . . . An unpaired t-test (one-sided) on the results yielded significance. . . . The number of deviations that had to be found had a strong influence on search time. An analysis was made for the cases of two and nine deviations per slide. . . . [An] analysis of variance with type (bar and stroke) and number (2 and 9) as variables showed a significant number effect . . . but strangly enough, no type effect. . . . Further an interaction effect was found . . . which indicated an advantage for the bar type in the case of nine deviations. . . . With the bar-type nearly twice as many errors were made as with the stroke-type. . . . From Experiment 1 a strong advantage appeared for the stroke type, not only in terms of search time but certainly also in terms of errors" (p. 41).

Experiment Two

Mixed: Technical students training to be process operators

Procedure

Independent Variables: Display format (bar graph or stroke-type), presentation time (1 or 2 seconds), number of deviating variables per slide (0, 1, 3, and 6), and search mode (begin searching from a fixed point or from any point chosen by the subject).

Dependent Variables: Proportion of errors

Task: In this experiment, subjects noted the deviating variables (under and over-flow alarms) and named their code numbers after the display had been removed. Displays lasted for 1 or 2 seconds.

Conclusions

"No significant differences in error rate were found between the two search modes (starting from centre fixation point or free strategy). A paired t-test on the differences between bar and stroke type showed significant results . . . in favour of the stroke-type. . . . With the stroke-type more errors were made in naming the deviating variables. . . . Also it seemed that with the stroke-type more low alarms were missed than with the bar-type, but these results were not significant. . . . In conclusion, for Experiment 2, one may state that the stroke-type could be scanned more completely than the bar-type in the same period of time. However, when under the quite extreme time stress of one or two seconds there is a danger that, with the stroke type, errors are made in finding the corresponding name of the variable" (p. 42).

Experiment Three

Domain: Detection of over and underflow alarms
Mixed: Undergraduate students
Procedure

Independent Variable: Display format (horizontal bar graphs or vertical bar graphs) number of choices per minute for the distractor task which was in the form of a binary choice task (0 choices per minute, 20 choices per minute, and 40 choices per minute), and task type (counting and identification tasks).

Dependent Variables: Mean search times.

Task: In this experiment, subjects both counted and named the deviating variables. Each display remained until an answer had been given. Twenty-four variables were presented at one time. In addition, "a distractor task was provided by means of a binary choice generator. . . . On both sides of the TV monitor a small lamp was mounted. At regular intervals one of the two lamps was randomly lit. A pedal corresponding to this lamp was then to be actuated by the subject. The frequencies used were 0, 20 and 40 choices per minute" (pp. 40-41).

Conclusions

"With the counting task, a comparison was made between horizontal and vertical. The counting time was shorter for vertical than for horizontal bars. . . . The search times for the counting task were shorter (6.7 s) than for the identification task (7.4 s). . . . However, with a parametric t-test no significance was found. Together with the identification task, a distraction task was included in the form of a binary choice task with 0, 20 and 40 choices per minute. For all three distraction task conditions no significant differences between horizontal and vertical bar-types were found. . . . There were, however, significant results in the mutual comparison of three distraction tasks. With 20 choices, search times were longer than with no choices; and with 40 choices, search times were longer than with 20 choices. . . . In conclusion, one may say that no preference may be expressed for horizontal or vertical bars. Only for the relatively simple counting task, a quick overview, do the vertical bars seem somewhat better. With the identification task, code numbers of deviating variables had to be memorised until all deviations were found. Apparently this memorising process is strongly affected by a binary choice task" (p. 43).

Experiment Four

Domain: Detection of over and underflow alarms
Mixed: Undergraduate students
Procedure

Independent Variable: Display format (t-type bar graphs or stroke-type bar graphs), number of choices (30 choices per minute and 60 choices per minute), number of deviating variables per slide (0, 1, 2, 3, 4, 5, and 6) and alarm level (slides with only low alarms and slides with only high alarms).

Dependent Variables: Mean search times and proportion of errors.

Task: In this experiment, subjects called out the code number of the deviating variable right after it had been detected. Sixty variables were presented at one time. In addition, a binary choice task was also used with 30 and 60 choices per minute.

Conclusions

"The differences between choice task levels (i.e., 30 and 60) are not significant [with regard to search times]. . . . The differences between stroke-type and T-type are significant [in favor of the stroke-type]. About the same level of significance is reached if one analyses the data for each case of deviations separately (0 to 4 only). If the number of deviations increases, search times increase as well. . . . Both for the stroke-type and for T-type the difference in search times is significant. . . . It is surprising that for the stroke-type, low alarms are detected more quickly than high alarms. . . . All subjects made more errors with the T-type than with the stroke-type. . . . Many errors were made during the first session in Experiment 4, especially if the subject started with the 60 choices condition. In total, more errors (omissions) were made in low alarms than in high alarms" (pp. 43-44).

Overall Conclusions

"An attractive alternative for bar graph process supervision displays on VDUs seems to exist in the form of a stroke-type display. . . . With respect to the alarm detection tasks as described, there is no difference in performance between horizontal and vertical bars. . .

. The memorising of code numbers is very much influenced by distraction. . . . Task complexity influences results of experiments as described in this paper. In Experiment 3, with counting tasks, a difference in results appeared in favour of the vertical bar. This difference disappeared with the more complicated identification task. . . . With bar (or T-) graphs, nearly twice as many alarm detection errors are made compared with the stroke type graphs. . . . Within stroke type presentation, most errors are misinterpreted of, at most, one place. Within bar type presentation most errors take the form of missed alarms. Low alarms are especially easily overlooked" (p. 44).