Journal of Parapsychology, Vol. 55, September 1991

PREDICTION OF FORCED-CHOICE ESP PERFORMANCE

PART III: THREE ATTEMPTS TO RETRIEVE CODED INFORMATION USING MOOD REPORTS AND A REPEATED-GUESSING TECHNIQUE

By James C. Carpenter

ABSTRACT: Three studies were done attempting to predict the hitting rate and the run-score variance that would be achieved by individuals taking, in solitude, a forced-choice ESP test. The predictors of performance were responses to the sheep-goat question and scales based on adjectives describing the subject's mood at the time of testing. Each study tested new mood scales, which were generated by stepwise multiple regression analyses on larger bodies of data. The targets of each study carried encryptions of information that the study was attempting to retrieve, and each subject in the study used the same targets repeatedly. Because past work found the data of low-authoritarian subjects (California F Scale) to be the most reliably related to mood reports, these subjects were of interest here. Extreme-quartile scores on the mood scales were used to generate predictions about the hitting rate and size of variance. Repeated-guessing analyses of calls, following fixed rules, were carried out to generate predictions about target content. When the scales expected to be the most reliable were used, each study showed some success at predicting hitting rate and at retrieving the encoded material. The prediction of variance was successful in only one study. Discussion focuses on an interpretation of the importance of the F Scale as a moderator variable, the sheep-goat results, the utility of individual self-testing, and an interpretation of the findings regarding psi-facilitative moods or states of mind.

Could ESP reliably serve the needs of human communication? If that peculiar ability is ever to move beyond the airy realms of science fiction and academic dispute, research will have to guarantee two qualities: a high degree of accuracy and sufficient consistency. By accuracy I mean how strongly or densely correct a clump of information is (how high a proportion of correctness versus incorrectness it contains), and by consistency I mean how much the retriever of information can be counted on to give extra-chance

Study 1 of this research was supported by a grant from the John E. Fetzer Foundation; Studies 2 and 3, by grants from the Parapsychology Foundation. The Foundation for Research on the Nature of Man provided help and facilities at several points. Jeff Brooks, Donald Burdick, and John Carroll provided helpful statistical and methodological advice. David Rogers, Gertrude Schmeidler, and two anonymous reviewers gave valuable suggestions for text revision.

results time after time. There are anecdotes about famous psychics and family stories about "Great-Aunt Helen who had second sight" that show a highly accurate and lucid grasping of information; but even famous psychics get so much wrong (Boerenkamp, 1985, 1986; Roll & Tart, 1965) and most of Aunt Helen's premonitions were probably worthless. Hunches that are occasionally powerful but often unreliable are a recipe for disaster in the practical world. On the other hand, there is a respectable measure of consistency that has been attained in the experimental literature about ESP. Meta-analyses (e.g., Honorton & Ferrari, 1989) show that certain ESP effects are reliable enough across many studies to be statistically significant and presumably real. But statistically reliable effects are generally very weak, an average of a hit or two in excess of chance expectation out of many trials. Although consistent, such statistical effects are not strong enough to be practically useful.

The fact seems to be that except for a few possible, rare exceptions (e.g., "B.D.," in Kanthamani & Kelly, 1974) ESP ability that is both highly accurate and consistent is hardly found naturally in the world. I *have* heard a few people describe their ESP ability as very accurate and absolutely consistent; but I met them in my role as a clinical psychologist, not as a parapsychologist, and all of them were psychotic.

Could the accuracy and consistency of the ESP capacity be enhanced artificially if enough understanding were gained about it? It seems a valid possibility.

This report describes three studies aimed at contributing to our understanding of both the problems of consistency (or reliability) and degree of accuracy. In these studies, I attempt to develop measures of the subject's state of mind in the hope that they may predict ESP performance in a reliable way; and I explore some techniques of combining a great deal of relatively weak guessing effort into a more powerfully correct distillation. These studies are a continuation of previous work (Carpenter, 1983a, 1983b) but with some major modifications.

Prior Work

Twelve experimental studies (3 pilot and 9 confirmatory studies) were done previously in which an ESP test, a self-report of mood on a mood adjective check list (MACL), and some attitude questions were used. These studies tested the validity of an empirically derived scale of mood adjectives (V scale) in predicting ESP run-score variance and used predictions from the scale in repeated-guessing analyses aimed at enhancing degree of accuracy. Subjects carried out self-testing alone. The testing mode was precognitive, as targets were always determined randomly after all subjects' calls for the studies had been collected. Two targets (usually "+" and "0") were always used.

Run-score variance (henceforth simply called *variance*) was calculated around the theoretical expected score for the run (12 for a run of 24 trials). Some moderator variables for the V scale were studied to see if certain subgroups of subjects or types of targets would be more predictive of variance.

Without the subjects' knowing it, each study had only a single target list against which all calls would be scored. This was done so that all calls could be combined in a single, overall set of majority decisions. Targets in the list were divided into "index" and "message" subsets, and the identity of "message" targets in the list was predicted by a set of rules using the observed performance on "index" targets, the V-scale predictions for each session, and a compilation of the subjects' guesses rendered into votes for the symbols + and 0. This predictive process is called a *repeated-guessing analysis.*

Thus, all these studies explored the reliability of the V scale for predicting variance as a function of subject type and so forth and also tested the utility of the repeated-guessing procedure, assuming that the V scale had the ability to predict variance.

At the conclusion of the 12 studies, some findings seemed consistent enough to be worth further study. The V scale seemed an effective predictor of variance for certain groups of subjects, and an index-sampling, repeated-guessing technique did seem to enhance the degree of accuracy of results for subjects whose scoring was adequately predicted by the V scale.

Three further studies are reported here. In them, an effort was made to maintain procedural consistency, but some changes were made in the methods of data collection, target selection, and analysis. In each study, as in the previous ones, two basic ways of analyzing the data were used: one aimed at directly assessing the reliability of mood-scale predictions; the other intended to test the usefulness of repeated-guessing methods for enhancing the degree of final accuracy. New mood scales were generated for each study.

Current Studies As an Attempt to Develop and Test Reliable Scales

Scale development procedure. The procedure used for scale derivation was forward-stepping stepwise multiple regression (Kleinbaum & Kupper, 1978; Pedhazur, 1982; Wilkinson, 1988). By this method, the most powerful set of variables predictive of some criterion are selected, one at a time, out of a larger set of potential predictors by virtue of the strength of their independent contributions to the regression equation upon the criterion. The set of variables can then be tested for its capacity to predict the criterion in a new set of data.

One factor that is expected to influence the reliability of the solution reached is the number of cases relative to the number of potential predictor variables examined. The larger the *N* the better. The analysis that produced the V scale used 283 cases and tested 54 variables against the variance criterion. The variables were mood items (coded "0" or "1"). No interaction terms were used. This data set will henceforth be referred to as "Set A." Such a small *N* would be expected to yield relatively unreliable results. For this reason, after the first study, new scales were developed with the larger *N* provided by including all cases used in earlier validating efforts. This practice is common in developing psychological instruments and can continue indefinitely until the judgment of "reliable enough" is reached. With relations as weak as those studied here, that process could continue on profitably for many hundreds more cases. Because this research strategy has not been commonly used in parapsychology, it is summarized in Figure 1.

In all these studies, scales were generated to predict hitting rate as well as variance.

Moderator variables. Two moderator variables are used in these studies in an attempt to find groups of subjects whose performance is more strongly predicted by the mood scales. The first and most important is the California F Scale (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950). Some defining characteristics of the "authoritarian syndrome" that the F (or fascism) Scale was designed to measure are anti-intraception, repression, and suggestibility. High-authoritarian people are assumed to be generally rigid in their perceptions of others and themselves (Jackson, Messick, & Solley, 1957; Scodel & Mussen, 1953), generally less likely to be interested in the personal inner life, are suspicious of introspection (Kogan, 1956), are intolerant of any but certain conventional feelings and attitudes in themselves and others (Siegel, 1956), are generally unaware of many of their actual feelings and motivations, and are assumed to be especially likely to attribute to themselves whatever perceptions or inner states they think are desired by authorities (such as experimenters) and by conventional morality (Barron, 1953; Kogan,

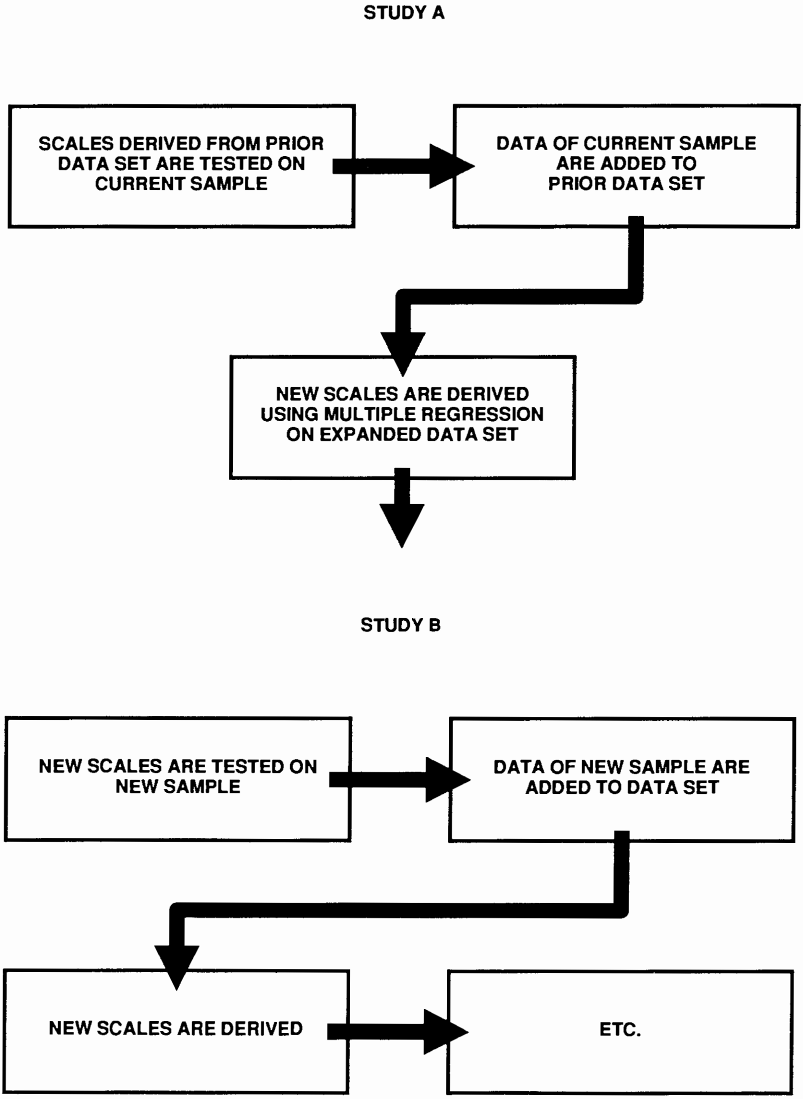
STUDY A
SCALES DERIVED FROM PRIOR DATA SET ARE TESTED ON CURRENT SAMPLE
DATA OF CURRENT SAMPLE ARE ADDED TO PRIOR DATA SET
NEW SCALES ARE DERIVED USING MULTIPLE REGRESSION ON EXPANDED DATA SET
STUDY B
NEW SCALES ARE TESTED ON NEW SAMPLE
DATA OF NEW SAMPLE ARE ADDED TO DATA SET
NEW SCALES ARE DERIVED
ETC.

Figure 1. Research strategy for deriving more effective mood scales in successive studies.

1956). Therefore, they should tend to be unreliable reporters of inner experience. The truth of this proposition was demonstrated in an especially pertinent way by Thayer (1971), who showed that mood reports similar to the ones used here matched physiologically measured changes in state only for the data of low-authoritarian subjects. The self-reports of high authoritarians did not reliably correlate with their own internal states of arousal and nonarousal.

It was found, and confirmed in several studies previously (Carpenter, 1983a), that the F Scale did discriminate between subjects whose mood reports were predictive of their ESP performance and those whose reports were not. For this reason, the F Scale is also used in the present series as a moderator for all mood scales.

The second moderating variable used in these studies is the "sheep-goat" attitude question of Schmeidler and McConnell (1958): "Do you believe that ESP is possible under the conditions of this experiment?"[1] It had been noted empirically, and confirmed, in earlier research (Carpenter, 1983a) that subjects who were students of the experimenter and who were goats (disbelievers) did not produce discriminative scores on the V scale, whereas sheep did. For subjects who were *not* students of the experimenter, the sheep-goat question did not matter: both groups produced discriminative scores. This variable is used as a moderator for the V scale only in Studies 1 and 2, since in Study 3 no subjects were drawn from classes taught by the experimenter (it is used as an independent predictor of ESP performance in Study 3).

Current Studies As Tests of a Repeated-Guessing Technique As a Method for Efficient Information Retrieval

Repeated guessing and majority vote. In part, these studies are intended to represent an approximation of a situation in which an ESP-testing procedure might be used to accurately retrieve unknown information. A first step in such an effort is to find some means to heighten the typically weak level of guessing success in ESP experiments. Most efforts to do this in the past have used some variation of a repeated-guessing, majority-vote procedure (Fisk & West, 1956, 1957; Kennedy, 1979; Ryzl, 1966; Taetzsch, 1962; Thouless, 1960). This research does also. These approaches all exploit the fact that even a relatively weak degree of above-chance calling of targets, if consistent, can be improved by making many calls

[1] Although this form of the sheep-goat question is the one used by Schmeidler, the response permitted the subjects in my studies is a bit different. In these studies only a "yes" or "no" was allowed, whereas Schmeidler also used an "uncertain" response and counted those subjects as sheep. A few subjects in these studies rejected either alternative and described themselves as uncertain: they were classified as sheep.

at each target. The numbers of calls for each symbol can be summed for each target (e.g., a total number of calls for + and for 0 for Target 1). These sums can be considered "votes," and the symbol with the larger number of votes can be nominated as the best guess for the whole body of data. These majority decisions, given above-chance data, will tend to be correct more often than the original body of data as a whole will be. In these studies, *two* repeated-guessing techniques are actually used. One technique uses predictions of variance and performance on index targets, and the other uses predictions of hitting rate. These are explained under "Study 1, Repeated-Guessing Analyses."

It was decided that for the first two of these studies, a verbal target-message would be chosen beforehand (changing the testing paradigm from precognitive to GESP) and that the verbal material would be translated into a sequence of binary forced-choice symbols (+ and 0). Thus, the targets could "carry" the verbal information in encoded form.[2] The third study used a randomly chosen number, encoded into bits, as a target-message. It was hoped that these "clumps" of verbal and numerical target information might be "retrieved" with a high degree of accuracy.

It was also decided that larger numbers of subjects would be tested in these as compared with previous studies, with the hope of achieving more reliable results.

General model for information exchange. Thus, these studies were carried out partly as an exercise in the exchange of information. It is as if one researcher in City A is trying to "send" some information to another researcher in City B, who is trying to "receive" it through an experimental protocol. The model uses persons playing three roles: an *originating experimenter,* who produces the information to be exchanged; a *receiving experimenter,* who has prior knowledge of some of the originator's information in the form of index targets, but not the rest; and a set of *subjects,* who attempt to guess the information over and over while also providing some independent predictor(s) of ESP performance (mood ratings, in these studies).

As in previous work, the originator's information in these studies was a list of 24 binary targets, each of which was comprised of two

[2] The target's being chosen by the experimenter is a clear violation of a procedural rule that has become a norm for parapsychological research: namely, that target order is picked at random. However, it is difficult to see a practical problem in this case. For one thing, not the target content but rather an encoded representation of the content is guessed at by the subjects. Also, the encoded content is broken up by other targets that are randomly chosen. Finally, the guesses as compiled are rendered in complex ways depending on mood responses, with many calls omitted and many others reversed according to criteria the subject could not know.

subsets of items: message and index trials. The subjects were asked to guess runs of the symbols + and 0 but were not told that target content was being repeated across runs or that some targets would be treated differently than others.

Contrary to previous work, in Studies 1 and 2 the items of the message subsets of targets were to stand for the dots and dashes of the Morse code: a dot was represented by a "+," and a dash by a "0." For both studies a word was chosen, the letters of which could be represented by an appropriate number of dots and dashes, thus permitting an attempt at retrieving coded verbal content. The following steps are involved in this information-exchange procedure for Studies 1 and 2. (The procedure used in Study 3 is an elaboration of these steps and is discussed in the section on Study 3 under "Target Preparation.")

1. The originating experimenter selects a verbal message codable into binary items (the Morse code dots and dashes). These are converted to equivalent ESP target symbols (+ and 0). The set of symbols representing each letter are arranged in contiguous strings within the whole target list, separated by other targets, randomly chosen, which will serve as indices.[3] Thus, a single target order is established for scoring each run in the study. This feature was changed in Study 3.

2. The receiving experimenter is given a target array containing blank spaces (unknown message items) and filled spaces (index items with target content given). This is illustrated in Table 1, in which 12 items are intended to represent a message and 12 are index items. This array is, in fact, the one used for Study 1.

3. Some subjects are enlisted to carry out an ESP task; their performance is presumed to be independently predictable by some means.

4. The receiving experimenter scores the predictor variables and uses them to make predictions about the subject's ESP performance. Some 5-run sets would be expected to show psi-hitting, some would show psi-missing, some would show large variance, and some small. Some sets would be assigned no predictions.

[3] The index targets are used here only in conjunction with independent mood scales intended to predict RSV (the average size of run-score deviations). If the size of deviation can be predicted, then the information from index target scoring can be used to predict the scoring of the remainder of the run. See Carpenter, 1983b, for a more detailed explanation. It must be remembered that the index targets are not intended to provide simply a measure of scoring success, with the assumption that this measure would predict the direction of scoring in the rest of the run. Though this idea seems intuitively appealing, it is generally wrong, because one cannot count on internal consistency in scoring within ESP runs.

TABLE 1

TARGET ARRAYS FOR ORIGINATOR AND RECEIVER ROLES IN STUDY 1

	Originator's array					Receiver's array	
Target no.	Target letter	Morse code	Message target	Index target	Target list	Target type	Content
1				+	+	index	+
2	P	·	+		+	message	?
3		–	0		0	message	?
4		–	0		0	message	?
5		·	+		+	message	?
6				0	0	index	0
7				0	0	index	0
8	E	·	+		+	message	?
9				0	0	index	0
10				0	0	index	0

Table 1 continued

	Originator's array					Receiver's array	
Target no.	Target letter	Morse code	Message target	Index target	Target list	Target type	Content
11	A	·	+		+	message	?
12		–	0		0	message	?
13				0	0	index	0
14				0	0	index	0
15	C	–	0		0	message	?
16		·	+		+	message	?
17		–	0		0	message	?
18		·	+		+	message	?
19				0	0	index	0
20				+	+	index	+
21				0	0	index	0
22	E	·	+		+	message	?
23				+	+	index	+
24				+	+	index	+

5. Index targets are scored.

6. Predictions (summary "best guesses") about the content of message items are generated by the receiver from the predictor variables, the subjects' performance on index items, and the subjects' calls on message items using a repeated-guessing technique. This process is spelled out below. See Figure 2 for a flowchart summarizing this information-transfer model.

In Studies 1 and 2, I played both the roles of the originating experimenter and the receiving experimenter. In Study 3, an independent experimenter played the originator role.

Study 1: The "Peace" Experiment

Two major changes were made in this study as compared with previous studies: a GESP target list was determined prior to data collection, and a new scale of mood items, aimed at predicting hitting rate, was tested.

Target Preparation

Prior to any data collection, I chose the target word *peace* as appropriate in both content and in Morse-code make-up. A message-target array was made up, as just described; it was composed of 12 pluses and 12 zeros separated into five blocks (letters). Twelve index items were derived by the same random procedure used in previous studies, using the temperatures recorded in that morning's newspaper as an entry point in the RAND book of random numbers (Rand Corporation, 1955). The five blocks of message items were spread sequentially down the target column, with at least one index target separating each block, and some attention being paid to spreading the blocks in a roughly even way all through the run. See Figure 1 for the originator's target list and the partial array accorded to the receiver role. Again, contrary to the procedure in most ESP experiments, only this one target list was used repeatedly for all runs in this study.

After constructing the target list, the experimenter sealed it in an envelope and placed it under a clutter of papers in a desk drawer. No data had yet been collected. No one was told of the intention to retrieve this word (or any word) until after all subjects had completed their work.

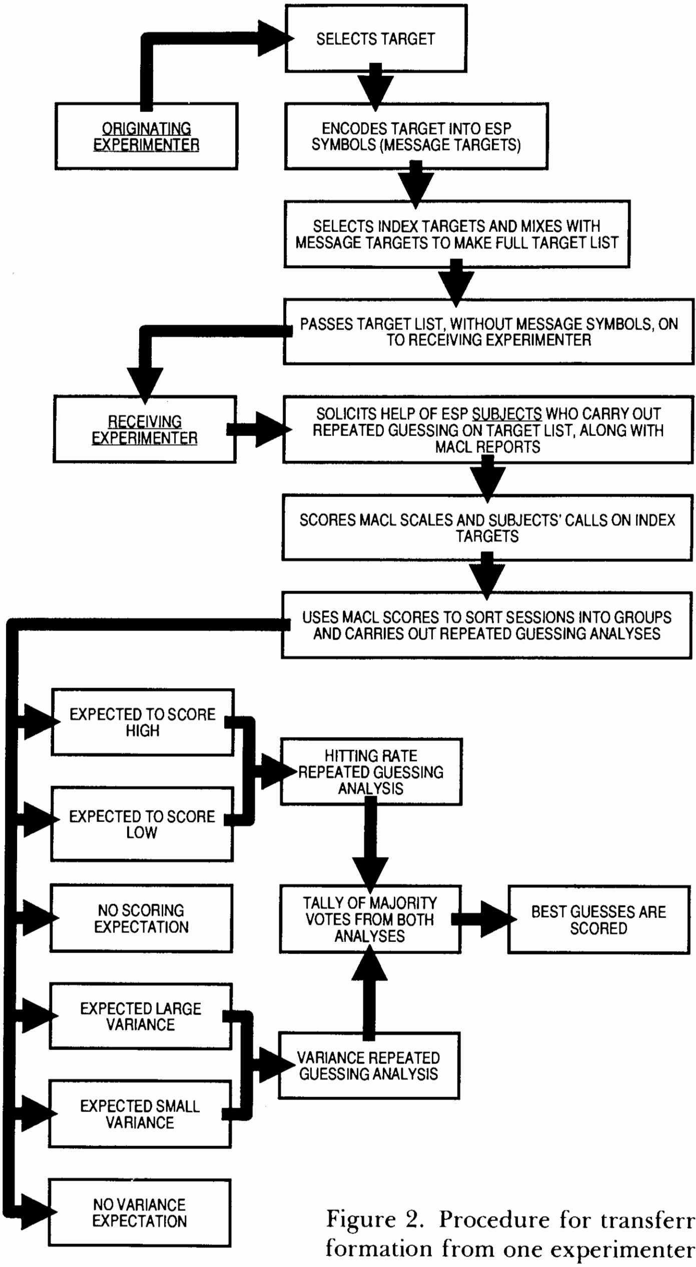
SELECTS TARGET
ORIGINATING EXPERIMENTER
ENCODES TARGET INTO ESP SYMBOLS (MESSAGE TARGETS)
SELECTS INDEX TARGETS AND MIXES WITH MESSAGE TARGETS TO MAKE FULL TARGET LIST
PASSES TARGET LIST, WITHOUT MESSAGE SYMBOLS, ON TO RECEIVING EXPERIMENTER
RECEIVING EXPERIMENTER
SOLICITS HELP OF ESP SUBJECTS WHO CARRY OUT REPEATED GUESSING ON TARGET LIST, ALONG WITH MACL REPORTS
SCORES MACL SCALES AND SUBJECTS' CALLS ON INDEX TARGETS
USES MACL SCORES TO SORT SESSIONS INTO GROUPS AND CARRIES OUT REPEATED GUESSING ANALYSES
EXPECTED TO SCORE HIGH
EXPECTED TO SCORE LOW
NO SCORING EXPECTATION
EXPECTED LARGE VARIANCE
EXPECTED SMALL VARIANCE
NO VARIANCE EXPECTATION
HITTING RATE REPEATED GUESSING ANALYSIS
TALLY OF MAJORITY VOTES FROM BOTH ANALYSES
BEST GUESSES ARE SCORED
VARIANCE REPEATED GUESSING ANALYSIS

Figure 2. Procedure for transferring information from one experimenter to another by the repeated-guessing protocol.

TABLE 2
SCALE ITEMS AND SCORING WEIGHTS OF V_A AND H_A SCALES

V_A scale	H_A scale
amiable (− 39)	adaptable (− 23)
bashful (51)	adventurous (− 39)
carefree (− 22)	aloof (− 29)
dizzy (30)	bashful (− 53)
downhearted (− 35)	cooperative (33)
drifting (− 31)	drifting (30)
fearless (37)	exultant (− 95)
intoxicated (− 94)	intoxicated (− 88)
lazy (43)	light headed (30)
task involved (− 25)	masterful (28)
tired (− 17)	sluggish (31)
	task-involved (− 20)
	tired (− 53)
	unsure (43)
	witty (34)

Note. V_A scores of 0 or more give a large RSV prediction; scores of − 53 or lower give a small RSV prediction. H_A scores of 28 or more give a psi-hitting prediction; scores of − 3 or lower give a psi-missing prediction.

Re-Derivation of Mood Scale Predictive of Variance

Unhappily, in cross-checking all analyses leading up to these studies, I discovered several key-punching errors in the data sets that had been used to generate the original V scale. A few mood items had been put in incorrect columns (in spite of double-checking by assistants at the time), but these were inconsequential in the regression analysis. One error was of consequence, and that was the misplacement of one variance value, which resulted in its being overestimated by a factor of 10. This did distort the original analysis. The corrected data were reanalyzed with forward-stepping multiple regression (*SPSS User's Guide*, 1983) using the same criteria for item selection as before. The three most weakly loaded items from the previous analysis dropped out, and two new ones emerged. The corrected V-scale items, along with Item B weights rounded to the nearest second digit, are given in Table 2. Generated from Data Set A, it is referred to, henceforth, as the V_A scale. Each MACL was to be given a total score based on weighted responses to these items.

Scores in the highest and lowest quartiles were used to generate dichotomous predictions about the size of variance.

Derivation of Mood Scale Predictive of Modal Trend

Up to this point, preoccupation with variance had led me to neglect the possibility of predicting directional trend (the overall tendency within the session toward hitting or missing) directly. In hopes of making the most of the data generated by the subjects in this series, I determined a set of MACL items by the same sort of forward-stepping multiple regression procedure, and on the same data, as I had used to produce the V_A scale. An experimental hitting scale (H_A scale) was generated, which postdicted optimally the session scores (scores summed across five runs) of the originating data (Carpenter, 1968, 1969). Fifteen mood adjectives comprised the scale. They are listed in Figure 2. For the repeated-guessing analysis, the extreme-quartile sets of scores on this scale were used to predict directional trend on the accompanying ESP data. Top-quartile sets were expected to show psi-hitting, and bottom-quartile sets were expected to show psi-missing.

Procedure

Subjects. There were 110 subjects, who were drawn from three psychology classes taught by me and from 5 classes taught by my colleagues at the University of North Carolina in Chapel Hill. The format of lecture, discussion, and instruction was kept unchanged from previous studies. The experimenter went to each class at the beginning of the period and delivered a 10- to 15-minute introductory lecture about ESP research, focusing on the procedures to be followed in the study.

Test materials and instructions. Subjects were told that the study was part of an exploration of possible relations between ESP performance and mood, and that previous work had yielded interesting results. They were told that their work was to be done outside of class, preferably in periods of solitude. It was also explained that participation in the study was strictly voluntary, had no bearing on their evaluations in class, and carried no departmental experimental credit. The experimenter promised to return to class with the results a few weeks after completion of the testing.

Volunteers were each given a packet containing the following materials: (*1*) an instruction sheet; (*2*) a face sheet for recording their name (or some code number if they preferred), age, and a response of "true" or "false" to the statement, "I believe that ESP is possible under the conditions of this experiment"; (*3*) a copy of a standard 30-item version of the California F Scale (Adorno et al., 1950) titled "general opinion survey"; (*4*) four ESP record sheets, each containing five double columns of 24 cells headed C (for call) and T (target); (*5*) accompanying each ESP call sheet was the 54-item mood adjective check list. The instruction sheet explained that on four different occasions the student was to pick a quiet time alone and fill in the five ESP "call" columns with +'s and 0's, trying to guess whatever would later be considered the correct target, much as one might make guesses on a carnival wheel of fortune before it is spun.

Subjects were not told anything about the repeated-guessing or verbal-content features of the experiment. A date was set 10 days to 2 weeks following the solicitation when the experimenter would return and pick up all materials.

Use of moderator variables. When all data were collected, subjects whose scores were expected to be most predictive were selected. On the basis of previous research findings on the V_A scale, data from subjects in the experimenter's three classes were held for V_A scale analysis if the subjects were low-F sheep; and in the other classes the data of all low-F subjects were held. Subsequently, the results of all other subjects were analyzed as well and examined separately. Because of the untested nature of the H_A scale, the data of all low-F subjects were to be analyzed, because a presumption was made that their MACL reports were more valid (Carpenter, 1983a).

To keep conditions as constant as possible, I used the same cut-off point for F-Scale scores to separate "high" and "low" groups as I had in previous research (F-Scale scores of −31 or less defined the low quartile). Forty-six subjects were "low-F" in this sample.

Analysis of Scale Success

Two sets of analyses based on the MACL scale predictions were carried out. One examined the differences between mean variances and mean hitting rates between the predicted-large and predicted-small groups; these were tested with the statistics described next. The other was a repeated-guessing, majority-vote analysis.

Because the same target order was used repeatedly for all runs in this study, an analysis of ESP performance must take into account the group calling-patterns for each target in the run. This "stacking-effect" problem obtains whether one is analyzing variance or hitting rate (Burdick, 1983; Burdick & Kelly, 1977). Hitting rates, variances, and differences between groups in terms of hitting rate and variance were therefore tested using z statistics that take the stacking-effect problem into account. The formulae used are given in Appendix A.

Because only unidirectional predictions were made, given these scales' predictive purpose, one-tailed p values are used throughout this report for significance tests.

Repeated-Guessing Analyses

Variance. The repeated-guessing analysis using variance predictions was carried out as follows:

1. For the selected subjects (in terms of F-Scale and sheep-goat responses), data sets were categorized according to their predicted variances by the V_A scores accompanying them. Some data sets were categorized as "predicted-large" (high-quartile V_A score), some as "predicted-small" (low-quartile score), and some were given no prediction (midrange score).

2. Sets with variance predictions were scored for index targets. Thus, each run in the set had an index score that was above chance (greater than 6), below chance, or exactly at chance.

3. In predicted-large variance sets, calls to message items were tallied as given if the calls on index targets for that run were above chance. If index scoring for the run was below chance, message calls were reversed and added to the tally. If index scoring was at chance, message calls for that run were omitted from the analysis.

4. In predicted-small variance sets, the opposite was the case. If index scoring for the run was above chance, the message calls were reversed and tallied, whereas if index scoring was below chance, message calls were tallied as given. Runs with index scores at chance level were again omitted.[4]

5. Tallies from both predicted-large and predicted-small data sets were combined into a summary number of votes for each target alternative for each message trial.

[4] These procedures are based on the fact that, in runs of relatively large overall deviations from chance expectation (large variance), the modal direction of scoring will tend to be the same for any two subsets of the run (e.g., index and message subsets), whereas for runs of very small or zero deviation from chance expectation (small variance), they will tend to be in the opposite direction. (The exact probabilities for various cases are given in Carpenter, 1983b.)

TABLE 3
RESULTS OF EXTREME QUARTILE PREDICTIONS ON V_A AND H_A SCALES IN STUDY 1

Scale	Mean	*N* runs	*z*	*p*
V_A				
Predicted-large var.	6.40	200	.62	.27
Predicted-small var.	4.94	170	1.32	.08
		$z_d = 1.57, p = .06$		
H_A				
Predicted psi-hit	12.36	280	2.12	.017
Predicted psi-miss	11.76	210	1.15	.125
		$z_d = 2.14, p = .016$		

Hitting rate. Extreme-quartile H_A scale scores were treated more simply, without regard to index calls.

1. Subjects were selected for analysis using the moderator variable (F Scale). H_A scores, depending on the quartile in which they fell, were used to classify the sets into "predicted-high," "predicted-low," and sets without any expectation.

2. For predicted-high sets, all calls to message items were tallied as given.

3. For predicted-low sets, all message calls were reversed and tallied.

4. All calls from both groups were combined into an overall set of votes for each message trial.

Combined predictions. All calls thus rendered by both repeated-guessing analyses were then summed for each message-target. The sums for each target alternative (+ or 0) were considered "votes," and the target alternative having more "votes" was chosen as the best guess for that target.

Results

Overall hitting. There were 424 total sessions in this study (some subjects failed to do all four sessions) with 2,120 runs. Overall, there were 25,573 hits (133 more than MCE), a hitting rate of 50.26%. No statistical test was applied to this overall result, because there was no hypothesis and an analysis corrected for the stacking effect would be expensive in worker-hours.

Scale success: (a) Variance. The results are given in Table 3. For the V_A scale, the predicted-large group produced an average variance of 6.40, as compared to the chance expectation of 6, which is a nonsignificant trend in the predicted direction. The predicted-small group produced an average variance of 4.94, which is also in the predicted direction but not significant. The two mean variances were different from each other at a marginal level of significance, using a one-tailed test (p = .06).

(b) Hitting rate. The H_A scale was more strongly predictive. The predicted-large data produced an average run score of 12.36 (MCE is 12), which is significantly high (p < .02, one-tailed). The predicted-low group produced a mean run score of 11.76, which is in the predicted direction but not significant. The difference between the two means is significant (p = .017, one-tailed).[5]

Repeated-guessing analysis. For the results of the repeated-guessing analysis, see Table 4. The V_A scale yielded 1,700 correct votes and 1,612 incorrect, for a 51.3% rate of accuracy. The V_A scale alone produced six correct majorities, three incorrect, and three ties. The H_A scale yielded 3,061 correct votes and 2,759 incorrect, for a 52.6% rate of correctness. It alone produced 11 correct majorities and one incorrect. Both combined yielded 4,761 correct votes and 4,371 incorrect, for an overall accuracy rate of 52.1%. Thanks partly to luck in the case of a couple of majorities that were extremely close, this yielded 12 correct majorities and none incorrect. Translated into Morse-code form, this captured the letters of the target word *peace* with complete accuracy.

Study 2: The "Info" Series

This study was similar in most respects to Study 1. The main differences were in the test of a new pair of mood scales derived from a larger data base and in the use of a different experimenter soliciting subject participation.

Target Preparation

Verbal material, in Morse-code form, was to be used again to constitute message target items, which would be combined with random index targets; the index targets would be determined by random numbers selected by newspaper weather numbers as in the last study.

[5] As mentioned, the above results for variance and hitting predictions only represent the work of low-F subjects. Data of high-F subjects were analyzed as well and, as expected, only very slight trends were found, which were often contrary to prediction in direction. The same is true for Studies 2 and 3.

Table 4
Majority-Vote Analysis, Study 1

Target				V_A scale		H_A scale		Sum			
no.	Correct target			+	0	+	0	+	0	Decision[a]	
1	+	(.)	(P)	148	128	262	223	410	351	+	*(P)**
2	0	(−)		141	135	222	263	363	398	*0*	
3	0	(−)		138	138	235	250	373	388	0	
4	+	(.)		131	145	269	216	400	361	+	
5	+	(.)	(E)	138	138	243	242	381	380	+	*(E)**
6	+	(.)	(A)	138	138	257	228	395	366	+	*(A)**
7	0	(−)		127	149	253	232	380	381	0	
8	0	(−)	(C)	133	143	231	254	364	397	*0*	*(C)**
9	+	(.)		151	125	265	220	416	345	+	
10	0	(−)		135	141	225	260	360	401	0	
11	+	(.)		136	140	246	239	382	379	+	
12	+	(.)	(E)	152	124	260	225	412	349	+	*(E)**

Note. Vote outcome was 4,761 right, 4,371 wrong, % = 52.10, z = 4.06, p = .00002. Binary decisions were 12 right, 0 wrong, % = 100.00. Letter decisions were 5 right, 0 wrong, % = 100.00.

[a] Italics and asterisks denote correct decisions.

This time, the placement, as well as the content, of the index targets was determined by systematic random means. Because there were four message-target blocks, I determined that they should be surrounded by blocks of index targets of lengths determined by the random numbers immediately following the set of numbers that determined the content of the index targets. By this method, it was determined that the first message block would be preceded by 7 index targets, the second by 2, the third by 2, the fourth by 2, which left room for no index targets following the last message item.

This time, I chose as the verbal target the word *info*, the abbreviated, colloquial term for *information.* As in Study 1, the index targets were determined by random numbers from the RAND book, with an entry point selected by the high temperatures of the first cities listed in that morning's newspaper. As in Study 1, the target list was prepared prior to any data collection. I was again to serve as both originator and receiver in the information-exchange model. The target list, with associated Morse-code content, is given in Table 5. Because only 11 binary units are used in this target word, the number of index targets was increased to 13.

After constructing the target list, a copy was sealed in an envelope and mailed in a larger envelope to K. R. Rao at the Foundation for Research on the Nature of Man, with a request that he put it away and not open it until after my data analysis was completed. Because the envelope was mailed and received before any data were collected or analyzed, this provided an independent verification of the target material.

Derivation of New Mood Scales

Two new mood scales, one to predict variance and one to predict modal trend, were derived using forward-stepping multiple regression on pooled data from all studies conducted thus far. This larger data set included 1,171 cases, and will be referred to as Set B. The two new scales are called V_B and H_B. When the "A" scales were derived, the default inclusion criteria of the original SPSS stepwise regression program were used for item selection. After consultation with a statistician, it was realized that these criteria were overly liberal for the purpose of constructing a predictive instrument. Therefore, the criterion of $p = .10$ was set as the significance level that had to be met by an item's independent, new contribution to the regression equation before it could be included in the scale of items.

TABLE 5

TARGET ARRAYS FOR ORIGINATOR AND RECEIVER ROLES IN STUDY 2

	Originator's array					Receiver's array	
Target no.	Target letter	Morse code	Message target	Index target	Target list	Target type	Content
1				0	0	index	0
2				0	0	index	0
3				0	0	index	0
4				0	0	index	0
5				+	+	index	+
6				0	0	index	+
7				0	0	index	0
8	I	·	+		+	message	?
9		·	+		+	message	?
10				+	+	index	+
11				0	0	index	0

Table 5 continued

	Originator's array					Receiver's array	
Target no.	Target letter	Morse code	Message target	Index target	Target list	Target type	Content
12	N	–	0		0	message	?
13		·	+		+	message	?
14				0	0	index	0
15				+	+	index	+
16	F	·	+		+	message	?
17		·	+		+	message	?
18		–	0		0	message	?
19		·	+			message	?
20				0	0	index	0
21				+	+	index	+
22	O	–	0		0	message	?
23		–	0		0	message	?
24		–	0		0	message	?

Table 6
Scale Items and Scoring Weights of V_A and H_B Scales

V_B scale	H_B scale
adaptable (− 2)	annoyed (2)
amibitious (2)	assertive (− 2)
annoyed (− 2)	bashful (− 4)
close-mouthed (2)	detached (1)
decisive (− 2)	dreamy (1)
dizzy (2)	fearless (4)
downhearted (− 3)	masterful (3)
drifting (− 3)	tired (− 2)
dull (2)	
fearless (3)	
hesitant (2)	
lackadaisical (2)	
satisfied (− 1)	

Note. V_B scores of one or more give a large-variance prediction; scores of − 3 or less give a small-variance prediction. H_B scores of 2 or more give a hitting prediction; scores of − 1 or less give a missing prediction.

The new scales are given in Table 6. Given the larger data base, it was assumed that these scales should be more reliable than the two older ones. Extreme-quartile groups on the scales were to be used to evaluate results.

The B scales were of primary interest in this study, although both pairs of scales were to be examined for their predictive effectiveness. The same F-Scale cut-off point that was used in Study 1 was to be used again to define the low-F group. For the A scales, the criteria for subject inclusion in the analyses were the same as those used in Study 1. For the V_B and H_B scales, the data of all low-F subjects were to be used.

Procedure

Subjects. Subjects were again UNC undergraduate volunteers recruited from classes following a brief lecture on ESP research and a description of the procedure of the study. There were 121 subjects who took part; 42 of them were low-authoritarian (F-Scale score of − 31 or lower).

Instructions and materials. In all previous studies, I had acted in the role of the experimenter giving the introductory lecture, describing the study, and soliciting subjects. To see if the effects found thus far could be found when I (or anyone with the authority of a professor) was not thus involved, a female honors student majoring in psychology carried out this part of the experiment. She solicited volunteers from eight classes, following roughly the same procedure as I had used previously. No professor was named as sponsoring the experiment. I was the instructor for two classes. The six other classes were taught by colleagues in the Psychology Department. All other details of procedure given about Study 1 apply here as well.

Analysis

To test the discriminatory power of the four scales, the same *z*-score analyses that were done with the data of the last series were to be carried out again.

The repeated-guessing procedure was tested by taking votes from the calls of extreme-quartile data sets in the same way that was done in the last study. Votes were to be tallied for each target item for each B scale separately, and for both combined.

Results

Overall hitting. There were altogether 476 sessions in this study, or 2,380 runs. Total hits were 28,397, 163 less than MCE. This is an overall hitting rate of 49.71%. No statistical test was applied.

Scale success. The B scales were the major focus of interest in Study 2, and were expected to perform more effectively than the A scales.

(a) Variance. using the V_B scale, the predicted-large variance set of data produced a mean variance of 6.64, which was in the right direction but not significantly different from chance expectation. The predicted-small set yielded a mean score of 5.06, a trend of suggestive significance. The difference between the two is significant. See Table 7.

The predicted-large variance data set on the V_A scale produced an average variance of 6.03, almost exactly equal to chance expectation. The predicted-small set produced a mean variance slightly larger than that: 6.12, a nonsignificant reversal from the prediction. The difference between the two was not significant.

(b) Hitting. The H_B scale discriminated data sets that were significantly different from chance in both directions. The predicted-hitting runs yielded an average run-score of 12.38; the predicted-missing runs averaged 11.69. The difference between the two was also significant: p = .005. See Table 7.

TABLE 7
RESULTS OF EXTREME-QUARTILE PREDICTIONS FOR V_B AND H_B SCALES IN STUDY 2

Scale	Mean	*N* runs	*z*	*p*	z_d	*p*
V_B						
Predicted-large var.	6.64	225	0.90	ns	2.03	.02
Predicted-small var.	5.06	235	1.35	.09		
H_B						
Predicted psi-hit	12.38	165	1.90	.03	2.55	.005
Predicted psi-miss	11.69	235	1.82	.04		

The H_A scale did not discriminate hitting rate significantly. The set of data predicted to show psi-hitting yielded a mean score of 12.02, and the predicted-missing set averaged 11.78.

Repeated-guessing analysis using the V_B and H_B scales. The V_B scale produced 2,752 correct votes and 2,583 incorrect ones, a 51.58% rate of accuracy. Its votes produced nine correct majorities and two incorrect ones. The H_B scale produced 2,291 correct votes and 2,109 incorrect votes, for a 52.07% rate. Eight of its majorities were correct, and three were wrong. When the votes from both scales were pooled, 5,043 were correct and 4,692 were wrong, an accuracy rate of 51.80% (z = 3.55, p = .0002). Of the 11 majority decisions, 10 were correct, allowing accurate retrieval of the first three letters of the target word: *I, N,* and *F*. See Table 8.

STUDY 3: THE BROUGHTON STUDY

This study differed in several ways from the prior two. Briefly, these were: (*1*) The target was determined by an independent co-investigator. (*2*) I was not the instructor for any of the classes in which subjects were solicited. The solicitation was done by a graduate research assistant. (*3*) Scoring and analysis were automated and utilized a program written by Richard Broughton of the Foundation for Research on the Nature of Man; this program removed the problem of the stacking effect. (*4*) Some subjects were UNC undergraduates as before, but others were visitors and students at FRNM. (*5*) Two new pairs of scales were derived and tested (although the four scales tested in Study 2 were examined again).

TABLE 8
MAJORITY-VOTE ANALYSIS OF STUDY 2 USING B SCALES

Target	Correct	H_A scale		V_A scale		Sum B		
no.	target	1	0	1	0	1	0	Decision[a]
1	+ (I)	212	188	268	217	480	405	+ *(I)**
2	+	220	180	268	217	448	437	+
3	0 (N)	177	223	241	244	418	467	0 *(N)**
4	+	210	190	248	237	458	427	+
5	+ (F)	208	192	238	247	446	439	+ *(F)**
6	+	199	201	245	240	444	441	+
7	0	186	214	240	245	426	459	0
8	+	215	185	252	233	467	418	+
9	0 (0)	205	195	237	248	442	443	0 (G)
10	0	199	201	221	264	420	465	0
11	0	206	194	253	232	459	426	+

Note. Vote outcome was: 5,043 right, 4,692 wrong, % = 51.80, z = 3.55. Binary decisions were 10 right, 1 wrong, % = 90.91. Letter decisions were 3 right, 1 wrong, % = 75.00.

[a] Asterisks and italics denote correct decisions.

Table 8
Majority-Vote Analysis of Study 2 Using B Scales

Target no.	Correct target	H_A scale		V_A scale		Sum B		Decision[a]
		1	0	1	0	1	0	
1	+ (I)	212	188	268	217	480	405	+ *(I)**
2	+	220	180	268	217	448	437	+
3	0 (N)	177	223	241	244	418	467	0 *(N)**
4	+	210	190	248	237	458	427	+
5	+ (F)	208	192	238	247	446	439	+ *(F)**
6	+	199	201	245	240	444	441	+
7	0	186	214	240	245	426	459	0
8	+	215	185	252	233	467	418	+
9	0	205	195	237	248	442	443	0
10	0 (0)	199	201	221	264	420	465	0 (G)
11	0	206	194	253	232	459	426	+

Note. Vote outcome was: 5,043 right, 4,692 wrong, % = 51.80, z = 3.55. Binary decisions were 10 right, 1 wrong, % = 90.91. Letter decisions were 3 right, 1 wrong, % = 75.00.

[a] Asterisks and italics denote correct decisions.

Two scales were shortened versions of the V_B and H_B scales, and two were new scales generated by the larger data base provided by including data from Series 2. In this larger data base were also included responses to a sheep-goat question, which had been collected previously but had never been included as a potential predictor before.

Target Preparation

In this study, I did not play both the originator and receiver roles. Instead, an actual transmission of information from one investigator to another was attempted. The target was determined independently by Broughton and was known only to him until all data had been collected and analyzed. I did not know what sort of information the target was to contain, only that it could be encoded into 12 binary units. Broughton decided to select randomly an octal number between 0000 and 7,777 (equivalent to 0 to 4,095, the decimal numbers), which had the advantage of being easily represented by the 12 binary units. Each octal digit was to be converted to three binary digits (ranging from 000 to 111). Then, the binary "0" was set as equivalent to the target symbol "0," and the binary "1" was set as equivalent to the symbol "+." Thus, a series of 12 +'s and 0's was to be determined, arranged in four sets of three digits each.

Using FRNM's standard computer-generated method for picking a random-number target (based on a proprietary adaptation of the Fortran IV rand function), Broughton drew the octal number 0625 (equivalent to decimal 405). Table 9 gives the numerical target, the target symbols aimed for by subjects, and the coding steps in between.

Target Shuffling and Scoring Procedures

These 12 message-target symbols were not to be arrayed in discrete blocks and interspersed with blocks of index targets in a static array as before. Instead, Broughton's program was able to "shuffle" the 12 message-targets for each run of guesses, and lay them out in a random mix with 12 index targets, newly picked for that particular run (using the same Fortran IV rand function). Thus, a subset of 12 message-items was mixed randomly with a new subset of 12 randomly selected index-targets for each run, and each target was scored automatically against its corresponding ESP call. The identity of each individual message-target was maintained across runs, permitting an accumulation of votes in the repeated-guessing procedure.

TABLE 9
NUMERICAL TARGET, MESSAGE-TARGET SYMBOLS, AND ENCODING STEPS FOR STUDY 3

Target no.	Message targets	Binary equivalents	Octal digits	Decimal equivalent
1	0	0	0	0405
2	0	0		
3	0	0		
4	+	1	6	
5	+	1		
6	0	0		
7	0	0	2	
8	+	1		
9	0	0		
10	+	1	5	
11	0	0		
12	+	1		

This individualized shuffling of targets for each run eliminates the stacking-effect problem, which made statistical analysis of the previous studies so arduous. The program also carries out index-target scoring, MACL scoring, categorization of sessions as to hitting or variance predictions using MACL-scale scores, and generation of votes for the repeated-guessing analyses, all in a rapid, automatic manner. It also carries out all of this in a way that leaves the receiver blind to the content of the message-targets while guesses are entered into the analysis and votes are generated.

Mood Scales: Old and New

The H_B and V_B scales were to be tested again to see if their predictiveness could be cross-validated in this new sample. Predictions using the A scales were also checked, although they were not expected to perform as strongly.

Four new scales were of major interest in this study. The first two were shortened versions of the H_B and V_B scales. Further consultation with a statistician led to the decision to make the criterion for item selection still more stringent than that which had been used for the B scales. Whereas the latter were made up of items whose independent contributions were significant at $p < .1$, this still probably led to the inclusion of unreliable items. It was decided to reanalyze the data that

TABLE 10
ITEMS AND SCORING WEIGHTS OF V_{BS} AND H_{BS} SCALES

V_{BS} scale	H_{BS} scale
adaptable (− 2)	assertive (− 2)
close-mouthed (2)	bashful (− 4)
downhearted (− 3)	dreamy (1)
drifting (− 2)	fearless (3)
dull (2)	masterful (3)
lackadaisical (2)	

Note. V_{BS} scores of 1 or more predict large variance; scores of − 2 or less predict small variance. H_{BS} scores of 1 or more predict hitting; scores of − 1 or less predict missing.

had produced the B scales, using the criterion of $p < .05$ for item inclusion unless too few items were selected. The scale would not be usable unless there were enough items so that total scores would produce a distribution allowing a reasonable extreme-quartile split. Otherwise it would not be possible to select groups of cases about which predictions could be made. For this reason, a total number of five or more items was set as a minimum, with the additional requirement that at least two items of each sign be included. If the criterion $p < .05$ did not produce enough items, then a reanalysis with $p <$ .075 would be carried out, and even a further analysis at the original $p < .1$ if necessary. The items making up the shortened scales (V_{BS} and H_{BS}) are given in Table 10. All items of the V_{BS} scale were independently significant at the .05 level, but the H_{BS} scale required a second pass at the .075 level of significance to attain a second, negatively weighted item ("bashful").

It was also decided to take advantage of the larger data base available since the completion of Study 2 for the generation of a new pair of scales that should be still more reliable. Before describing these analyses, however, another matter must be mentioned: the sheep-goat question.

All subjects in Studies 1 and 2, and most of those tested previously (Carpenter, 1983a, 1983b) were asked to respond yes or no to Schmeidler's sheep-goat question: "Do you believe that ESP is possible under the conditions of this test?" Responses to this question were initially used as a moderator variable in the predictiveness of the V_A scale, but these results were inconsistent, and this analysis was discontinued in later studies. Retrospective examination, however, showed that for low-authoritarian subjects (those whose results have been found to be predictable in all this work) the sheep-goat question had the discrim-

TABLE 11
ITEMS AND SCORING WEIGHTS OF V_C AND H_C SCALES

V_C scale	H_C scale
annoyed (− 1)	bashful (− 4)
close-mouthed (1)	dreamy (1)
decisive (− 1)	fearless (3)
drifting (− 2)	masterful (2)
dull (1)	unsure (1)
fearless (2)	sheep/goat (1)
genial (− 1)	

Note. V_C scores of 1 or more predict large variance; scores of − 2 or less predict small variance. H_C scores of 2 or more predict hitting; scores of 0 or less predict missing.

inating effect that Schmeidler found it to have and many others have confirmed (Palmer, 1977; Schmeidler & McConnell, 1958). Sheep tended to score above chance, and goats below. These analyses have not been done in a way that can give exact statistics because of the stacking-effect problem. But standard statistics made it clear that the expected discrimination was tending to obtain. A descriptive analysis, not corrected for the stacking effect, of the low-authoritarian data of all research prior to Study 3 showed that sheep produced in 293 sessions an average session score of 60.85 (where MCE is 60.00) and that goats in 208 sessions produced an average of 59.44 (t = 2.83). Although not corrected for stacking, it is based on 10 target orders, not a single one, and is large enough to appear to be real. Also, it is as strong as any of the discriminations produced by mood items. Because of this, responses to the sheep-goat question (scored as + 1/ − 1) were included in the last regression analyses as another potential predictor.

The data from Study 2 were included with all the previous data, making up Data Set C with a total of 1,647 cases. New scales were generated with forward-stepping multiple regression, using the more stringent criteria that were used with the shortened B scales. These analyses generated the scales about which one would have the strongest expectation for success: the V_C and H_C scales. They are given in Table 11. The sheep-goat variable did emerge as an item on the scale to predict hitting rate. All items on both scales were independently significant at $p < .05$ in the regression analyses.

Procedure

Subjects. It was decided to collect data from at least 150 subjects in this study, in hopes of securing enough low-authoritarian cases to permit a highly reliable overall result in the event that the scales being tested worked effectively. There were 124 UNC undergraduates and 28 visitors and students at the Foundation for Research on the Nature of Man who took part.

Of the 152 subjects participating, only 34 were categorized as low-authoritarian using the cut-off point that had become standard (among university students, there appears to have been a sizable drift toward more authoritarian attitudes over the period of years spanned by this research). This was a disappointing fact, because it meant that this study, despite the large number of subjects, was actually being conducted with fewer usable participants than either the "Peace" series (47) or the "Info" series (42). Given this, less reliable results on the majority-vote analyses for a given size of effect could be expected.

Instructions and material. The solicitation of subjects was not done by me in this study, in order to test again the robustness of the mood scale-ESP relations when I was not in a position to exert any personal influence on the volunteers. A female clinical psychology graduate student who was working as my research assistant solicited participants from eight undergraduate psychology courses at UNC. She gave a brief introduction to ESP research and a description of the particular experiment, covering the same points as had been done in all previous solicitations. James Perlstrom, a research fellow at FRNM, solicited participants there.

Analysis

Subjects were included in the analyses involving the A and B scales according to the same criteria used in the last study. On the new B (shortened) and C scales, data were kept for analysis if the subjects were low-F (-31 or lower) as before. ESP guesses and mood-scale responses were entered into a computer file and cross-checked. As just described, all data were scored automatically and blindly, with mood-scale scores, index-target scores, total run scores and session scores, and votes for the repeated-guessing analyses being printed out. The effectiveness of the scales was to be tested by pooling data from extreme quartiles. For the three scales predicting scoring direction, single-mean t tests were to be done on the data of each extreme-quartile group, and the difference between the means of the "predicted-psi-hitting" and "predicted-psi-missing" groups for each scale would be tested by a t test. For the three scales predicting variance, a nonparametric test was chosen because of the skewed distribution such scores have. The difference between the data from the predicted-large-variance and predicted-small-variance groups for each scale would be tested using the Kruskal-Wallis Test. Because the z analysis is not used here, the statistics are calculated on session scores, not run scores as in Studies 1 and 2. Hence, the means that are tabled in the Results section are about five times larger than corresponding ones reported for the earlier studies.

The repeated-guessing procedure was tested by selecting calls from the appropriate data sets, as determined by extreme-quartile mood scores, and rendered and tallied automatically by computer, using the same procedural rules that were used in the previous studies.

Results

Overall hitting rate. There were a total of 600 sessions in this study, or 3,000 runs. The total number of hits was 35,883, 117 fewer than MCE. This is a hitting rate of 49.84%. As the stacking effect does not obtain in this study, this can be tested statistically, but it is far from significant: $t = 0.90$.

Scale success. The major interest in this study was in the scales expected to be most effective: the shortened B scales (BS scales) and the C scales. Results with the A scales and long B scales are mentioned as well, for the sake of comparison.

(a) Variance. As expected, the V_A scale was again ineffective in discriminating scoring trends in this sample. It produced mean variances that were slightly contrary to expectation: 25.23 for the predicted-large set, and 25.71 for the predicted-small.

The V_B scale also did poorly. The average variance for the predicted-large set was again slightly smaller than chance expectation, and the predicted-small mean was slightly larger. See Table 12. This reversal was not significant.

Perhaps it can be said that the V_{BS} scale was slightly more effective than the V_B scale, as expected, in that it produced a trend that was in the expected direction, but it was very slight and quite insignificant. See Table 13.

The V_C scale produced a trend that was also nonsignificant, but it was at least larger than that given by the V_{BS} scale. The means of both extreme-quartile groups show trends that are in the expected directions. See Table 14.

In summary, none of the scales designed to predict variance were able to do so in this sample. The scales designed with more stringent inclusion criteria and a larger data base did show trends that were more nearly in line with expectation.

TABLE 12

RESULTS OF EXTREME-QUARTILE PREDICTIONS FOR SET B MOOD SCALES IN STUDY 3

Scale and data set	Mean	*N*	*SD*
V_B			
Predicted-large var.	26.34	32	
Predicted-small var.	30.15	27	
	$U = 496, p = .33$		
H_B			
Predicted-high score	61.16	19	5.00
Predicted-low score	59.11	35	4.89
	$t[52] = 1.46, p = .076$		

Note. Test statistic for variance is Mann-Whitney *U*; the statistic for hits is Student's *t* test.

TABLE 13

RESULTS OF EXTREME-QUARTILE PREDICTIONS FOR SHORTENED B SCALES IN STUDY 3

Scale and data set	Mean	*N*	*SD*
V_{BS}			
Predicted-large var.	28.33	15	
Predicted-small var.	27.09	47	
	$U = 375, p = .71$		
H_{BS}			
Predicted-high score	61.70	30	4.64
Predicted-low score	58.71	21	4.29
	$t[49] = 2.33, p = .01$		

TABLE 14

RESULTS OF EXTREME-QUARTILE PREDICTIONS FOR C SCALES IN STUDY 3

Scale and data set	Mean	*N*	*SD*
V_C			
Predicted-large var.	32.20	20	
Predicted-small var.	26.41	22	
	$U = 186, p = .19$		
H_C			
Predicted-high score	62.13	38	3.99
Predicted-low score	58.77	26	4.56
	$t[62] = 3.12, p = .002$		

Hitting rate. The H_A scale yielded trends that were in the right directions but were not significant. The observed mean for predicted-high was 61.12, and for predicted-low was 59.94. The difference between them was not significant.

The H_B scale was marginally successful. Both groups produced means that differed in the expected directions from chance expectation, but neither difference was significant. See Table 12. The difference between the two means yielded $t = 1.46$, $p = .076$, one-tailed.

The H_{BS} scale was more effective than the H_B scale was, as expected. The mean session score of 61.70 for the predicted-high group was significantly large, $t\ (30) = 2.00$ ($p < .05$), whereas the mean of 58.71 for the predicted-small group was nonsignificantly small. See Table 13. The difference between them was significant: $t\ (49) = 2.33$, $p = .01$.

The H_C scale did well, as expected given the relatively large data base that generated it. The predicted-high group scored significantly highly, $t\,(38) = 3.25$ ($p < .001$), and the predicted-small group showed a nonsignificant trend toward psi-missing. See Table 14. The difference between them was the strongest for any scale discrimination made in the data of this study: $t\ (62) = 3.12$, $p = .002$.

The scales to predict direction of scoring were generally successful, except for H_A as expected. The scales that were expected to be more effective because of a larger data base and more stringent inclusion criteria were more effective.

Repeated-guessing analysis using the B scales. The results using the B scales are given in Table 15. The H_B scale produced 1,652 correct votes and 1,576 incorrect ones, a rate of 51.18%. Seven majorities were right, five wrong. The V_B scale, as would be expected given its failure to discriminate variance trends, produced null results in the repeated-guessing analysis: 1,345 votes were right, 1,343 wrong (50.04%). Seven majorities were right, and five were wrong.

The tallied votes from both scales yielded 2,997 correct votes and 2,919 incorrect votes, a hitting rate of 50.66% ($z = 1.00$). Seven majorities were correct, and five were incorrect. Two octal digits were captured correctly, and two were wrong.

Repeated-guessing analyses using the B (shortened) and C scales. The analyses using the shortened B scales (BS scales) are given in Table 16. The H_{BS} scale yielded a 51.08% hitting rate in votes, with 1,557 correct and 1,491 incorrect. Nine majorities were right and three were wrong. The V_{BS} scale gave a hitting rate for votes of 51.51%, with 1,434 correct and 1,350 incorrect. Seven majorities were right, three were wrong, and two were ties. Pooled votes yielded a 51.29% rate

TABLE 15
MAJORITY-VOTE ANALYSIS OF STUDY 3 USING B SCALES

Target no.	Correct target	H_B scale		V_B scale		Sum B		Decision[a]
		1	0	1	0	1	0	
1	0	140	129	108	116	248	245	1
2	0 (0)	133	136	110	114	243	250	*0* (4)
3	0	136	133	110	114	246	247	*0*
4	1	150	119	102	122	252	241	*1*
5	1 (6)	130	139	132	92	262	231	*1* *(6)**
6	0	120	149	110	114	230	263	*0*
7	0	130	139	131	93	261	232	1
8	1 (2)	131	138	111	113	242	251	0 (5)
9	0	139	130	122	102	261	232	1
10	1	136	133	116	108	252	241	*1*
11	0 (5)	126	143	114	110	240	253	*0* *(5)**
12	1	146	123	121	103	267	226	*1*

Note. Vote outcome was 2,997 right, 2,919 wrong, % = 50.66, z = 1.00. Binary decisions were 7 right, 5 wrong, % = 58.33. Octal decisions were 2 right, 2 wrong, % = 50.00.

[a] Italics and asterisks denote correct decisions.

TABLE 16
MAJORITY-VOTE ANALYSIS OF STUDY 3 USING SHORTENED B SCALES

Target no.	Correct target	H_{BS} scale		V_{BS} scale		Sum BS		Decision[a]
		1	0	1	0	1	0	
1	0	137	117	122	110	259	227	1
2	0 (0)	116	138	113	119	229	257	*0* (4)
3	0	125	129	113	119	238	248	*0*
4	1	130	124	117	115	247	239	*1*
5	1 (6)	132	122	134	98	266	220	*1 (6)**
6	0	121	133	107	125	228	258	*0*
7	0	113	141	120	112	233	253	*0*
8	1 (2)	118	136	125	107	243	243	None (None)
9	0	137	117	117	115	254	232	1
10	1	130	124	116	116	246	240	*1*
11	0 (5)	121	133	116	116	237	249	*0 (5)**
12	1	139	115	126	106	265	221	*1*

Note. Vote outcome was 2,991 right, 2,841 wrong, % = 51.29, z = 1.95, p = .026. Binary decisions were 9 right, 2 wrong, % = 81.82. Octal decisions were 2 right, 1 wrong, % = 66.67.

[a] Italics and asterisks denote correct decisions.

of correctness, with 2,991 correct and 2,841 incorrect. The sign test yields a $z = 1.95$, $p = .026$. Nine majorities were correct, two were incorrect, and one was a tie. Two of the four octal digits were captured correctly ($p = .016$). One octal decision was wrong, and one (because of the tie) was indeterminate.

The repeated-guessing analysis using the C scales was the most successful of this study in terms of hitting rate for votes, but not successful in octal decisions. See Table 17. The H_C scale produced 1,990 correct votes and 1,850 incorrect votes, a 51.82% rate. Nine H_C majorities were correct, and three were incorrect.

The V_C scale yielded a 51.43% hitting rate for votes, with 938 correct and 886 incorrect. Eight majorities were correct, four were wrong. When the votes from the C scales are pooled, they yield a correctness rate of 51.69% ($z = 2.54$, $p < .005$). In spite of the relatively high hitting rate, votes happened to be distributed such that only seven majorities were correct, four were incorrect, and one was indeterminate. These seven binary hits were themselves distributed such that none of the four octal digits was retrieved correctly.

Discussion

Original Scale

These studies were preceded by several others that were aimed at validating the V_A scale (in an earlier, slightly erroneous form) and at using variance predictions in repeated-guessing analyses. Viewed from the perspective of the present results, the V_A scale appears to have been useful as a starting point for developing research, but is not itself a powerful predictor in new samples. The H_A scale, generated from the same early sample, has similarly failed to be cross-validated in the majority of these studies. This is not surprising. The original sample was too small to expect a reliable regression solution for either criterion variable. The use of larger samples has proven to be helpful, and still larger samples would probably allow the development of more reliable instruments.

In particular, a larger sample of data from low-authoritarian subjects would be useful. The scales studied here have all been found to be useful only with low-authoritarian subjects. At the same time, to have a sufficiently large sample to permit stepwise regression, I included the data from all subjects (including high-F and others who did not take the F Scale) in each step of scale generation. Presumably

Table 17
Majority-Vote Analysis of Study 3 using C Scales

Target no.	Correct target	H_C scale		V_C scale		Sum C		Decision[a]
		1	0	1	0	1	0	
1	0	165	155	72	80	237	235	1
2	0 (0)	148	172	71	81	219	253	*0* (?)
3	0	159	161	77	75	236	236	None
4	1	175	145	72	80	247	225	*1*
5	1 (6)	162	158	90	62	252	220	*1* (7)
6	0	167	153	78	74	245	227	1
7	0	129	191	75	77	204	268	*0*
8	1 (2)	163	157	83	69	246	226	*1* (3)
9	0	155	165	83	69	238	234	1
10	1	177	143	82	70	259	213	*1*
11	0 (5)	167	153	74	78	241	231	1 (7)
12	1	163	157	77	75	240	232	*1*

Note. Vote outcome was 2,928 correct, 2,736 incorrect, % = 51.69, z = 2.54, p = .005. Binary decisions were 7 right, 4 wrong, % = 63.64. Octal decisions were 0 right, 3 wrong, % = 0.00.

[a] Italics denote correct decisions.

this has added "noise" to the solutions reached. A large sample of low-F subjects should permit more reliable scales.

Prediction of Variance and Hitting Trend

This research was preceded by several other studies that had the aim of developing other ways of predicting variance by subjects' mood reports (Carpenter, 1968, 1969). The original hypothesis guiding the research was that energetic versus dull moods might be predictive of variance, and mood items that had been found by Nowlis (1961, 1965) to discriminate pharmacologically elated and sedated states were combined with other "filler" items to comprise the MACL. These early approaches were not successfully cross-validated in later samples. The prediction of hitting trend was not attempted until the inception of the current work. The current series of studies, which used presumably more reliable scales on relatively large samples, suggests that variance, in spite of its being the main focus of most of the research, is actually not as successfully predicted by these means in this situation as hitting rate is. Only one variance prediction in these studies was significant at $p < .05$; it was that for the V_B scale in Study 2. The scales that predicted hitting succeeded more consistently. Future work using mood reports to predict both parameters is worth pursuing, but prediction of hitting rate may be expected to be more successful.

California F Scale

At this point it is worth asking: why are the ESP data of low-F subjects more discriminable with these mood items? One answer could be that the high-F subjects, for some reason, do not express ESP ability in this procedure. However, there is no evidence to support this idea. The scoring rate of high-F subjects did not differ significantly from that of their low-F counterparts in any of these studies. Another possibility is that high-F subjects are expressing ESP ability but that the mood states that are associated with different patterns of scoring are not the same for them as for low-F subjects. This idea was tested by generating an experimental pair of scales (one for hitting rate and one for variance) from the data of all the high-F subjects of Data Set C. These scales were used as predictors for the high-F data of Study 3. The results were both slight reversals from expectation, and far from significant. If mood states are reliably related to these parameters of ESP performance for high-F subjects, that fact could not be captured by these mood ratings.

The possibility appearing to be the most likely is that while high-F subjects may be exhibiting some sort of ESP effects in their data, their mood reports are invalid and hence could not be expected to predict performance. This conclusion is consistent with the findings of Thayer (1971).

The same lack of introspective validity may also affect the act of self-report required to respond to the sheep-goat question. In these data, sheep scored more highly than goats, as predicted, only among low-F subjects. This is discussed more fully in the next section.

Sheep-Goat Effect

Responses to the sheep-goat question were collected originally in this research because it was thought that the variable might be a useful moderator in mood-scale prediction of variance. It was only with the advent of these current studies that it was examined as a predictor of hitting rate in its own right. The analysis of this effect is not an exact one, as it has not been done in a way that can correct for the stacking effects in the data. Counting the current studies, 12 have been carried out with a single target order per study, and one (the last) used reshuffled orders for each run, eliminating any stacking effect for it. If all data from low-F subjects are pooled and analyzed as if the run orders were independently shuffled, sheep scored in 407 sessions an average score of 60.69, and goats in 226 sessions an average of 59.37, with $t = 2.73$. This is associated with $p = .003$. This p value is inexact, but it would seem to be small enough to make it likely that a correct analysis would still show a significant difference. A correct analysis would be preferable. But so much hand-keying of data at the guess level would be needed that to date it has not seemed to merit the cost required.

It is of interest for future research on the sheep-goat effect that it is found here only for low-authoritarian subjects. For the high-F data, only a slight trend in the predicted direction is noted (for sheep, who had 745 sessions, mean = 59.94; for goats, who had 398 sessions, mean = 59.70; $t = 0.70$). In trying to understand why this might be so, the same reasoning that was applied with mood reports may be useful. A response to the sheep-goat question requires the subject to look within and give an accurate statement of his or her actual belief, perhaps along with some assessment of past personal experiences that might have been supportive of an ESP hypothesis. Research has found that the sheep-goat response is generally modifiable by instructions

(Layton & Turnbull, 1975) and by the apparent attitudes of the experimenter (Crandall, 1985). The response of the high-F subject may be so contaminated by external considerations and repression as to be an inaccurate predictor of his or her own performance. If this is so, one would expect it to be a matter of degree. That is, among the "high-authoritarian" subjects, the responses of the relatively less authoritarian should be more valid (and predictive) than those of the more authoritarian. To test this, the "high-F" data were split at that group's median F score, and the halves were examined separately. For the "lower" high-F group, a stronger trend was observed: $t = 1.51$ (p would be .07 if targets were independent). For the "higher" group, there was a slight reversal of expectation: $t = -0.52$.

Other variables have been found to moderate the sheep-goat effect. The most pertinent one here is Rorschach-rated adjustment level (Schmeidler, 1960). Better adjusted subjects demonstrated the sheep-goat effect, less well-adjusted ones did not. F-Scale scores have been found to correlate positively with neuroticism, as measured by the Taylor Manifest Anxiety Scale (Davids, 1955; Singer & Feschbach, 1959), although the relationships were not large.

One might speculate that authoritarian tendency would be expected to moderate the sheep-goat effect *in this testing situation* (solitary self-testing) but might not have the same effect in individual testing by an experimenter, where the desire to please the experimenter could motivate good performance by the high-authoritarian sheep.

The sheep-goat response was collected from each subject before any guessing was done, relatively close in time to the first session, further from each successive one. Because of this, it seemed likely that the question should be most effective as a predictor of performance with the session nearer in time. With no feedback forthcoming and other experiences intervening, the subject's attitude could well change as the work went on, but there were no restatements of the question to capture that. To examine that possibility, an analysis by session of the effect was done. The results are presented in Figure 3.

Comparable analyses for the two subsets of the "high-F" group are also presented for comparison in Figure 3. It can be seen that the sheep-goat effect for low-F subjects was, in fact, confined mostly to the initial session ($t = 3.68$) with a trend in the second session ($t = 1.37$), and virtually no effects at all in the later sessions. Interestingly, the effect for the "lower high-F" group was also strongest for the initial session ($t = 1.93$). These analyses, too, are only descriptive and should be done in a way that eliminates the stacking

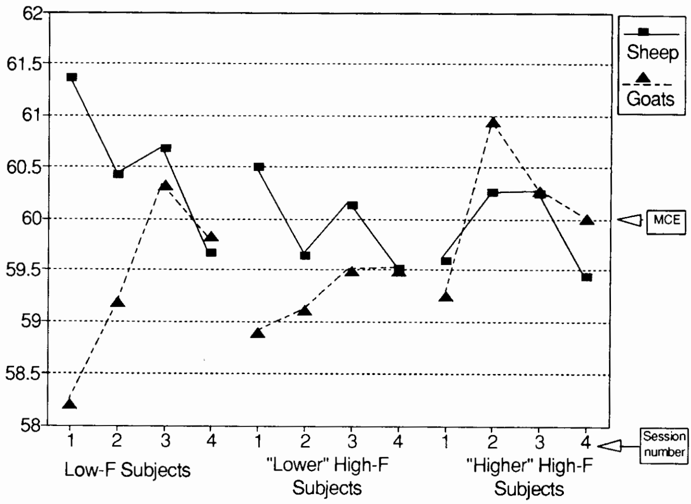

Figure 3. Average scores of sheep and goats on successive sessions, by F-Scale group.

effect. But they suggest that the sheep-goat response may itself be altered by repeated testing and the passage of time, and may be expected to be most predictive when testing shortly follows it.

Utility of Self-Testing

Although individual testing by an experimenter is expensive in experimenter-hours, testing in groups (although faster) has not tended to produce results of sufficient reliability (Honorton et al., 1990; Palmer, 1977).

This difference in reliability between group and individual testing, if generally true, is itself a potentially important phenomenon that has not been explicitly studied and is not really understood. One can speculate, however, that the distractions inherent in group testing may preclude the states of mind most conducive to the expression of ESP, and the anonymity of functioning as part of a mass of people may make individual difference variables (traits, moods, etc.) unlikely to be expressed in performance.

A useful alternative for many research problems may be individual self-testing, whether forced-choice or free-response. If target material is securely unavailable to subjects, and if the task is sufficiently pleasant and interesting, it may combine some of the virtues of the individual testing situation with the economy of group administration of materials and instructions. Self-testing, guided by appropriate instructions, can provide a quiet atmosphere perhaps conducive to more facilitative states, and an individualized situation in which individual-difference variables would more likely be expressed. Surprisingly, it has not been used very much in recent years, although its virtues and feasibility were pointed out many years ago by Fisk (1951). In addition, self-testing by volunteers also provides implicitly for self-selection, perhaps assuring a higher level of helpful motivation.

What Moods Affect Performance?

The precursors of this research were inspired by the observation that variance seemed to vary as a function of the subject's mood of interest and energy compared with disinterest and withdrawal (Carpenter, 1966; Rogers, 1966; Rogers & Carpenter, 1966; Whittlesey, 1960) and attempted to measure those states using an MACL (Carpenter, 1968, 1969). When this intuitively appealing idea failed to cross-replicate, theory was forsworn and blind empiricism (via multiple regression) was turned to for guidance. Since then, mood items have been selected and studied solely on the basis of their empirical relation to the criteria. Still, it is the actual state of mind of the subject while being tested that is of real interest, and the moderation of effects by authoritarianism adds support to the assumption that it is the real internal state, and not merely the report, that is important in influencing performance. Enough work has been done that it is of interest now to turn to the content of the items that have been found to be important and try to understand what states of mind, or moods, they represent.

To assist in this, the 54 items of the MACL were factor analyzed, with varimax rotation and a factor selection criterion of eigenvalue equal to 1 or more. Using the data of all subjects who did not have high F scores, a 13-factor solution was obtained. The items constituting the factors (weights of .40 or greater) and a nominated title for each factor are given in Table 18.

To get the best possible estimate of which mood items discriminate hitting rate and variance, I pooled data from all subjects and subjected them to multiple-regression analyses.

Items predicting hitting rate. Items predicting psi-hitting were, in descending order of significance: *forceful, masterful, drifting,* and *amiable.* Items predicting psi-missing were *bashful* and *adaptable.* When the factor analysis is consulted, it is clear that hitting is associated with

TABLE 18
ITEMS AND LOADINGS OF 13 MOOD ADJECTIVE CHECK LIST FACTORS

Factor 1	*Factor 2*	*Factor 3*
(Agreeable and outgoing)	*(Indifferent to task)*	*(Strong-willed)*
amiable (.74)	indifferent (.70)	forceful (.65)
genial (.71)	disinterested (.65)	decisive (.63)
friendly (.68)	lazy (.51)	masterful (.62)
cooperative (.58)	lackadaisical (.50)	ambitious (.61)
adaptable (.57)		assertive (.58)
cheerful (.56)		enterprising (.56)
satisfied (.52)		industrious (.51)
warm-hearted (.49)		task-involved (.43)
sociable (.45)		business-like (.41)
		fearless (.40)
Factor 4	*Factor 5*	*Factor 6*
(Exultant)	*(Uncertain)*	*(Altered)*
exultant (.79)	unsure (.73)	drifting (.64)
ecstatic (.75)	hesitant (.52)	dreamy (.61)
		languid (.53)
		detached (.42)
Factor 7	*Factor 8*	*Factor 9*
(Socially Anxious)	*(Annoyed)*	*(Inebriated)*
bashful (.77)	annoyed (.68)	intoxicated (.71)
aloof (.60)	critical (.68)	light-headed (.60)
		dizzy (.52)
Factor 10	*Factor 11*	*Factor 12*
(Garrulous)	*(Glum)*	*(Careful)*
talkative (.74)	close-mouthed (.66)	careful (.67)
sociable (.58)	withdrawn (.62)	business-like (.48)
witty (.46)	down-hearted (.55)	(not) fearless (– .41)
	dull (.46)	
	quiet (.40)	
Factor 13		
(Sleepy)		
drowsy (.65)		
tired (.64)		
sluggish (.59)		
dizzy (.45)		
dull (.42)		

Factor 3 (strong-willed), Factor 6 (altered), and Factor 1 (outgoing). Examination of Pearson *r*'s shows that the other items of Factors 3 and 6 strongly tend to correlate positively with hitting rate, whereas the other items of Factor 1 show mixed and generally weak relationships. Psi-missing is associated with Factor 7 (socially anxious) and also with an item of Factor 1 (outgoing). Correlation with the other item on Factor 7 is also negative. Thus, it seems safe to say that hitting rate is facilitated by being in a mood of strong confidence and determination, and by being in a relaxed, somewhat altered state; whereas hitting is diminished by a state of anxiety.[6] The meaning of the relationships with *amiable* and *adaptable* cannot be clarified by their common factor membership (agreeable-outgoing) because of contrary relationships with other items of the factor.

There is a fairly large (mostly old) literature on psi-facilitative states of mind, but none of it was consulted in constructing this MACL. If it had been, many other propitious items might have been included. Even so, there are interesting parallels to these findings. White (1964), in her examination of the older literature on subjective methods of response, found that successful percipients stressed the importance of a state of profound relaxation combined with an intense wish to succeed. Any sort of fretful, over-conscious laboring was said to be detrimental. Murphy, making reference to similar material (1962) and to research on comparable processes involved in creativity (1966), emphasized the importance of deep relaxation and strong motivation, combined with dissociation and an openness to one's own "associative network." Rhine (1934) noted early in his work that states of "detachment," "abstraction," and "relaxation" were helpful to his subjects, as were confidence and an intense will to succeed. Honorton (1977) argues persuasively that considerable current research demonstrates the importance of "altered" states of consciousness, in which the subject is removed from ongoing, external sensory experience. Palmer (1977) has summarized a relatively large body of research that shows that anxiety is detrimental to psi-hitting.

The relationships found here are not, for the most part, new; but their similarity to previous findings gives added confidence in their validity. Combined with the previously described sheep-goat results, the finding regarding hitting rate can be summed up: hitting is fa-

[6] It might be argued that the items "bashful" and "aloof" provide a very inadequate measure of anxiety. However, the MACL was constructed without any intention of assessing anxiety, and the occurrence of a factor comprised of these two make it seem reasonable to assume that they provide as good a measure of anxiety (in particular, "social anxiety") as this collection of items can provide.

cilitated by a commitment to a belief that ESP is possible under the test conditions (a "yes" to the sheep-goat question), by an enthusiastic determination to succeed (*forceful, masterful,* and other items of Factor 3), and by a state of deep relaxation and altered consciousness (*drifting* and other items of Factor 6); and it is impaired by a commitment to disbelief and by a state of social anxiety (*bashful).*

Items predicting run-score variance. An analysis similar to that for hitting was carried out for the criterion of variance. In descending order of significance, items predicting large variance were *fearless, dull,* and *carefree.* Items predicting small variance were *drifting, annoyed,* and *task-involved.* Correlations with other item-members of the factors with which these items are associated are not as consistent in direction as they were with hitting tendency. No factor had a generally positive relationship, but Factor 6 (altered) and Factor 12 (careful), on which *carefree* was loaded negatively and *task-involved* was loaded positively, were generally negative.

There is no older, subjective literature to consult on states associated with different levels of variance, and relatively little contemporary, experimental research. Several studies (Carpenter, 1968; Carpenter & Carpenter, 1967; Rogers & Carpenter, 1966) have reported declines of variance over sustained periods of guessing without feedback. Whittlesey (1960) found very constricted variance in the data of his subjects who had been given LSD-25, and who found the test to be "ridiculously petty." Rogers (1966, 1967) found large variance on test days when his subjects felt enthusiastic and positive about taking the test and small variance when they felt unenthusiastic, disinterested, and lacking in confidence. Stanford (1966) found large variance when his subjects were encouraged to respond spontaneously and small variance when they were asked to keep track of their calls and balance their frequencies. He also reported another study (1967) showing that subjects produced larger variance in sessions within which they also showed an increase of alpha activity on EEGs taken during the time of testing and smaller variance when alpha decreased. The latter situation is consistent with more ongoing cognitive work.

No generalizations leap out of this collection of findings, but I have ventured a model (Carpenter, 1983a) that may be consistent with them. Psi-hitting and psi-missing may be imagined to be alternating or oscillating functions, which switch from one to the other, unconsciously, at some rate during the response-work of most ESP subjects. If the rate of switching is slow (slower than the length of time taken to perform a single run), the deviations of run-scores will be large because a predominance of either hitting or missing will characterize each run's work. On the other hand, if the rate of switching is fast (faster than the length of a run), the hitting and missing tendencies will tend to cancel each other out and the deviations will be small. What could affect the rate of such a switching of unconscious tendencies? Perhaps the switches could be triggered by worry and conscious calculation and by inward shifts of attention, strategy, and conceptual work, all of which might tend to constrict variance by mixing brief periods of hitting and missing. In short, any self-reflective attention to what one is doing and how it is being done might trigger the unconscious switch. Periods of responding that have little self-reflection and few such shifts of attention, strategy, and thinking would tend to produce large variance (large deviations). Whittlesey's tripping subjects probably had quickly shifting attention and thinking, and Rogers's "negative" subjects may have as well, as they struggled along in their disagreeable task. Stanford's "balanced-calling" subjects had more active conceptual work to do than their spontaneous counterparts, and the subjects in the decline-of-variance studies may have become more analytical, self-doubting, and inwardly inconsistent in strategy as they labored on and on without feedback as to their success.

The older literature might have some pertinent hints after all, in spite of the fact that scoring variability was not an explicit concern for it. One feature shared by the suggestions of Warcollier (1938), Sinclair (1930), Sidgwick (1924), and others reviewed by White (1964) is a kind of faithful discipline, a *method*, of inward self-preparation for seeking the extrasensory information. Perhaps inspired by the rigorous introspective methods that were then in vogue in psychology, these approaches are all diametrically opposed to the sort of unstructured "spontaneity" usually asked of subjects in more recent times. They all require a very patient passivity and disciplined blankness of attention. Self-reflection, self-doubt, conscious calculation, experimental changes of inward tack, all of which may by the model suggested here decrease scoring extremity, are soundly renounced by these older techniques. These techniques might thus have encouraged scoring that is relatively consistent in direction, as well as a tendency toward psi-hitting.

As to the results of the present analysis, the items *fearless, dull,* and *carefree* might all seem to represent relatively unreflective states in which little internal inconsistency, self-doubt, and conceptual work would be applied to the ESP task. *Forceful* may be too bold and confident, *dull* too apathetic, and *carefree* too cheerfully unconcerned to worry over the task very much. On the other hand, *drifting* (and other items of Factor 6) suggests a state of shifting approach and flexible perspectives, *annoyed* could represent an irritable tendency to work at the task and resent it alternately, and *task-involved* seems an ideal term to convey the kind of conscious reflection and calculation just described—all of which might trigger the shifts of psi-hitting and psi-missing, producing small variance.

These speculations may be correct to some extent, but caution is needed. They are not tied as firmly to previous findings as the relationships found for hitting tendency were. This and the weak findings of the third study make the prediction of variance using these mood items seem relatively uncertain.

Retrieval of Coded Information Using Repeated-Guessing Analyses

The results of the repeated-guessing analyses in the three studies reported here all tend to support the usefulness of this way of enhancing the reliability of an effort to acquire some piece of information using ESP. The work of a suitable group of subjects can be combined and rendered by using appropriate independent predictions, resulting in a level of reliability superior to that of the raw data of the whole sample. The additional steps of encoding ESP targets to represent verbal or numerical information apparently present no barrier to the procedure.

Reflection makes it clear that the use of such repeated-guessing analyses is not limited to predictions based on mood items or the sheep-goat question. *Any ESP effect*, if reliable, and based on a predictor that is independent of the ESP data themselves, can be used in similar analyses, provided that targets are presented repeatedly and their identity preserved.

There is an obvious caution, perhaps unnecessary. These efficiency-heightening procedures are only as good as the independent discriminations about performance that precede them. The problem of developing such discriminations (the "replicable phenomena" long sought by parapsychology) is still as much a problem as ever. In these studies, scales that worked well in earlier studies did not do so in later ones. This was not unexpected, as already discussed, but even the latest scales are only as good as replication proves them to be. Nor are there other discriminators that are nearly as reliable as we would wish. "Applied ESP" is not just around the corner.

Weaknesses of Method and Future Directions

This program of work was begun with some incorrect assumptions about what mood-items might best predict variance. The prediction of hitting rate was not considered at all. Many changes in item content might heighten the power of predicting these parameters of performance. The content has been held constant until now for the sake of carrying out a sustained, programmatic body of research. Enough seems to have been done with these items to permit culling the list of several that show no promise at predicting performance and adding others that might be useful. A new MACL has been designed for future studies, with particular attention being paid to the more adequate measurement of anxiety and of the kinds of self-consciousness, self-doubt, and conceptual work (and their opposites) that are hypothesized to be pertinent to variance.

Response to the MACL has been a very simple matter in these studies. The subject simply checked the items that described his or her mood at the time of testing and left the others blank. This led to a great variation in the number of items checked per page, and the simple binary response yields very crude psychometric information. This method was chosen in the belief that such a quick and nonintrusive form of response would provide the least possible interference with the mood phenomena the scale was intended to measure. The method has been held constant until now for the sake of programmatic consistency. The new MACL will require a response to each item, but one that is still very simple and quick. It is hoped that this compromise will provide psychometrically stronger data and still be relatively nonintrusive.

The studies carried out until this point have all used minimal instructions and limited attention to subjects. The later studies have used as experimenters people who had little knowledge of the research and little investment in it. This was done deliberately to prove that the relationships studied were not somehow due to the influence of a single experimenter. It seems likely, however, that this increasingly indifferent treatment of subjects could discourage the very moods that are being found to facilitate success. Future research will use experimenters who attempt to convey genuine enthusiasm over the research and encourage an attitude of personal exploration on the part of the subjects. There will also be a focus on populations that are likely to have higher densities of low-authoritarian subjects.

Appendix A

The significance of a single variance was tested by use of the following *z*:[7]

$$z = \frac{RSV - \frac{T}{4}}{\sqrt{var\ (RSV)}}$$

For testing the difference between two variances (*RSV*s) with nonindependent samples, the following is used:

$$z = \frac{RSV_1 - RSV_2}{\sqrt{var(RSV_1) + var(RSV_2) - 2cov(RSV_1, RSV_2)}}$$

The following definitions are used:

R = number of runs
T = number of trials per run
P = ½ (with two target symbols)
X_i = number of hits in *i*th run, where $X_i \sim B(T,½)$, i.e., the distribution of X_i follows the binomial

$$D_i = X_i - \frac{T}{2}$$

M_{ij} = number of common calls in *i*th and *j*th call sequences

$$\alpha_{ij} = \frac{M_{ij}}{T}$$

$$RSV = \frac{1}{R}\sum_{i=1}^{R} D_i^2$$

$$E(RSV) = \frac{T}{4}$$

The equations for variance and the covariance term are:

$$RSV = \frac{T(T-1)}{8} - \frac{T^2(R-1)}{2R}\left[\frac{1}{R(R-1)}\sum\sum_{i \neq j} \alpha_{ij}(1-\alpha_{ij})\right]$$

and

$$cov(RSV_1, RSV_2) = \frac{T(T-1)}{8} - \frac{T^2}{2}\left[\sum_{i=1}^{R1}\sum_{j=1}^{R2} \alpha_{ij}(1-\alpha_{ij})\right]$$

[7] For clarity, the term *var* is used in these formulae to indicate an empirical variance, and *RSV* is used to denote run-score variance, the parameter being predicted, which is a measure of average dispersion around the theoretical mean.

The total score (mean number of hits per run) of a set of runs is tested by:

$$z = \frac{\overline{X} - \frac{T}{2}}{\sqrt{var(\overline{X})}}$$

$$var(\overline{X}) = \frac{T}{4R} + (R-1)\frac{T}{2R}\left[\sum\sum \alpha_{ij}(1-\alpha_{ij}) - \frac{1}{2}\right]$$

The difference between two means is tested by:

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{var(\overline{X}-\overline{Y})}}$$

Where the additional definitions obtain:
R_1 = number of runs in Group 1
R_2 = number of runs in Group 2
X_i = number of hits in the *i*th run of Group 1
Y_j = number of hits in the *j*th run of Group 2
α_1 = average of α_{ij} for Group 1
α_2 = average of α_{ij} for Group 2

$$\alpha_{12} = \frac{1}{R_1R_2}\sum_{i=1}^{R1}\sum_{j=1}^{R2}\alpha_{ij}$$

= between-group average α_{ij} for Groups 1 and 2

$$\overline{X} = \frac{1}{R_1}\sum_{i=1}^{R1}X_i \qquad \overline{Y} = \frac{1}{R_2}\sum_{j=1}^{R2}Y_j$$

$$var(\overline{X}-\overline{Y}) = var(\overline{X}) + var(\overline{Y}) - 2cov(\overline{X}-\overline{Y})$$

$$cov(\overline{X}-\overline{Y}) = \frac{T}{2}\left[\alpha_{12} - \frac{1}{2}\right]$$

$$var(\overline{X}) = \frac{T}{4R_1} + \frac{T(R_1-1)}{2R_1}\left[\alpha_1 - \frac{1}{2}\right]$$

$$var(\overline{Y}) = \frac{T}{4R_2} + \frac{T(R_2-1)}{2R_2}\left[\alpha_2 - \frac{1}{2}\right]$$

For further discussion of these formulae, see Carpenter (1983a) and Burdick (1983).

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, J. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper and Row.

Barron, F. (1953). Some personality correlates of independence of judgment. *Journal of Personality*, **21,** 287–297.

BOERENKAMP, H. G. (1985). A study of paranormal impressions of psychics: Part II. The standard series. *The European Journal of Parapsychology*, **5,** 349–371.

BOERENKAMP, H. G. (1986). A study of paranormal impressions of psychics: Part V. The group of control series with nonpsychics. *The European Journal of Parapsychology*, **6,** 259–284.

BURDICK, D. S. (1983). Analyzing binary data in the presence of the stacking effect: A technical note. *Journal of Parapsychology*, **47,** 237–242.

BURDICK, D. S., & KELLY, E. F. (1977). Statistical methods in parapsychological research. In B. Wolman (Ed.), *Handbook of parapsychology* (pp. 81–130). New York: Van Nostrand Reinhold.

CARPENTER, J. C. (1966). Scoring effects within the run. *Journal of Parapsychology*, **30,** 73–83.

CARPENTER, J. C. (1968). Two related studies on mood and precognition run-score variance. *Journal of Parapsychology*, **32,** 75–89.

CARPENTER, J. C. (1969). Further study on a mood-adjective check list and ESP run-score variance. *Journal of Parapsychology*, **33,** 48–56.

CARPENTER, J. C. (1983a). Prediction of forced-choice ESP performance: Part I. A mood-adjective scale for predicting the variance of ESP run-scores. *Journal of Parapsychology*, **47,** 191–216.

CARPENTER, J. C. (1983b). Prediction of forced-choice ESP performance: Part II. Application of a mood scale to a repeated-guessing technique. *Journal of Parapsychology*, **47,** 217–236.

CARPENTER, J. C., & CARPENTER, J. C. (1967). Decline of variability of ESP scoring across a period of effort. *Journal of Parapsychology*, **31,** 179–191.

CRANDALL, J. E. (1985). Effects of favorable and unfavorable conditions on the psi-missing displacement effect. *Journal of the American Society for Psychical Research*, **79,** 27–38.

DAVIDS, A. (1955). Some personality and intellectual correlates of intolerance of ambiguity. *Journal of Abnormal and Social Psychology*, **51**, 415–420.

FISK, G. W. (1951). Home-testing ESP experiments: A preliminary report. *Journal of the Society for Psychical Research*, **36,** 369–370.

FISK, G. W., & WEST, D. W. (1956). ESP and mood: Report of a "mass" experiment. *Journal of the Society for Psychical Research*, **38,** 1–7.

FISK, G. W., & WEST, D. W. (1957). Towards accurate predictions from ESP data. *Journal of the Society for Psychical Research*, **39,** 157–164.

HONORTON, C. (1977). Psi and internal attention states. In B. Wolman (Ed.), *Handbook of parapsychology* (pp. 435–472). New York: Van Nostrand Reinhold.

HONORTON, C., BERGER, R. E., VARVOGLIS, M. P., QUANT, M., DERR, P., SCHECHTER, E. I., & FERRARI, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology*, **54,** 99–140.

HONORTON, C., & FERRARI, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology,* **53,** 281–308.

JACKSON, D. N., MESSICK, S. J., & SOLLEY, C. M. (1957). How "rigid" is the "authoritarian"? *Journal of Abnormal and Social Psychology,* **54,** 137–140.

KANTHAMANI, H., & KELLY, E. F. (1974). Awareness of success in an exceptional subject. *Journal of Parapsychology,* **38,** 355–382.

KENNEDY, J. E. (1979). Redundancy in psi information. *Journal of Parapsychology,* **43,** 290–314.

KLEINBAUM, D. G., & KUPPER, L. L. (1978). *Applied regression analysis and other multivariable methods.* Belmont, California: Wadsworth Publishing.

KOGAN, N. (1956). Authoritarianism and repression. *Journal of Abnormal and Social Psychology,* **53,** 34–37.

LAYTON, B. D., & TURNBULL, B. (1975). Belief, evaluation, and performance on an ESP task. *Journal of Experimental Social Psychology,* **11,** 166–179.

MURPHY, G. M. (1962). A qualitative study of telepathic phenomena. *Journal of the American Society for Psychical Research,* **56,** 63–79.

MURPHY, G. M. (1966). Research in creativeness: What can it tell us about extra-sensory perception? *Journal of the American Society for Psychical Research.* **60,** 8–22.

NOWLIS, V. (1961). Methods for studying mood changes produced by drugs. *Revue de Psychologie Applique,* **11,** 373–386.

NOWLIS, V. (1965). Research with the mood-adjective check list. In S. S. Tomkins & C. E. Izard (Eds.), *Affect, cognition and personality* (pp. 352–389). New York: Springer Publishing.

PALMER, J. (1977). Attitudes and personality traits in experimental ESP. In B. Wolman (Ed.), *Handbook of parapsychology* (pp. 175–201). New York: Van Nostrand Reinhold.

PEDHAZUR, E. J. (1982). *Multiple regression in behavioral research.* New York: CBS College Publishing.

RAND CORPORATION (1955). *A million random digits with 100,000 normal deviates.* Glencoe, IL: Free Press.

RHINE, J. B. (1934). *Extra-sensory perception.* Boston: Bruce Humphries.

ROGERS, D. P. (1966). Negative and positive affect and ESP run-score variance. *Journal of Parapsychology,* **30,** 151–159.

ROGERS, D. P. (1967). Negative and positive affect and ESP run-score variance: Study II. *Journal of Parapsychology,* **31,** 290–296.

ROGERS, D. P., & CARPENTER, J. C. (1966). The decline of variance within a testing session. *Journal of Parapsychology,* **30,** 141–150.

ROLL, W. G., & TART, C. T. (1965). Exploratory token object tests with a "sensitive." *Journal of the American Society for Psychical Research,* **59,** 226–236.

RYZL, M. (1966). A model of parapsychological communication. *Journal of Parapsychology,* **30,** 18–30.

SCHMEIDLER, G. R. (1960). ESP in relation to Rorschach test evaluation. *Parapsychological Monographs No. 2*. New York: Parapsychology Foundation.

SCHMEIDLER, G. R., & MCCONNELL, R. A. (1958). *ESP and personality patterns*. New Haven: Yale University Press.

SCODEL, A., & MUSSEN, P. (1953). Social perceptions of authoritarians and nonauthoritarians. *Journal of Abnormal and Social Psychology*, **48,** 181–184.

SIDGWICK, MRS. H. (1924). Report on further experiments in thought-transference carried out by Professor Gilbert Murray. *Proceedings of the Society for Psychical Research,* **4,** 212–274.

SIEGEL, S. M. (1956). The relationship of hostility to authoritarianism. *Journal of Abnormal and Social Psychology,* **52,** 368–372.

SINCLAIR, U. (1930). *Mental radio*. Monrovia, California: Upton Sinclair.

SINGER, R. D., & FESCHBACH, S. (1959). Some relationships between manifest anxiety, authoritarian tendencies, and modes of reaction to frustration. *Journal of Abnormal and Social Psychology*, **59**, 404–408.

SPSS user's guide. (1983). Chicago, Illinois: SPSS Marketing Dept.

STANFORD, R. G. (1966). The effect of restriction of calling upon run-score variance. *Journal of Parapsychology*, **30,** 160–171.

STANFORD, R. G. (1967). Shifts in EEG alpha rhythm as related to calling patterns and ESP run-score variance. *Journal of Parapsychology*, **31,** 313.

TAETZSCH, R. (1962). Design of a psi communication system. *International Journal of Parapsychology*, **4,** 35–66.

THAYER, R. E. (1971). Personality and discrepancies between verbal reports and physiological measures of private emotional experiences. *Journal of Personality*, **39,** 57–69.

THOULESS, R. (1960). The repeated guessing technique. *International Journal of Parapsychology*, **2,** 21–36.

WARCOLLIER, R. (1938). *Experimental telepathy*. Boston: Boston Society for Psychical Research.

WHITE, R. A. (1964). A comparison of old and new methods of response to targets in ESP experiments. *Journal of the American Society for Psychical Research,* **58,** 21–56.

WHITTLESEY, J. R. B. (1960). Some curious ESP results in terms of variance. *Journal of Parapsychology*, **24,** 220–222.

WILKINSON, L. (1988). *Systat: The system for statistics*. Evanston, IL: Systat, Inc.

727 Eastowne Dr.
Suite 300B
Chapel Hill, NC 27514