Journal of Parapsychology, Vol. 50, June 1986

THE CHECKER EFFECT REVISITED

BY DEBRA H. WEINER AND NANCY L. ZINGRONE

ABSTRACT: Feather and Brier found that the outcomes on precognition runs, arbitrarily divided between them, varied depending on who scored ("checked") the runs. They asked subjects to predict which runs would be checked by the test administrator and which by an unnamed colleague; differential scoring between these runs occurred only when the checker was indeed the test administrator. Further, Feather and Brier obtained significantly different scoring rates. These results were not due to a difference in checkers' accuracy but seemed to be psi-mediated.

The present work is a two-series conceptual replication of this research. From the Feather-Brier work, two measures of checker influence were planned. In addition, by manipulating whether checkers were blind to subjects' predictions of the checker, Series II tested the hypothesis based on the observational theories that checkers influence the data during their observation of the call-to-target match.

Series I found significant evidence of checker influence on one measure and indirect evidence on the other. Series II found both measures significant but *only* when checkers were not blind to subjects' predictions of who the checker would be. The results of these not-blind data resembled those of Series I, also checked under not-blind conditions.

Although the data best support the view that the checkers were the psi sources in these experiments, there is secondary evidence for psi from the subjects. We propose that the occurrence of checker influence only when checkers were aware of subjects' predictions of the checker may be because this knowledge made the observation of the call-to-target match *meaningful* or *complete*.

By 1968, researchers at the Institute for Parapsychology had noticed that on occasion the person who scored precognition tests (the "checker") seemed to play some role in the results. In a paper published that year, Feather and Brier related two anecdotes that had suggested this. In them, two persons had shared the checking task: one had checked the calls of half the subjects; the other had

A preliminary report of this research was presented at the 27th Convention of the Parapsychological Association in Dallas, Texas, during August, 1984.

We thank Ginia Jahrke, Norm Bell, and Shirley Koczan (Chicago City-Wide College), Vicki Newsome (Durham Technical Institute), and John Near (Elgin Community College) for helping us recruit subjects, and Allison Amoroso, Linda Ironside, Paul DiGiovanni, Shelly Anderson, and Linda Vann for being our assistants. We also thank James Perlstrom and George Hansen for their help, J. E. Kennedy, Richard S. Broughton, and John Palmer for their comments on an earlier draft of this report, and Donald S. Burdick for statistical consultation. We especially thank Patric V. Giesler, whose extensive and insightful commentary inspired new thinking and led to major improvements in all aspects of this paper.

checked the other half. Although the division of subjects had been arbitrary, the outcomes of the two sets of data had differed significantly. Intrigued, Feather and Brier (1968) tested whether this was related to subjects' conscious expectations or unconscious precognition of who the checker would be.

In their pilot series, Feather gave a four-run precognition test to each student of a parapsychology course she was teaching. Apparently in an effort to assess simultaneously the subjects' conscious expectations and their ability to precognize the checker, she also asked them to predict which two runs would be checked by her and which two by an unnamed colleague (Brier). Later, Feather and Brier used digits from a random number table to determine which of them would be the checker for each run and what the sequence of targets would be.[1] In a confirmatory series, they switched roles: Brier was the test administrator and Feather the anonymous colleague.

The same result was observed for both series. There was a significant difference between the runs predicted for the test administrator and the runs predicted for the colleague, but *only* in the data actually checked by the test administrator (pilot: $CR_d = 2.2$, $p < .03$, two-tailed; confirmatory: $CR_d = 2.08$, $p < .02$, one-tailed). No such difference was found for data checked by the colleague.

This result is not an artifact of checking errors. Although their published report does not state whether the data were double-checked, Feather has confirmed that double-checking was standard procedure at the Institute and had been carried out (S. R. Feather, personal communication, July 19, 1984).

The intriguing aspect of these results is that they depend not just on who the subjects *thought* would check their runs but also on who, in fact, *did* check the runs. If the results had been just an effect of the subjects' expectation to perform better on the runs they thought the test administrator would check, subjects would have scored higher on these runs no matter who checked the data. Furthermore, these results cannot be explained by a simple precognition hypothesis that correctly predicting the checker (whether it be

[1] Feather and Brier stated that they used "the standard method" (p. 169) for obtaining an entry point into the random number table to select these digits but did not describe this method. Morris (1968) described in detail the standard method used at the Institute for Parapsychology at the time, and we based our procedure on his description. That Morris's method is the same as Feather and Brier's is supported by Freeman (1968), another researcher working at the Institute at the time, who also referred to the "standard procedure" (p. 178) and who cited Rhine and Pratt (1957). The comparison of Rhine and Pratt's method (pp. 162–163) with Morris's shows them to be the same in essential aspects.

the test administrator or the anonymous colleague) causes significantly different run scores from those obtained when the checker is not correctly predicted. Instead, the difference was significant only on the runs actually checked by the test administrator. As Feather and Brier interpret it:

> Since the significance occurs only on those runs which the experimenter checked, it appears that the person who actually checks the test is having some effect upon the scores of the test he is checking. This contradicts the notion that once the tests are alphabetized and an entry point into the tables is obtained by random means, the scores are determined and that from that point on nothing can affect them. After the entry point is obtained, it seems, there is still one factor at least which may affect the score—the checker. (p. 174)

Interpretations of the Checker Effect

Who is responsible for the checker effect? Feather and Brier saw it as a consequence of the subjects' unconscious (i.e., psi-mediated) reaction to the person who would be checking their data, though they also alluded to the possibility of backward causation ("the past appears to be affected by the present," p. 174). More recently, however, parapsychologists have suggested an alternative view: the *checkers*, not the subjects, were the psi sources in Feather and Brier's experiments. Both Kennedy and Taddonio (1976) and White (1976a, 1976b) discussed Feather and Brier's work in their landmark reviews of experimenter effects in parapsychological research. Kennedy and Taddonio, for example, interpreted the Feather-Brier results as a possible case of a psi-based experimenter effect, bolstering this interpretation with a reanalysis of the data showing overall significant missing when Brier was the anonymous checker ($CR = -3.04$, $p < .002$, two-tailed); for these runs the scores were significantly lower than those obtained when Feather was the anonymous checker ($CR_d = 3.18$, $p < .002$, two-tailed). Presumably, Kennedy and Taddonio compared only these data because they were uncontaminated by the tester- versus colleague-predicted difference in the data of the test administrator. But when they later analyzed all the data for both checkers, they still found a significant difference in scoring rate between Feather and Brier (Kennedy, O'Brien, O'Brien, & Kanthamani, 1977). If the checkers were truly neutral parties in the experiment, we would not expect to see a difference in their scoring rates.

Kennedy and Taddonio formed no conclusions about how Feather and Brier had affected their data. They did state, however, that they considered the hypothesis that the checkers unintentionally used psi during target generation (presumably, by using PK on the dice tosses that obtained the entry point into the random number table) to be more reasonable than the hypothesis that the checkers influenced the results (Kennedy & Taddonio, 1976, p. 13). White (1976a) offered a slightly different possibility by considering the results in the context of Schmidt's (1974) writings on the relationship between quantum physics and parapsychology. Schmidt had suggested that the outcomes of random events become determined not when they are generated but when they are *observed.* White, speculating from this view, suggested: "Even after the entry point into the random number table has been obtained, the observer (checker) may still influence the results" (White, 1976a, p. 144). Because the present research considers the sort of observer-effect hypothesis White alluded to, we shall briefly discuss its theoretical rationale.

An Observer Model of the Checker Effect

In the last decade, a set of theories grouped under the rubric *the observational theories* has attempted to explain psi phenomena through radical interpretations of the "measurement problem" in quantum mechanics (Schmidt, 1975; Walker, 1975). (For a nontechnical introduction to the theories, see Millar [1978]; see also Walker [1984] for a discussion of Millar's treatment.) These theories hold that a random event is indeterminate—a collection of all possible outcomes—until it is observed by a conscious being. At the time of observation, one state (e.g., a hit, a miss) is selected from this collection and becomes the final outcome. Because this selection is thought to occur as part of the process of observation itself, sensory feedback about the outcomes of random events is considered necessary for the observer to influence these events. Further, the theories hold that psi is time-independent; that is, the time of initial observation relative to the time of target generation is immaterial.

Phillips (1984) has criticized these theories for resting on an unjustified interpretation of quantum mechanics, and Braude (1979) has criticized them for containing logical paradoxes; others have defended them (Bierman, Houtkooper, & Millar, 1981; Walker, 1984). Our intention here, however, is not to evaluate these theories in

their entirety but to see to what extent they can help us understand possible checker effects.

If we look at the Feather-Brier study from an observational-theory point of view, the important moment of their study is the moment of initial observation. But at what stage in their experimental procedure did that first observation take place? During the precognition tests, subjects observed their guesses as they wrote them on the call sheets. Was this the moment of first observation? Although the question of what constitutes an observation in these theories is not resolved (e.g., Bierman, Houtkooper, & Millar, 1981; Braude, 1979; Rao, 1977; von Lucadou & Kornwachs, 1980a; Walker, 1984; Weiner & Bierman, 1982), nevertheless, both the major observational theorists (Schmidt, 1975; Walker, 1975) agree that in the case of an ESP experiment, the crucial observation takes place *when the subject's response is matched to the target.* In other words, these theories hold that ESP occurs not when subjects make their guesses but when the guesses are compared with the targets.

This view has obvious implications for a checker effect. When precognition calls are hand-scored, the checker is the first person to observe the match between call and target. Therefore, according to the observational theories, the checker may play a particularly important role, if not the only role, in producing significant results.[2] Thus, not only is the idea that checkers can influence precognition test outcomes consistent with the observational theories, it is predicted by them.

If we take the observational-theory view and consider Feather and Brier as being the psi sources who are operating while they are checking their data, how might we understand their experiment? Their design becomes reconceptualized as a nonintentional psi test of two individuals, each performing with two "task types": runs predicted for himself or herself and runs predicted for the partner. Kennedy et al.'s (1977) findings of a significant difference in Feather's and Brier's scoring rates can then be interpreted as evidence of a difference in their psychic functioning, in the same way that a sig-

[2] The question whether the first observer plays the *only* role or whether subsequent observers can also influence results has received theoretical and empirical attention (e.g., Bierman & Houtkooper, 1981; Bierman & Weiner, 1982; Hartwell, 1977; Houtkooper, 1982; Millar & Hartwell, 1979; Schmidt, 1984; Walker, 1977; Weiner, 1982; see especially Houtkooper, 1983). Although there is some evidence of second-order observer effects (e.g., Weiner & Bierman, 1979), Schmidt (in press) has recently obtained support for a first-observer-only model. If this model continues to receive support, it will enhance the theoretical importance of the checker effect for the observational theories.

nificant difference in scores between two groups of subjects is considered evidence for *their* psychic functioning.

To interpret Feather and Brier's original finding (better scoring on correctly predicted runs than on incorrectly predicted runs but only for data checked by the test administrator), we can propose the following speculation: Feather and Brier both stated that at the time of their study, they had seen a checker effect as a consequence of the subjects' contact with the test administrator and that they had therefore expected significant results to occur only in that person's data (R. Brier, personal communication, November 25, 1985; S. R. Feather, personal communication, July 19, 1984). Further, they believed that the likely effect of this contact would be to increase subjects' motivation to do better on runs they correctly predicted for the test administrator, so they apparently also expected to see high scores in that condition. From our reading of their report, confirmed by the researchers (R. Brier, personal communication, November 25, 1985; S. R. Feather, personal communication, May 8, 1985) and by our inspection of the original call sheets, the subjects' notations of which checker they had predicted for that run were visible to the checkers as they checked the calls. Could it be, then, that the test administrator, expecting to be the one to obtain significant results (especially in correctly predicted runs) and knowing for whom the run had been predicted, had a different, perhaps more positive, expectation or attitude when checking runs predicted for him or her than when checking runs predicted for the anonymous checker? And could it follow that no such psychological difference occurred for the colleague who did not expect to obtain significant results?

This speculation rests not only on the supposition that Feather and Brier had certain expectations but also on the supposition that checkers' expectations can influence the outcomes of the data they check. There is evidence to support this latter supposition. O'Brien (1979) randomly divided between two checkers precognition guesses collected under identical circumstances. One checker was given a cover story to induce an expectation of above-chance scores. The other was given a cover story to induce an expectation of below-chance scores. A pilot study showed significant above-chance scores for the hitting-expectancy checker ($CR = 2.32$, $p < .02$) and chance scores for the missing-expectancy checker, with a marginally significant difference between the two ($p = .06$, two-tailed). A second series with two new checkers yielded scores each significantly in the

directions of the induced expectation. Because O'Brien double-checked the data of both series, these differences cannot be attributed to differences in accuracy between the two checkers but instead suggest the influence of checker expectation on psi scores.

What specific thoughts Feather and Brier had while checking their runs cannot now be known. Still, because they had information regarding the subjects' predictions, we can at least say this knowledge may have allowed for a different psychological *context* (i.e., cognitive, emotional, motivational, attitudinal context) when they observed call-to-target matches on runs predicted for them than when they did this on runs predicted for their partner. It is possible that this context is an important variable.

We know of only one other study (Kennedy et al., 1977) that has attempted to replicate the Feather-Brier work, but it was not successful.[3] Because of the lack of research in this area and the implications of this work as a means for testing the observational-theory model of ESP, we undertook a two-series conceptual replication of the Feather-Brier research.

In Series I we tested the hypothesis that the checker can influence the outcome of data she checks. Our test took the form of two predictions: (1) The difference between tester-predicted and colleague-predicted runs would vary depending on who actually checked the data. This prediction follows conceptually from Feather and Brier's results, but because we also considered an observational-theory interpretation of the checker effect (which would allow for the colleague influencing psi test results) we did not specify that the difference would necessarily occur only for runs checked by the test administrator. (2) The scoring rates of the two checkers would differ significantly. This follows from Kennedy and Taddonio's (1976) and Kennedy et al.'s (1977) reanalyses of the Feather-Brier data.

In Series II, we retested these predictions and added a third prediction based on our speculations about the potential importance of checkers' awareness of subjects' predictions. This prediction and its rationale will be presented in the introduction to Series II.

---

[3] Fisk and West (1958; West & Fisk, 1953) also found significant scoring differences between data they had checked. However, because their checker manipulation was confounded with other factors, we cannot unambiguously attribute their results to their roles as checkers. Additionally, Houtkooper and Haraldsson (1985) reported a "checker effect," but as the authors themselves noted, the checkers in their experiment were the *second* observers of the data. Therefore, their results more properly pertain to the question of second-order observer effects (e.g., Weiner, 1982) and are not directly relevant to the present research.

SERIES I

Method

Testing and Subject Selection

As part of guest lectures on parapsychology given in psychology classes at Durham Technical Institute (Durham, NC) during the spring of 1983, NLZ administered to all interested students four 25-trial runs of a standard precognition test that used ESP symbols. Each run was printed on a separate call sheet for easy distribution to checkers. Each student was asked to guess which two of their four runs would be checked at a later date by NLZ and which two would be checked by an unnamed colleague (DHW). Subjects made their predictions by marking either "yes" or "no" in answer to a question printed at the top of the call sheet: "Do you think this run will be checked by the Experimenter?"

Consent forms were also included in the packet. Students were given the opportunity to request a report of their scores. They were also able to choose their level of participation by not completing the runs if they did not want to or by not completing the consent form if they wanted feedback of their ESP scores but did not want their data used in the study.

Four selection criteria were planned: (1) Students 18 years of age or older must have signed a consent form. (2) Students under 18 years of age must have a consent form signed by a parent or guardian. (3) Students must have completed all 25 trials on each of the four runs. (4) Students must have correctly followed the instructions to indicate two runs to be checked by NLZ. As will be explained later, the third criterion was modified.

NLZ initially tested 64 students. We eliminated 44 of them because they had not followed instructions to predict two (and only two) runs to be checked by NLZ, 3 of them because they had no consent form, and 2 because they had made only 5 calls per run. It is not clear why so many students in this and the subsequent series failed to follow instructions, for NLZ had repeated these instructions frequently during the test.

As we will explain later, an assistant inadvertently overlooked 5 subjects in the original selection of usable data. The final sample of usable subjects was 15, ranging in age from 19 to 50, with a mean age of 27.6 (median = 27). Of them, 8 were male and 7 were female.

Preparation of Call Sheets

NLZ designed the procedures for organizing call sheets for checking. To remain as blind as possible to subjects' calls, she had an assistant review the call sheets and consent forms and separate out data sets (i.e., the sets of four runs) for all students who met the selection criteria. Because of the large number of subjects who did not meet these criteria, only the data from 10 subjects were initially considered usable. The assistant then made two photocopies of each usable call sheet. The originals were stored in a locked cabinet not accessible to either experimenter, and only the photocopies were used by the checkers and assistants.

To prepare the data for checking, the assistant arranged the data sets in alphabetical order by subject surname. She then separated the runs predicted to be checked by NLZ (runs with "yes" in answer to the "experimenter as checker" question) from those predicted for the anonymous checker (runs with "no" in answer to the question), maintaining the alphabetical order. The "yes" runs were then divided into two sets, one for NLZ and one for DHW, using an ABBA order (NLZ, DHW, DHW, NLZ). The "no" runs were divided in a like manner. Keeping the subjects in alphabetical order, the assistant interfiled NLZ's yes and no runs so that the two runs NLZ received from a given subject were together. She then rearranged the runs within the pair so that the yes and no runs were alternated in an ABBA fashion throughout NLZ's data set (yes, no, no, yes). DHW's runs were treated in the same manner. In short, each checker received two runs from each subject, one predicted for her and one predicted for the other checker, with the order of these predictions counterbalanced across the pairs of runs. Like Feather and Brier, we were *not* blind to subjects' predictions while we checked the runs.

This procedure for assigning runs to checkers was a deliberate deviation from that of Feather and Brier, who divided runs *randomly* between checkers. We imposed an equal division to guarantee equal cell sizes for our planned method of analysis, an analysis of variance (ANOVA), which is a more powerful statistical technique than the CR_d.

Target Generation

DHW was responsible for generating target sequences. A separate target order was created for each run for each subject. After

the subjects of both series had been tested, DHW, blind to subjects' calls, obtained an entry point into a computer file containing the RAND table of a million random digits (RAND Corporation, 1955). The procedure for obtaining the entry point was a modification[4] of the standard method described by Morris (1968). (See Footnote 1.) A 10-sided die was rolled eight times to produce four two-digit numbers. These were multiplied together; the product was then multiplied by itself backwards and the square root of *that* product was taken. The five digits to the left of the decimal point were taken as the entry point. To put the entry point within range of the line numbers of the random number table, the first digit (i.e., the fifth one to the left of the decimal point) was changed to a 0 if it was even and to a 1 if it was odd. The process of obtaining this entry point was witnessed by a colleague, who retained a copy of the entry point for security purposes.

After the appropriate section of the random number table file was accessed, a computer program translated the digits into ESP symbols using a standard code (Rhine & Pratt, 1957, p. 151): 1,6 = "O"; 2,7 = "+"; 3,8 = "="; 4,9 = "L"; 5,0 = "*." The program printed target sequences separately for the two checkers and stored their coded sequences into separate computer files for later double-checking. In accordance with Feather and Brier's procedure, all of the test administrator's (NLZ's) targets followed the entry point, with DHW's targets following immediately after NLZ's targets.

Data Checking

Once the targets and call sheets were ready, NLZ received both sets of call sheets (hers and DHW's) from the assistant. She gave DHW her set and at the same time received from DHW her own target sheets. (By this type of exchange, neither checker had prior access to both call sheets and targets at the same time.) Upon re-

[4] The modifications and their rationale are as follows: (a) The entry point was the beginning of a numbered line (from 0 to 19999) of the computerized RAND table rather than a digit within a line. This was done because the program used to access the RAND table did not allow for entry within a line. (b) Four *two*-digit numbers rather than four *three*-digit numbers were obtained so that their product would not exceed the capacity of the electronic calculator being used to multiply them together. We felt that exceeding the capacity could result in a nonrandom distribution of digits (e.g., an excess of zeroes) among the least significant digits. (c) The five digits to the left of the decimal point were used rather than the five least significant digits (to the right of the decimal point) in case modification b did not completely take care of potential nonrandom distribution of digits.

ceiving these materials, the checkers proceeded to check the data. No particular rules for checking were imposed other than the two checkers agreeing to check the runs in the order given and not to discuss their results until both parties had completed the task.

During the process of checking, it became apparent that 3 of the 10 subjects whose calls had been designated as usable by the assistant had, in fact, not fulfilled the selection criteria, for only 24 trials had been completed on one or more runs. Eliminating these subjects would have reduced the sample size to seven, which was considered too low to show any meaningful trends. Instead of aborting the series, we decided to modify our original criteria and to include any subject who had completed at least *24* trials in each run. To compensate for the unequal number of trials across subjects, we transformed the run scores into *z* scores (not corrected for continuity), which then became the dependent variable.

When this decision was made we were blind to each other's results but not blind to our own (though no analyses had been made and the checker who subsequently obtained significant results was not consciously aware of any trends in her data). Still, the possibility existed that decisions about how to handle this problem could have been influenced by subconscious awareness of the results. To minimize this possibility, we (1) followed our original, fixed protocol as closely as possible and avoided decisions that could be influenced by any subconscious awareness of results and (2) consulted colleagues, blind to our results, who concurred with our procedures.

A new clerical assistant retrieved from the discarded data any sets that met this new criterion. No additional 24-trial data sets were found; however, five sets that had met the original criteria but had apparently been overlooked by the first assistant were discovered.

Because the calls for the first 10 subjects had already been checked, incorporating the new data into the original 10 sets by realphabetizing the entire group of 15 subjects would have resulted in new target assignments for most of the original 10 subjects. We considered this to be an improper procedure because it would be tantamount to generating new target sequences for the subjects. This would not only have been confusing from a parapsychological point of view (i.e., which set of targets were the subjects supposed to precognize?), but, more importantly, it would have afforded the experiment an improper "second chance" for significant results. We therefore left the 10 sets intact and added the five new sets to the end of the original set. These five sets were arranged in alphabetical order among themselves and divided between checkers by the

method described earlier. Targets for these sets were created from the digits immediately following those used for the initial 10 sets. No new entry point was obtained.

Because of computer malfunction, the data were not double-checked by computer, as planned, but were independently hand-checked by two new assistants. These assistants used clean, unmarked copies of the call sheets. Discrepancies between the checkers' and assistants' results were resolved by each checker working in consultation with her assistant. (These discrepancies involved at most only a few percent of the trials and, of course, were not always the checkers' errors. Therefore, we can safely consider the effect of the assistants as first observers of the correct call-to-target matches to be negligible.)

Run-score feedback, along with information on how to evaluate the scores and a statement describing the results of the study, was sent in the winter of 1985 (after completion of Series II) to all who had requested it. Thirteen subjects (86.7%) had requested feedback. To give feedback to persons whose data had been unusable, after completing Series II and making a preliminary report of this research (Zingrone & Weiner, 1984) we generated a special set of targets by the method described earlier. Subjects were *not* told which checker had scored which run. It would not have been possible to give this information to students whose data had been unusable (because these runs were not hand-scored), and we did not wish to embarrass them by telling them they had failed to follow instructions properly.

Predictions and Method of Analysis

The data were analyzed by a 2 × 2 ANOVA using actual checker (C) and subjects' prediction of checker (P) as the independent variables. There were two predictions:

1. *The scoring difference between tester-predicted and colleague-predicted runs would vary depending on who actually checked the run.* As already explained, this prediction follows conceptually from the original Feather-Brier finding, but we did not restrict our prediction to the data of the test administrator only. Further, as runs predicted for the test administrator or for the anonymous checker could be viewed simply as two task "types" confronting the checker, we did not prespecify that a significant difference between them would necessarily show better performance on correctly predicted runs.

Therefore, Prediction 1 would be supported by a significant checker-by-prediction (C × P) interaction of any form.

2. *A significant difference in overall scoring rate between the two checkers would be found.* This prediction follows conceptually from Kennedy and Taddonio's (1976) and Kennedy et al.'s (1977) reanalyses of the Feather-Brier data and would be confirmed by a significant main effect for the checker factor. It was not prespecified which checker, if either, would score significantly differently from mean chance expectation and if so, in which direction above or below chance.

Although an analysis of variance is the correct method of testing these predictions, the *form* of ANOVA that is most appropriate depends on the assumptions one makes about the source of the psi effect. If the subject is considered the psi source, then it is correct to use a dependent ANOVA, in which prediction of checker and actual checker are repeated measures. However, if the checker is thought to be the source, then a fixed-effects ANOVA is appropriate. It was planned to analyze Series I with both methods and then to choose only one method for the analysis of Series II. It should be stressed that these ANOVAs do not *test* the hypotheses that the subject or the checker, respectively, are in fact the psi sources. Instead, the choice of ANOVA has to do with the appropriateness of the statistical method for the assumptions underlying the analysis. (This point—that the researcher's views regarding the psi source have consequences for the method of analysis—may be relevant for a large range of psi experiments.)

The data were analyzed with the BMDP package (analyses of variance) and the Institute's standard statistical programs for t tests. Analyses were carried out by both of us working together and were double-checked by assistants who independently organized the data and reran the analyses.

Results

The first prediction, that of a checker-by-prediction interaction, was not confirmed by either the checker-source or subject-source model. (See Tables 1 and 2.) However, there was indirect support for this prediction in that the difference between correctly and incorrectly predicted runs was significant in DHW's data ($t[14] = 2.42$, $p = .03$, two-tailed) but not in NLZ's data.

TABLE 1
RESULTS OF SERIES I: CHECKER-SOURCE MODEL

Source	Sum of squares	*df*	Mean square	*F*
Checker	6.43	1	6.43	8.14*
Prediction	0.95	1	0.95	1.21
C×P	0.95	1	0.95	1.20
Error	44.23	56	0.79	
Total	52.56	59		

*p = .006.

The second prediction, that of a significant difference in scoring rate between checkers, was confirmed under both the checker-source model ($F[1,56] = 8.14$, $p = .006$) and the subject-source model ($F[1,14] = 8.72$, $p = .01$).

This difference between checkers was based on suggestive hitting in runs checked by NLZ, with a mean z score of .36 (MCE = 0, $t[29] = 1.95$, $p < .061$, two-tailed) and significant missing in runs checked by DHW, with a mean z score of $-.30$ ($t[29] = -2.13$, $p < .05$, two-tailed). The missing in DHW's data was concentrated in those runs correctly predicted to be checked by her, which were significantly below chance (mean $z = -.55$, $t[14] = -2.81$, $p = .014$, two-tailed) and, as reported earlier, significantly different from those that she had checked but had been predicted for NLZ. NLZ

TABLE 2
RESULTS OF SERIES I: SUBJECT-SOURCE MODEL

Source	Sum of squares	*df*	Mean square	*F*
Checker	6.43	1	6.43	8.72*
Error	10.32	14	0.74	
Prediction	0.95	1	0.95	1.05
Error	12.75	14	0.91	
C×P	0.95	1	0.95	1.37
Error	9.68	14	0.69	
Total	41.08	45		

Note: The discrepancy in total sum of squares and degrees of freedom between this table and Table 1 is due to our not reporting here the sum of squares (11.47) and degrees of freedom (14) for the error term associated with the mean.

*p = .01.

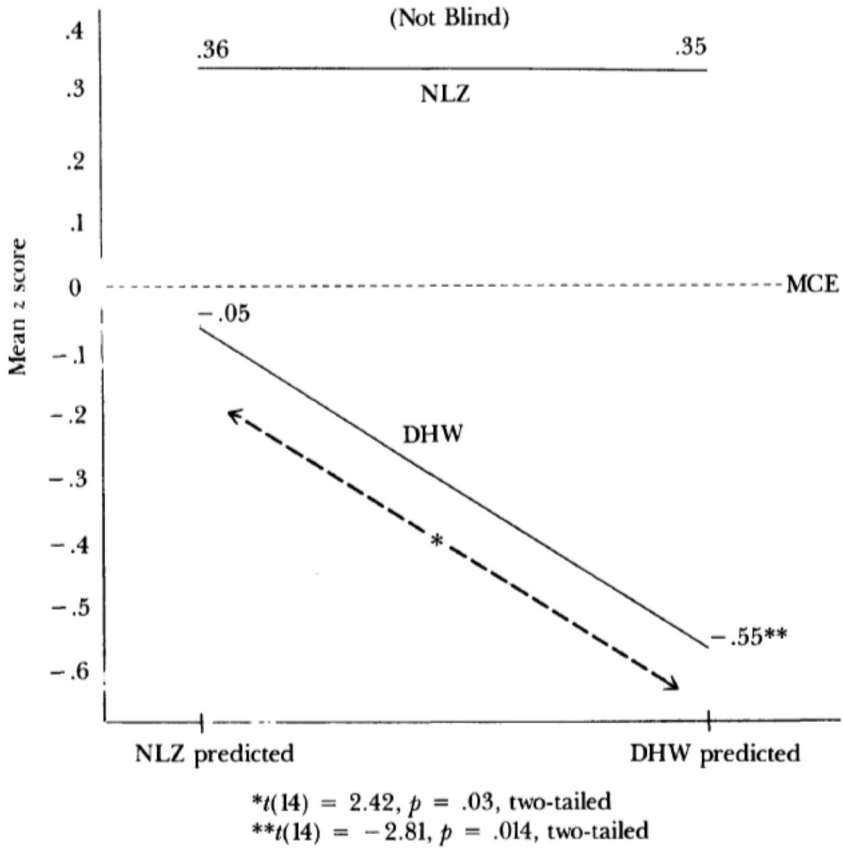

Figure 1. Results of Series I showing the difference in scoring depending on who the checker was.

obtained nearly identical mean scores for runs predicted for her and runs predicted for the anonymous checker (.36 and .35, respectively). Neither of these means differed significantly from chance. (See Figure 1.)

It may seem strange that although the magnitude of the z score for runs checked by NLZ was higher than that for runs checked by DHW (.36 vs. .30, respectively), the latter mean was significantly different from chance whereas the former was not. This occurred because of unusually low variance among DHW's scores. The variance of these scores around their own mean was 0.576 when the expected variance is 1.00. Comparing these values (Hays, 1973) shows that the variance was significantly small: $\chi^2(29) = 10.30$, $p < .01$, two-tailed. We will not speculate on this other than to note that low

variance around the empirical mean has been cited as a possible mechanism for observer effects (Weiner & Bierman, 1979; Weiner, 1982).

This low variance in DHW's data raises the question of nonhomogeneity of variance, which would violate the assumptions of the ANOVA. Cell variances were tested for nonhomogeneity by Hartley's method (Winer, 1962, pp. 92–94) and were found to be adequately homogeneous: $F_{max} = 2.41$, with the F_{crit} for the given parameters being 4.01.

Secondary Analyses

Feather and Brier found that the significance of their results was concentrated in the first runs completed by each subject. In their pilot series, Run 1's in which Feather was correctly predicted to be the checker were significantly higher than Run 1's in which Brier was predicted but Feather was the checker: $CR_d = 3.84$, $p < .0001$, two-tailed. They do not present separate statistics for Runs 2 to 4, but our calculations on their published figures show that the difference was negligible ($CR_d = .38$). The same result was found in their confirmation experiment, though to a lesser degree: the checker effect in subjects' first runs yielded a CR_d of 2.47 ($p < .01$, one-tailed), whereas the CR_d for Runs 2 to 4 was 1.06 (our calculation).

As a secondary analysis, we also investigated the checker effect in the subjects' first runs. Neither checker obtained a significant difference between runs predicted for herself or for her partner when only Run 1's were analyzed. A post hoc test found that scoring on correctly predicted Run 1's (either for NLZ or the other checker) was suggestively lower than that on incorrectly predicted runs (mean $z = -.49$ and .40, respectively; $t[13] = 1.90$, $p = .08$, two-tailed).[5] Because the evidence of checker influence here was a significant difference in scoring rate between checkers (Prediction 2), we tested whether *this* result was particularly pronounced in subjects' first runs. Subjects' first runs checked by NLZ were nonsignificantly higher than first runs checked by DHW ($z = .15$ vs. $z = -.20$, $t[13] = .70$, corrected for nonhomogeneity of variance). Instead, the difference in scoring rate was concentrated in Runs 2 to 4: $t[43] = 2.94$, $p = .005$, two-tailed. Thus, if anything, our results contradict those of Feather and Brier in that our evidence of

[5] Probability values for post hoc analyses, when given, are on a per-comparison basis and are not corrected for selection from multiple analyses. These probability values should be considered with the usual precautions regarding post hoc results.

checker influence (albeit one that is not the conceptual analog of theirs) was more pronounced in the subjects' later runs.

Series II

In much the same way as the first, the second series was carried out with a new pool of subjects. Though the procedural problems of Series I did not compromise its integrity as a test of its predictions, we felt that, given these problems, further testing was warranted. Therefore, both Predictions 1 and 2 were retained for Series II. However, in Series II, a third prediction was added, one that more directly addresses the observational-theory interpretation of the checker effect.

We had speculated that in the Feather-Brier work the different expectations of the test administrators for runs predicted for themselves as opposed to runs predicted for their partner may have produced their results. In our study, it was impossible to test this idea directly. (Both of us are too familiar with the checker-effect literature to manipulate our expectations through an artificial cover story.) However, it *was* possible to vary our awareness of subjects' predictions. This manipulation would not specifically vary our expectations but it would vary the opportunity for expectations to be formed, or—more generally and more accurately—it would vary the informational context (and any consequent cognitive, attitudinal, motivational, or emotional context) under which we would observe the call-to-target match.

This manipulation would test the hypothesis that the checker effect is an observer effect. If the observer-effect hypothesis is true, a change in the conditions of observation may reasonably be expected to affect the outcomes. On the other hand, if the checker effect is due, say, to subjects' psi-mediated reaction to the checker, the context of the call-to-target observation is not likely to matter. Further, not only would a manipulation of the information available at the point of observation throw some light on the basic problem under study (whether the subject or the checker is the psi source in these experiments), but it would illuminate psi-conducive observational conditions.

Following this line of reasoning, we decided that on half the runs the checkers would be kept blind to subjects' predictions of the checker while they checked the runs. Because this is a new manipulation in checker-effect research, we made no specific predictions about its effect.

Method

Testing and Subject Selection

In October, 1983, NLZ tested 133 students during guest lectures on parapsychology given to psychology and sociology classes at Elgin Community College (Elgin, IL) and to college-level classes taught at two girls' high schools by the College Acceleration Program of Chicago City-Wide College. Of them, 54 students were eliminated for not having correctly followed instructions for predicting the checkers, and 26 for having no consent forms. It was discovered during checking that one subject had fewer than 25 trials per run. (This subject was the last in the set; therefore, eliminating this subject in no way disturbed the target assignments for the remaining subjects.) Of the 52 subjects remaining in the sample, 8 were male and 44 female. They ranged in age from 16 to 48, with a mean age of 19.7 (median = 18).

Subjects were tested before Series I was analyzed—indeed, before the targets for Series I had even been generated. Consequently, NLZ had no knowledge of the results of Series I when she interacted with the subjects of Series II. Subjects were not aware of the blind versus not-blind manipulation.

Preparation of Materials and Data Checking

An assistant arranged and divided the call sheets for checking in the same way as in Series I. To manipulate checker blindness of subject prediction, however, she performed an additional step. After the assistant had created each checker's set of runs, consisting of a yes-no run pair from each subject, she went through each set and designated the *pair* of runs that the checker received from a subject as "not blind" or "blind" in an ABBA order (not-blind, blind, blind, not-blind). Taking the not-blind pairs as a group, the runs were alternated by checker prediction using an ABBA order as described for Series I (yes, no, no, yes). Pairs designated as blind were rearranged such that 50% of them began with the run predicted to be checked by NLZ and 50% began with the run predicted for the anonymous checker. To accomplish this, the assistant was instructed to devise an ordering system and to *not* use an ABBA order so that checkers could not infer which of the two blind runs had been predicted for her.[6]

[6] This order was as follows, with A indicating a run predicted for NLZ and B a

After the assistant coded the blind call sheets for later retrieval, she cut the sheets in half so that information about checker prediction was not available to the checkers. Blind and not-blind sets were then interfiled back into the original alphabetical order for presentation to the checkers.

In summary, the pile of call sheets received by each checker began: not-blind/yes; not-blind/no; blind/?[yes]; blind/?[no]; blind/?[yes]; blind/?[no]; not-blind/no; not-blind/yes; and so on. In this way, checker, blindness, and prediction category were manipulated orthogonally, and the number of runs in all possible combinations of conditions was the same. All four runs from a given subject were in the same blind or not-blind category.

Targets were generated in the same manner as for Series I. DHW obtained a new entry point into the RAND table, again in the presence of a colleague, who retained a copy of the entry point for security purposes.

Hand-checking was double-checked by computer. Two assistants using fresh, unmarked copies of call sheets entered calls into computer files. The calls were then computer scored against each checker's target file. Discrepancies between computer scores and hand-checked scores were resolved by each checker working together with one of the assistants.

All persons who had requested feedback, including those who had been excluded from the study, received their run scores and debriefing information (see p. 96) in the winter of 1985. Fifty subjects (96.2%) requested feedback. As in Series I, targets for persons not included in the study were generated after the main analyses had been conducted and a preliminary report made public (Zingrone & Weiner, 1984).

Predictions and Method of Analysis

The data were analyzed by a 2 × 2 × 2 fixed-effects ANOVA with checker, prediction, and blindness as the three independent variables. For two reasons, we planned to analyze Series II with the fixed-effects ANOVA only. (1) Although the F ratio of Series I's repeated-measures ANOVA was slightly larger than that of the fixed-effects ANOVA, this was due to a minor difference between the two mean square error terms (.74 vs. .79); and as the fixed-effect F ratio

---

run predicted for the other checker: 3(AB), 2(BA), 2(AB), 3(BA), 3(AB), 2(BA), 2(AB), 3(BA), 3(ABBA). (The letters refer to the order of runs within the pair; the numbers refer to the number of consecutive pairs with that order.)

TABLE 3
RESULTS OF SERIES II

Source	Sum of squares	*df*	Mean square	*F*
Checker	10.62	1	10.62	2.69*
Prediction	20.31	1	20.31	5.15‡
Blindness	5.89	1	5.89	1.49
C×P	15.62	1	15.62	3.96†
C×B	12.50	1	12.50	3.17
B×P	2.12	1	2.12	<1
C×P×B	15.62	1	15.62	3.96†
Error	788.19	200	3.94	
Total	870.87	207		

*p = .10. ‡p < .03 (not predicted).
†p < .05.

was significant (p = .006), this model was not contraindicated. (2) More importantly, the observer-effect hypothesis we were testing assumes that the checker is the psi source; thus, the fixed-effect ANOVA is more appropriate to the assumptions underlying the design.

It was predicted that: (1) *The scoring difference between tester-predicted and colleague-predicted runs would vary depending on who actually checked the run.* (2) *A significant difference in overall scoring rate between the two checkers would be found.* (3) *The checker effect described in Prediction 1 would vary depending on whether checkers were blind to subjects' predictions.*

As in Series I, Prediction 1 would be confirmed by a significant C × P interaction, and Prediction 2 by a significant main effect for the checker variable. Prediction 3 would be tested by the checker-by-prediction-by-blindness (C × P × B) interaction. No specific predictions about the effect of checker blindness on the C × P interaction were made.

All analyses were independently double-checked.

Results

Prediction 1 was confirmed: $F(1,200) = 3.96$, $p < .05$. Prediction 2 was not confirmed, though the result was suggestive: $F(1,200) = 1.69$, $p = .10$. (As will be discussed, Prediction 2 *was* confirmed in the not-blind data, which replicates the results of Series I.) Prediction 3 was confirmed: $F(1,200) = 3.96$, $p < .05$. (See Table 3.)

TABLE 4
SERIES II: SIMPLE INTERACTION EFFECTS

Source	Sum of squares	*df*	Mean square	*F*
C × P: Not blind	31.64	1	31.64	7.93*
C × P: Blind	.0001	1	.0001	$\ll 1$
Error	788.19	200	3.94	

*p = .01.

Though the C × P interaction (Prediction 1) was significant, its interpretation is affected by the significant three-way interaction of Prediction 3. Examination of the simple interaction effects of this three-way interaction clarifies the picture. (See Table 4 and Figure 2.) When checkers were blind to subjects' predictions, the checker effect (C × P interaction) was far from significant ($F \ll 1$).[7] However, when the checkers knew which runs had been predicted for which checker, the checker effect was significant: $F(1,200) = 7.93$, $p < .01$.

Breaking down this significant checker effect in the not-blind data, we see that, as in Series I, the difference in scores between the two prediction categories (NLZ or anonymous checker) in DHW's not-blind data was significant ($F[1,200] = 12.20$, $p < .001$) whereas that for NLZ was not ($F < 1$). However, there is a reversal in the direction of this difference. In this series the difference was due to significant missing in runs that DHW had checked but had been predicted for NLZ ($t[25] = -4.33$, $p < 2 \times 10^{-4}$, two-tailed) instead of in runs correctly predicted for DHW. (See Figure 2.)

The not-blind data also replicated Series I's difference in scoring rates between the two checkers ($F[1,200] = 5.86$, $p < .025$). As in

[7] It is curious that the checker effect is *so* nonsignificant in the blind data. The F ratio for the C×P interaction, 2.54×10^{-5}, is in fact significantly *small* at the .01, two-tailed level (Hays, 1973, p. 447). Although we would not want to impart any meaning to a "significantly nonsignificant" outcome, the result invites speculation about possible causes. Two possibilities are (1) there was an unknown intervening variable that systematically reversed the C × P interaction on half the runs, thereby cancelling out a C × P interaction in the remaining runs; or (2) the violation of the laws of probability represented by the significant C × P interaction in the not-blind data is somehow being "balanced" or compensated by the close fit to these laws of the C × P interaction in the blind data (see Bierman, 1985; Carpenter, 1983; Eisenbud, 1963; Rao, 1978). It should be noted, however, that George (1981) found no evidence that an increased variance in an ANOVA caused by a presumed psi effect was balanced by a decreased variance elsewhere.

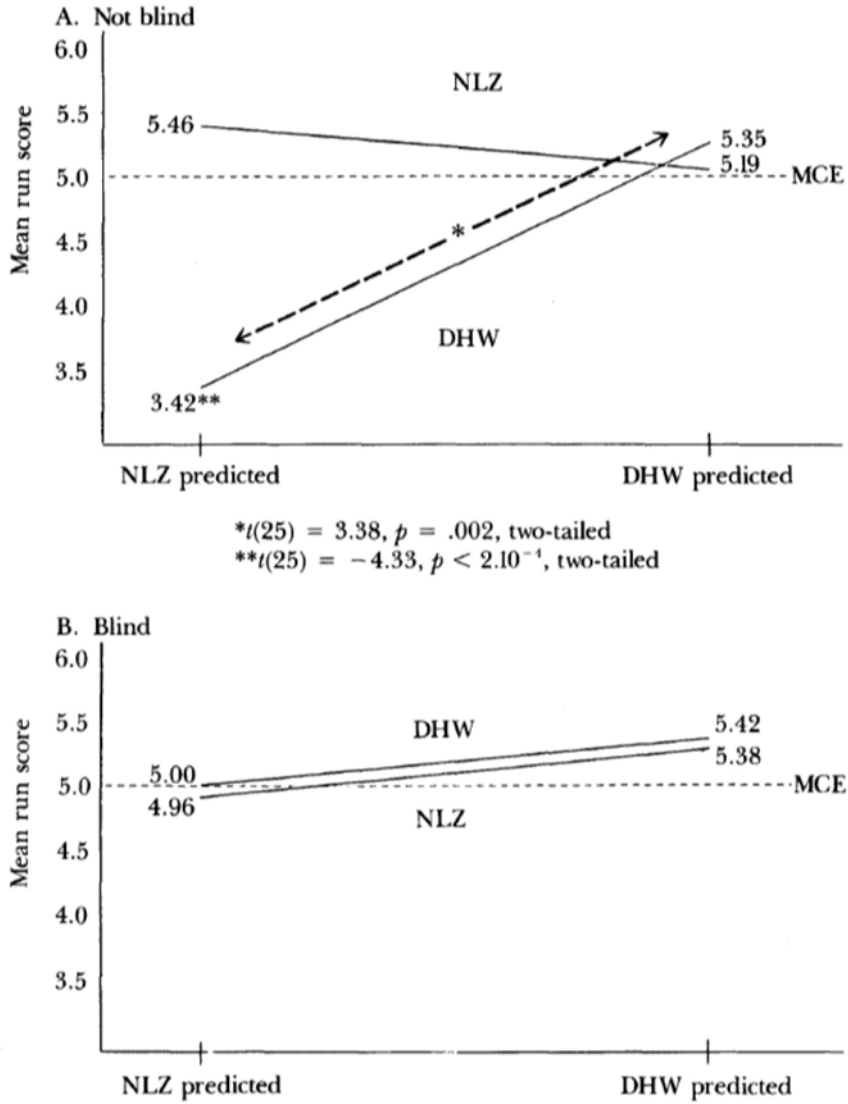

Figure. 2. Results of Series II showing the impact of the informational context of checking on the "checker effect" (C × P interaction).

Series I, runs checked by NLZ showed overall above-chance scores ($t[51] = 1.11$) though this was not near significance, whereas runs checked by DHW showed marginally significant missing ($t[51] = -2.00$, $p < .051$, two-tailed).

Secondary Analyses

In testing for a possible concentration of the checker effect in subjects' first runs, we wished at the same time to explore possible causes for such a finding. Because Feather and Brier had checked runs in the same order as subjects had completed them, subjects' Run 1's would also be the first run from that subject the checker observed. Therefore, the "Run 1 effect" could be due to the subject's completing Run 1 first or to the checker's observing it as the first of the pair of runs assigned to him or her. Our method of assigning runs to checkers allowed us to distinguish these two possibilities. However, because of a confounding of conditions with run order, we had to generalize the analysis beyond an examination of Run 1's only and compare scoring on runs subjects completed earlier versus later in the test.

A 2 × 2 × 2 × 2 fixed-effect ANOVA, using unweighted means to compensate for unequal cell sizes (Keppel, 1973), was carried out on not-blind runs only. (These data were selected because Feather and Brier checked runs only under not-blind conditions and because the blind runs contributed nothing to the overall C × P interaction.) The four factors were: (1) actual checker, (2) predicted checker, (3) subjects' run order (SRO), dichotomized as the earlier- versus later-completed run within the pair the checker received from a given subject, and (4) checker observation order (COO), the first versus second run checked within the pair.

Three terms of the ANOVA were of interest: The C × P × SRO interaction (if the Run 1 effect had to do with the run being the first one the subjects completed), the C × P × COO interaction (if it had to do with the run being the first the checker sees), and the four-way interaction (if it had to do with some combination of the two). Of these terms, only that of the C × P × SRO interaction was significant ($F[1,88] = 5.54$, $p = .021$). Inspection of the cell means showed that the C × P interaction occurred only for the *earlier* runs that the subjects completed. In these runs, DHW's data showed a significant difference between prediction categories ($p = .004$) because of psi-missing on runs predicted for NLZ ($p = .0007$), whereas NLZ's data showed a suggestive difference ($p = .087$) in the opposite direction, with suggestive hitting ($p < .08$) on runs predicted for her. In the later runs, the difference between prediction categories in DHW's data was in the same direction as just described but much weaker ($p < .11$) and the nonsignificant difference be-

tween categories in NLZ's data was now in the same direction as DHW's data.

Post Hoc Analyses

The analysis of variance showed an unpredicted main effect for subjects' prediction of checker. Post hoc analyses showed that scores on runs predicted for NLZ were nonsignificantly below chance ($t[103] = -1.31$), whereas scores of runs predicted for the colleague showed suggestive hitting ($t[103] = 1.91$, $p < .10$, two-tailed). This result would suggest the influence of subjects' psi, since the prediction factor is based solely on subject behavior and has nothing to do with either of the checker-related variables. This main effect, however, cannot be interpreted independently of the significant higher order interactions. It can be shown, for example, that the difference can be explained by cancellation or reinforcement of scoring direction when cells are pooled to form the NLZ-predicted or DHW-predicted data, respectively. (See Figure 2, top.)

Discussion

The first question to address is whether these results can be explained as a procedural or statistical artifact. Our entry-point procedure, use of the RAND table of random numbers, and method of dividing runs between checkers make the possibility of nonrandom target sequences remote. The problem of conscious or unconscious checking errors is eliminated by our independently verifying all checking. Finally, though it can never be completely ruled out in any experiment, fraud on the part of the individual experimenters is made less viable by the strict division of labor (NLZ handled calls, DHW handled target generation), the use of a fixed system to arrange call sheets (e.g., alphabetical and ABBA orders) which did not allow checkers to rearrange them to their own advantage, the use of assistants (blind to our predictions) to carry out clerical procedures, and the storage of original call sheets out of reach of both checkers. The resemblance between the results of Series I and those of the comparable data of Series II (i.e., the not-blind runs) suggests that our results are not due to statistical fluctuation, but only further replication can verify this contention.

Assuming, then, that our results represent real findings about the nature of psi, what have we found? To summarize, Series I did

not find evidence of checker influence in the form of a C × P interaction but did find a significant difference in scoring rates between checkers. Here, the anonymous checker obtained overall significant missing, concentrated in runs correctly predicted to be checked by her. In Series II, the C × P interaction was significant, and the difference in checkers' scoring rate was suggestive. Of particular interest, however, is that the C × P interaction, the conceptual test of Feather and Brier's findings, occurred *only* when the checkers, like Feather and Brier, were not blind to subjects' predictions. Furthermore, the results of the data checked under not-blind conditions resembled those of Series I, which was also checked under not-blind conditions: (1) the checkers' scoring rates were significantly different, with the test administrator again obtaining hitting and the anonymous checker again obtaining missing; and (2) a significant difference between categories was again found in runs checked by the anonymous checker, though the direction of the difference was reversed from that of Series I. A secondary analysis in Series II found that the C × P interaction was significant only for runs subjects completed relatively early in the test.

The Feather-Brier Replication

As explained earlier, we did not predict an exact replication of Feather and Brier's results. Still, it is of interest to compare our findings with theirs to see to what extent we did replicate their study. Since Feather and Brier checked all their data with knowledge of subjects' predictions, this comparison will be restricted to data checked under similar conditions (i.e., Series I and the not-blind condition of Series II).

To make this assessment, let us break down Feather's and Brier's results into seven findings: (1) significant difference between prediction categories (2) in favor of correctly predicted runs and (3) independently significant in subjects' Run 1's *but* (4) occurring only in data checked by the test administrator. Further, Kennedy and Taddonio found (5) a significant difference in overall scoring rates between checkers attributable primarily to (6) significant missing by Brier when he was the anonymous checker. Our analysis of Feather and Brier's published figures also reveals (7) that the low scores cited in Finding 6 are independently significant only for runs correctly predicted to be checked by Brier ($CR = -2.70$, $p = .007$, two-tailed). Findings 1, 2, 3, and 4 occurred in both their pilot and

confirmatory series but 6 and 7 occurred only in the former. (Finding 5 involved the *pooled* data of both series.)

In our work, both series found a significant difference between prediction categories for one checker, in one case in favor of correctly predicted runs and in one case not. Series I did not show a concentration of the checker effect in subjects' Run 1's, but, in Series II, analogous to the Run-1 effect, the difference between prediction categories was significant only in runs subjects completed relatively early in the test. It is important to note, however, that, unlike in Feather and Brier's work, the checker to obtain these results was not the test administrator. Both series replicated the significant difference in scoring rates between checkers and the significant below-chance deviation for the anonymous checker (Findings 5 and 6). Further, Series I replicated Finding 7, though this was reversed in Series II.

Why did we find the significant differences between prediction categories in the data of the anonymous checker instead of in those of the test administrator? There was nothing we could see in NLZ's interaction with subjects that would clearly suggest why subjects might be "oriented" toward the anonymous checker. One could argue that we obtained this result because we believed that the anonymous checker could influence the data. But although we entertained this possibility, both of us actually expected a significant checker effect to occur in *NLZ's* runs, albeit for different reasons. (NLZ believed that because she had had contact with the subjects, the checker effect would occur in her data. DHW believed that NLZ had more psi ability than herself and therefore as an observer would be more likely to "produce" the checker effect.) These expectations were, of course, the ones we consciously held; it may well be that *subconscious* attitudes, beliefs, expectations, and reactions play a more important role in the formation of psi scoring. However, we cannot discuss such expectations because we did not at the time engage in any psychodynamic exploration that might have uncovered them, and we would not want now to circularly infer "unconscious expectations" on the basis of the obtained results (Weiner & Geller, 1984).

It is interesting that, like Brier, the anonymous checker in our work obtained significant psi-*missing*. Brier (personal communication, November 25, 1985) interpreted his below-chance scores as subjects' psi-mediated reaction to him as an "outsider" to Feather's class. This interpretation, however, does not fit our case, because NLZ was herself a relative outsider to the classes and could not be

expected to have the same rapport with the students as Feather had had with hers. More important, missing did *not* occur when DHW was blind to subjects' predictions ($M = 5.21$). If the missing were due to subjects' psi-mediated reaction to her, we would expect this reaction for all runs she checked.

The absence of negative scoring when DHW was blind to subjects' predictions also rules out a variety of explanations having to do with checker characteristics (e.g., personality) that would not be expected to vary systematically between blind and not-blind runs. At this point, the best we can say is that the anonymous checker's awareness of subjects' predictions seemed to be important in producing the overall negative scores.

In their report, Feather and Brier did not speculate on why their checker effect was concentrated in subjects' Run 1's. To the extent that our analysis—an indirect test of the Run-1 effect—pertains to their study, it suggests that the effect is related to the subject rather than the checker. This point will be discussed later.

Last, it should be noted that there are a number of differences in the subject pools and procedures of our study and the Feather-Brier work that may account for the difference in our results, though no clear connection is apparent.

The Mechanism for the Checker Effect

Earlier we outlined three interpretations of Feather and Brier's results culled from the literature: (1) the psi sources are the subjects who are responding unconsciously (i.e., precognitively) to the checker; (2) the psi source is the researcher who unintentionally uses PK on the entry-point procedure to obtain "favorable" targets; and (3) the psi sources are the checkers who use psi during the observation of the call-to-target match. Which of these best explains our results?

Subject as source. This interpretation can be applied to the results of Series I by assuming, for example, that the overall missing in DHW's data resulted from subjects' psi-mediated negative response to her. The subject-source model, however, is difficult to apply to the main results of Series II, showing that a checker effect occurred only when checkers had information about subjects' predictions. Given the arbitrary method of dividing the runs between blind and not-blind conditions (ABBA order applied to the alphabetical listing of subjects by surname), it seems unlikely that the division corre-

sponded to some real difference between subjects that would account for the results.

A formulation of a subject-source model that avoids these problems is one that holds that subjects can respond precognitively to the checker *only* when the checker has information about subjects' predictions. In other words, the subject precognizes not the identity of the checker but the checker's *experience*. If we assume that for some reason checkers' knowledge of whom a run was predicted for is an integral part of that experience, this hypothesis would explain why no checker effect occurred for subjects whose data formed the blind condition. This hypothesis—by saying that subjects precognize the checker's experience (i.e., DHW checking a "yes" run under not-blind conditions) and are influenced in their scoring ability by that precognition—is less parsimonious than a checker-source hypothesis, which says that the checker (as the psi source) is influenced by her *own* experience and produces results accordingly. Because both hypotheses are tested by the same manipulation, it may not be possible to empirically distinguish between them.

A subject-source model is supported by the secondary analysis in Series II showing that the checker effect (C × P interaction) is moderated by an aspect of the subjects' experience (the relative order in which they completed the runs). The meaning of this result, however, is not completely clear. As it happened, the "earlier completed runs" were comprised primarily of *Run 1's* and *Run 3's* (77% of Run 3's ended up in this category). What could it be about subjects' first and third runs, then, that would be experienced differently from the remaining two? Experimenter-PK and observer models likewise are at a loss to explain this result parsimoniously. Future research will need to replicate this result and explore how this aspect of the subjects' experience impacts on the checker effect.

Experimenter PK. The second interpretation is that our results were caused by experimenter PK operating during the process of target generation. The algorithm for obtaining the entry point into the table of random numbers was based on the outcomes of dice tosses. Did DHW unwittingly use PK on the dice to get an entry point that would lead to significant results?

There is some evidence that one can use psi to obtain an entry point to produce targets corresponding to a prestated goal (Harley & Sargent, 1980; Morris, 1968). But there are reasons to believe that this hypothesis does not adequately explain our results: (1) In Morris's study, the method relying on dice-throwing to obtain the initial number to be processed through the algorithm (the method

used here) was not successful in either of his series. (2) Morris found that even with his successful "verbal" (ESP) method, the best results occurred within relatively short sequences (100 digits) after the entry point. If we assume that large sequences following an entry point are less likely to show significant biases, it is important to note that in Series I the entry point produced *1500* targets and in Series II the entry point produced *5200* targets. (3) The form of target nonrandomness that the entry point would have needed to uncover could not be the simple odd-versus-even type tested by Morris and Harley and Sargent but would have had (a) to relate to subjects' response biases and (b) to occur only in those blocks of 25 digits (the length of a run) that corresponded to the runs responsible for the significant results, given our method of dividing data among conditions. In other words, the nonrandomness would have had to occur in arbitrarily scattered blocks of 25 digits following the entry point. Though we cannot conclusively rule out the existence of such a bias—and the hypothesis that by PK on the entry-point process DHW located just those sections showing it—this interpretation strikes us as unparsimonious, resting as it does on ad hoc assumptions about the composition of the RAND table.

Observer effect. We tend to favor the third interpretation, that the checker effects in this study occurred as a function of the *observational process* that checking represents. As Broughton (1977) has demonstrated, the observational-theory model of ESP can be tested by seeing whether a manipulation applied to the observer at the moment of observation has an effect on the experimental outcome. Varying whether checkers knew subjects' predictions during checking was just this sort of manipulation, and, because it did affect the results, our study adds support to the observational theories' model of ESP.

The observer interpretation is further supported by our use of different strategies during checking. DHW recalls that while checking the data of Series I, for example, she played a mental game of noting whether that run had been predicted for her and making some mental comment such as "This one's mine!" or playfully cheered one prediction category and "booed" another. In Series II, she used a similar strategy, though less consistently. (On some occasions she neglected to take note of the subject's prediction until after she had finished checking the run.) Still, her awareness of prediction categories during checking seems to have been greater than that of NLZ, who reports that she paid relatively little attention to this information. This greater awareness may explain why in both

series DHW found significant differences between prediction categories whereas NLZ obtained nearly equal scores, and why in Series II DHW found a significant difference between prediction categories only when she had access to this information. This point can be no more than a casual comment, because we did not formally evaluate or manipulate checkers' strategies. But it does suggest that state-of-mind factors in the checker *during observation of the call-to-target match* may play a role in determining the outcome of the match.

Checker Knowledge of Subjects' Predictions: Implications for the Observational Theories

Even if the subjects are the psi sources and are precognizing the checkers' experience, the C × P × B interaction suggests that at least one aspect of that experience—checkers' knowledge of subjects' predictions—was necessary for psi to have taken place in our experiment. What is it about this knowledge that is important? We consider here two possibilities: (1) this knowledge made the observation of the call-to-target match *meaningful*; and (2) this knowledge made the observation of the call-to-target match *complete*.

Meaningful context of observation. Knowing for whom the run had been predicted may have made the context of observing the call-to-target match more meaningful than the same observation made without this knowledge. This is not exactly the same thing as von Lucadou and Kornwachs' (1980a) concept of *meaningful observation.* In our case, the crucial observation (the call-to-target match) is meaningful in von Lucadou and Kornwachs' terms under both blind and not-blind conditions: Both checkers understood the meaning of a correct or incorrect match and its evaluation (whether the total number of hits is high or low compared with the expected value), which von Lucadou and Kornwachs (1983) imply is an important component of meaningful observation. This understanding does not change when we are blind to subjects' predictions. What *does* change is the meaning of the call-to-target match for the larger goal of confirming the experimental predictions. Here we can focus on the experience of the checker who obtained significant differences between prediction categories. DHW recalls that while checking both the blind and not-blind data of Series II, she had an interest in finding consistent differences between the scores of each subject's yes-no run pair. When she was not blind to subjects' predictions she could of course see whether the differences were consistent (i.e.,

"no" runs > "yes" runs) and did make mental note of this. Additionally, after she noticed the trend in the not-blind data toward lower scores in runs predicted for NLZ, she came to expect this trend as she continued through the data. These interests, intentions, or expectations were necessarily of a weaker or less complete sort when checking blind data. She might have wished to see a difference between run scores within a subject's pair, but without knowing which category each run belonged to, she could not have a clear intention of *which* run she would want to see have a high score.

Of course, significant psi effects can occur in the absence of clear intentions and expectations, as we know from the literature on nonintentional psi (e.g., Schechter, 1977; Stanford, Zenhausern, Taylor, & Dwyer, 1975). But we should not over-generalize that research. It may well be that for certain individuals or in certain situations a clearly formed intention is necessary for psi to occur.

Obviously, the topic of meaningfulness is an exceedingly complex one; here, we can only hope to raise it in the most superficial way. For example, we described earlier DHW's observational strategy. Did her ascription of ownership in some cases ("This one's mine!") give the run a *personal* meaningfulness that could not exist when she did not know if the run had been predicted for her? Was there something meaningful about our not-blind runs being observed under conditions that were more similar to those of Feather and Brier's experiments and Series I? These are only some of the possibilities raised by the concept of a meaningful context.

Complete observation. Although consistent with other types of observational theories as well, our second argument follows from von Lucadou and Kornwachs's system-theoretical approach to psi (Kornwachs & von Lucadou, 1979; von Lucadou & Kornwachs, 1977a, 1977b, 1980a, 1980b, 1983). Under their model, our experiment can be thought of as a system consisting of various elements (e.g., subjects, calls, targets, checkers). The complexity of this system changes over time as certain procedures are carried out and add new levels in the system's hierarchical structure. (For example, comparing calls to targets adds the new level "run scores.") Von Lucadou and Kornwachs argue that a complex system, by virtue of its inability to be completely described, will display a certain inherent unpredictability in its behavior and that this unpredictability is analogous on a macrocosmic scale to the Heisenberg uncertainty principle in the prediction of electron behavior on a microcosmic scale. Thus, under this view we can discuss the "observational histories" of the blind and not-blind runs in Series II according to how various

observations reduced the uncertainty of the system without requiring that this uncertainty be quantum mechanically indeterminate.

Let us consider, then, the observational histories of blind and not-blind runs as they were processed through the system (Series II) from the time of first observation (checking) until the data were analyzed. For the "final output" of the system (i.e., the ANOVA), there were four important pieces of information that needed to be obtained (observed) for each run: (1) Who checked the run? (2) Was it checked under blind or not-blind conditions? (3) Was it predicted for NLZ or not? and (4) What was the run score? These four items of information "fixed" the composition of the cells of the ANOVA. Once these cells were fixed, there was no longer any uncertainty in the output of the system because the mathematical procedures for calculating the ANOVA and the p values associated with the F ratios are deterministic.

When each checker observed her data, the answer to the first question was constant; only the remaining three items of information varied from run to run. When runs were checked under not-blind conditions, these three items were obtained at approximately the same time and during the same procedure. On the other hand, when runs were checked under blind conditions, the checkers obtained only two pieces of information (blind or not blind and run score). The remaining item (prediction category) was observed at a later stage when the experimenters obtained the assistant's ordering scheme and identified the runs predicted for NLZ and those predicted for the anonymous checker. In other words, the degree of uncertainty reduced by the observation of call-to-target matches under not-blind conditions was greater than that reduced under blind conditions. More important, for the not-blind runs complete reduction of uncertainty (i.e., observation of the three variable items of information) occurred within a relatively short time and as part of a single procedural step: checking the calls. For the blind runs, on the other hand, complete reduction of uncertainty occurred in *two* steps (checking and later decoding into yes and no runs) that took place at separate times and possibly under different psychological and observational conditions. These factors may explain why the checker effect was not found in the blind data.

Conclusions

So little checker-effect research has been conducted that a firm assessment of the robustness of our results cannot be made. Still, an

overview of this research shows the following: Of the 11 data sets[8] that tested for a main effect of the checker, significant differences were found in four (pilot series of Feather & Brier, 1968; Series II of O'Brien, 1979; Series I and the not-blind data of Series II of this report). A marginally significant difference was found in an additional experiment (Series I of O'Brien's study, 1979). Of the nine data sets testing for a checker effect in an interaction between actual checker and predicted checker, three were significant (the pilot and confirmation studies of Feather & Brier, 1968; Series II, not-blind data of this study). These trends are at least promising and suggest that the checker may play an influential role in precognition experiments. But as Kennedy has commented (personal communication, August 5, 1985), experiments on checker effects are more like tests of special subjects than like tests of unselected volunteers and therefore may produce highly idiosyncratic results and be particularly hard to replicate and generalize.

All checker-effect research to date has been done only with precognition tasks. This is not to say that the checker effect cannot occur in present-time ESP studies, merely that this has not yet been tested. But if it turns out that the checker plays a role in precognition studies but not in present-time ESP studies, the checker effect may have implications for Tart's observation of generally poorer scoring in precognition than in present-time ESP research (Tart, 1983). That is, unmeasured and idiosyncratic factors related to the checker (e.g., the checker's identity, psychic ability, and approach to the task, the information he or she has during checking) may be detrimental to successful results in precognition experiments, whereas the same factors would have no effect in present-time ESP experiments. As Giesler (personal communications, October 30, 1985, and February 16, 1986) has pointed out, if this is so it may be due either to an inherent difference between precognitive and present-time ESP or to observational factors (e.g., the impact of the checker may be especially reduced in present-time ESP experiments in which the subject receives immediate feedback and therefore is the first observer).

The present results lend support to the idea that the checker effect is an observer effect, as modeled by the observational theo-

[8] The 11 data sets are: Feather and Brier (1968), two series; Kennedy et al. (1977), four series; O'Brien (1979), two series; and this report (Series I, Series II/blind data, Series II/not-blind data). O'Brien did not have subjects predict the checker; therefore, her two series were not included when assessing evidence for a C × P interaction.

ries. This conclusion, however, is tempered by the results of the secondary analysis of Series II, which suggested that the subject also played an active role in producing these results. Giesler (personal communication, October 30, 1985) has noted that the "checker" effects in the present research may not be due simply to the psi of either subjects or checkers alone but instead be due to some combination or interaction of the two. Future research might test the relative contributions of subjects and checkers by substituting randomly determined "pseudocalls" for those of real subjects (e.g., Bierman, 1978) if this can be done with checkers blind to the manipulation and with all other conditions remaining equal.

Our work suggests that the context under which calls are matched to targets may be important. We have discussed two potentially important aspects of that context: its *meaningfulness* and its ability to make the observation *complete*. This discussion begets a host of research questions. For example, how might the checker effect be changed if the meaningfulness of the observation is altered? What would happen if "complete" observation takes place over a number of stages when each stage has precisely defined and manipulated conditions? The answers to these questions may help parapsychologists address the important question of how (and from whom) significant results in experimental psi testing are created.

References

BIERMAN, D. J. (1978). Observer or experimenter effect? A fake replication. *Journal of Parapsychology*, **42**, 55–56. (Abstract)

BIERMAN, D. J. (1985, January). *Physics and parapsychology*. Paper presented at the joint conference of the Parapsychological Association and Andhra University, "Parapsychology: Eastern and Western Perspectives," Waltair, India.

BIERMAN, D. J., & HOUTKOOPER, J. M. (1981). The potential observer effect, or the mystery of irreproduceability. *European Journal of Parapsychology*, **3**, 345–372.

BIERMAN, D. J., & WEINER, D. H. (1982). Multiple preobservation of data as a method for screening off future-observer effects. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology 1981* (pp. 132–134), Metuchen, NJ: Scarecrow.

BIERMAN, D. J., HOUTKOOPER, J. M., & MILLAR, B. (1981, April). *Triple observation of the observational theories*. Paper presented at the Fifth International Society for Psychical Research Conference, Bristol, England.

BRAUDE, S. (1979). The observational theories in parapsychology: A critique. *Journal of the American Society for Psychical Research*, **73**, 349–366.

BROUGHTON, R. S. (1977). An exploratory study on psi-based subject and experimenter expectancy effects. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 173–177). Metuchen, NJ: Scarecrow Press.

CARPENTER, J. C. (1983). 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.

EISENBUD, J. (1963). Psi and the nature of things. *International Journal of Parapsychology*, **5**, 245–273.

FEATHER, S. R., & BRIER, R. (1968). The possible effect of the checker in precognition tests. *Journal of Parapsychology*, **32**, 167–175.

FISK, G. W., & WEST, D. J. (1958). Dice-casting experiments with a single subject. *Journal of the Society for Psychical Research*, **39**, 277–287.

FREEMAN, J. A. (1968). Sex differences and primary mental abilities in a group precognition test. *Journal of Parapsychology*, **32**, 176–182.

GEORGE, L. (1981). *Psi as a universal homeostatic function: Toward a testable theory*. Unpublished manuscript.

HARLEY, T. A., & SARGENT, C. L. (1980). Can experimenters localize non-random target sequences in random number tables? *European Journal of Parapsychology*, **3**, 247–252.

HARTWELL, J. (1977). A bound for the observational theories of psi. *European Journal of Parapsychology*, **2**(1), 19–28.

HAYS, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, & WINSTON.

HOUTKOOPER, J. J. (1982). Methodological aspects of space-time independence in the observational theories. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology 1981* (pp. 66–68). Metuchen, NJ: Scarecrow.

HOUTKOOPER, J. M. (1983). *Observational theory: A research programme for paranormal phenomena*. Lisse, Holland: Swets & Zeitlinger, B. V.

HOUTKOOPER, J. M., & HARALDSSON, E. (1985). Experimenter effects in a plethysmographic ESP experiment. *European Journal of Parapsychology*, **5**, 313–326.

KENNEDY, J. E., & TADDONIO, J. L. (1976). Experimenter effects in parapsychological research. *Journal of Parapsychology*, **40**, 1–33.

KENNEDY, J. E., O'BRIEN, J. T., O'BRIEN, D., & KANTHAMANI, H. (1977). *An attempt to replicate the checker effect*. Unpublished manuscript. Abstracted in *Journal of Parapsychology*, **41**, 377–378.

KEPPEL, G. (1973). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.

KORNWACHS, K., & VON LUCADOU, W. (1979). Psychokinesis and the concept of complexity. *Psychoenergetic Systems*, **3**, 327–342.

MILLAR, B. (1978). The observational theories: A primer. *European Journal of Parapsychology*, **2**, 304–332.

MILLAR, B., & HARTWELL, J. (1979). Dealing with divergence. In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 91–93). Metuchen, NJ: Scarecrow.

MORRIS, R. L. (1968). Obtaining non-random entry points: A complex psi task. In J. B. Rhine & R. Brier (Eds.), *Parapsychology Today* (pp. 75–86). New York: Citadel.

O'BRIEN, J. T. (1979). An examination of the checker effect. In W. G. Roll (Ed.), *Research in parapsychology 1978* (pp. 153–155). Metuchen, NJ: Scarecrow.

PHILLIPS, P. R. (1984). Measurement in quantum mechanics. *Journal of the Society for Psychical Research*, **52**, 297–306.

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

RAO, K. R. (1977). On the nature of psi: An examination of some attempts to explain ESP and PK. *Journal of Parapsychology*, **41**, 294–351.

RAO, K. R. (1978). Psi: Its place in nature. *Journal of Parapsychology*, **42**, 276–303.

RHINE, J. B., & PRATT, J. G. (1957). *Parapsychology: Frontier science of the mind*. Springfield, IL: Charles C. Thomas.

SCHECHTER, E. I. (1977). Nonintentional ESP: A review and replication. *Journal of the American Society for Psychical Research*, **71**, 337–374.

SCHMIDT, H. (1974). A new role of the experimenter in science suggested by parapsychology research. In A. Angoff & B. Shapin (Eds.), *Parapsychology and the Sciences* (pp. 266–279). New York: Parapsychology Foundation.

SCHMIDT, H. (1975). Toward a mathematical theory of psi. *Journal of the American Society for Psychical Research*, **69**, 301–319.

SCHMIDT, H. (1984). Comparison of a teleological model with a quantum collapse model of psi. *Journal of Parapsychology*, **48**, 261–276.

SCHMIDT, H. (in press). Human PK effect on prerecorded targets previously observed by goldfish. In D. H. Weiner & D. I. Radin (Eds.), *Research in Parapsychology 1985*. Metuchen, NJ: Scarecrow.

STANFORD, R. S., ZENHAUSERN, R., TAYLOR, A., & DWYER, M. A. (1975). Psychokinesis as psi-mediated instrumental response. *Journal of the American Society for Psychical Research*, **69**, 127–133.

TART, C. T. (1983). Information acquisition rates in forced-choice ESP experiments: Precognition does not work as well as present-time ESP. *Journal of the American Society for Psychical Research*, **77**, 293–310.

VON LUCADOU, W., & KORNWACHS, K. (1977a). Can quantum theory explain paranormal phenomena? In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology 1976* (pp. 187–191). Metuchen, NJ: Scarecrow.

VON LUCADOU, W., & KORNWACHS, K. (1977b, September). *Fitting of PK results with a quantum mechanical model?* Paper presented at the 1977 Parascience Conference, London, England.

VON LUCADOU, W., & KORNWACHS, K. (1980a). Development of the system theoretic approach to psychokinesis. *European Journal of Parapsychology*, **3**, 297–314.

VON LUCADOU, W., & KORNWACHS, K. (1980b, September). *On the structure of observational theories.* Paper presented at the 1980 Parascience Conference, London, England.

VON LUCADOU, W., & KORNWACHS, K. (1983). On the limitations of psi: A system-theoretic approach. *Psychoenergetics*, **5**, 53–72.

WALKER, E. H. (1975). Foundations of paraphysical and parapsychological phenomena. In L. Oteri (Ed.), *Quantum physics and parapsychology* (pp. 1–44). New York: Parapsychology Foundation.

WALKER, E. H. (1977). Comparison of some theoretical predictions of Schmidt's mathematical theory and Walker's quantum mechanical theory of psi. *Journal of Research in Psi Phenomena*, **2**(1), 54–70.

WALKER, E. H. (1984). A review of criticisms of the quantum mechanical theory of psi phenomena. *Journal of Parapsychology*, **48**, 277–332.

WEINER, D. H. (1982, February). *The once and future psi source: A review of the literature on future-observer effects.* Paper presented at the Southeastern Regional Parapsychological Association Conference, Chapel Hill, NC. Abstracted in *Journal of Parapsychology*, **46**, 58–59.

WEINER, D. H., & BIERMAN, D. J. (1979). An observer effect in data analysis? In W. G. Roll (Ed.), *Research in parapsychology 1978* (pp. 57–58). Metuchen, NJ: Scarecrow.

WEINER, D. H., & BIERMAN, D. J. (1982). Toward a definition of observation: A test of the effects of information specificity and meaningfulness on PK. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in parapsychology 1981* (pp. 134–135). Metuchen, NJ: Scarecrow.

WEINER, D. H., & GELLER, J. (1984). Motivation as the universal container: Conceptual problems in parapsychology. *Journal of Parapsychology*, **48**, 27–37.

WEST, D. J., & FISK, G. W. (1953). A dual ESP experiment with clock cards. *Journal of the Society for Psychical Research*, **37**, 185–197.

WHITE, R. A. (1976a). The influence of persons other than the experimenter of the subject's scores in psi experiments. *Journal of the American Society for Psychical Research*, **70**, 133–166.

WHITE, R. A. (1976b). The limits of experimenter influence on psi test results: Can any be set? *Journal of the American Society for Psychical Research*, **70**, 333–369.

WINER, B. J. (1962). *Statistical principles in experimental design.* NY: McGraw-Hill.

ZINGRONE, N. L., & WEINER, D. H. (1984, August). *The checker effect revisited.* Research brief presented at the twenty-seventh Parapsychological Association Convention, Dallas, TX.

Institute for Parapsychology
Box 6847, College Station
Durham, NC 27708