SUMMARIZING RESEARCH FINDINGS: META-ANALYTIC METHODS AND THEIR USE IN PARAPSYCHOLOGY

CHARLES HONORTON

What is the value of a scientific research literature? We know how to evaluate the outcomes of individual studies, but what can we conclude regarding an entire research domain?

This question is important from a variety of perspectives. Policy-makers, funding agencies and research administrators all need the most complete information possible in order to make informed decisions regarding allocation of limited resources. For the scientific community, informed evaluation of new knowledge claims requires integration of all the available evidence. Investigators pursuing basic research, as well as those seeking reliable applications, need to realistically assess previous findings, suggestions regarding particularly successful approaches, the degree to which specific research practices may lead to unreliable outcomes, and so on. The caveat "Further research is necessary," is nearly always true, but what can we learn from all of the research that already exists?

Meta-analysis applies the methods of data analysis to the assessment of findings across all available studies in a given research domain. Regardless of the specific intent of any meta-analysis, there are two general activities common to all meta-analytic investigations, *cumulation* and *blocking*. Cumulation addresses the question "Is there an effect and, if so, how strong is it?" It involves assessing the statistical significance and magnitude of the effect under study, and the extent to which the cumulative effect is vulnerable to the selective reporting of "significant" results. Blocking subdivides the research domain on the basis of differences across studies that might account for their variability. By coding variations in procedures, subject populations, stimulus conditions, etc., the meta-analyst can address a variety of important questions such as: "Is the effect systematically related to study quality?" "How robust is the effect?" "Can we identify variations in procedures, subject populations, stimulus conditions, etc., that are particularly successful, as-

sociated with especially strong or reliable outcomes, or which have been consistently unproductive?"

Meta-analytic procedures are described in a number of excellent volumes (e.g., Cooper, 1984; Glass, McGaw & Smith, 1981; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Rosenthal, 1984). Investigators wishing to embark on meta-analytic investigations should consult these sources, especially the Rosenthal and Cooper texts. Light and Pellemer provide a particularly readable discussion of meta-analysis that will appeal especially to those interested in the value of research integration for policy-making.

Meta-analysis of Experimental Precognition

A meta-analysis of forced-choice precognition experiments will serve to illustrate some of the major characteristics of meta-analytic investigations. This meta-analysis was performed by Diane Bailey, George Hansen, and myself at PRL, under a subcontract from SRI International. In the limited time available for this presentation, my discussion of this study will be limited to an overview of its essential features. As such, it represents somewhat of an oversimplification, and I will avoid discussion of some of the complexities inherent in any study of this type. A more detailed report is in preparation for publication elsewhere. My purpose at this time is simply to concretize the meta-analytic process for those of you who are not already familiar with it and to convey, through this example, some appreciation of its power.

We addressed four major questions through our meta-analysis of the precognition literature:

1. Is there overall evidence for accurate target identification (i.e., above chance scoring) in experimental precognition studies?
2. What is the magnitude of the overall (directional and predicted) precognition effect?
3. Is the observed precognition effect related to variations in methodological quality that could pose serious threats to validity?
4. Does precognition performance vary systematically with potential moderating variables, such as differences in subject populations, stimulus conditions, experimental setting, knowledge of results, and temporal distance?

Delineating the Domain

Source of Studies and Criteria for Inclusion. The source of studies was restricted to forced-choice precognition studies published in the peer-

reviewed English-language parapsychology journals: *Journal of Parapsychology Journal (and Proceedings) of the SPR, Journal of the ASPR, European Journal of Parapsychology* (including the *Research Letter* of the Utrecht University Parapsychology Laboratory) and *Research in Parapsychology*. We restricted our review to studies in which significance levels and effect sizes based on direct hitting could be calculated. Reports using outcome variables other than direct hitting, such as run-score variance, displacement, etc., were included only if they provided relevant direct hits information (i.e., number of trials, hits, and probability of a hit). We also excluded studies by two investigators, S. G. Soal and Walter J. Levy, whose work has proven to be unreliable. Many published reports contained more than one experiment or experimental unit. Experiments involving multiple conditions were treated as separate study units.

General Characteristics of the Domain. We located 309 studies in 113 separate publications. These studies were contributed by 62 different senior authors and were published over a 52-year period, between 1935 and 1987. Considering the half-century time-span over which the precognition studies have been conducted, it is not surprising that the studies are quite diverse. The data base comprises nearly 2 million individual trials and more than 50,000 subjects. Study sample sizes range from 25 to 297,060 trials with a median of 1194 trials. The number of subjects ranges from 1 to 29,706 with a median of 16 subjects. The precognition domain encompasses a diverse range of subject populations. Student populations comprise the largest grouping (approximately 40%), while studies with the experimenter as subject and animal studies comprise the smallest groupings (each representing about 5% of the studies).

Outcome Measures. Significance Levels: We calculated two significance estimates for each study. The *directional z-score* (Zdir) measures the subjects' success in scoring in the direction of their intention.

Effect Sizes: Significance levels are a function of sample size and comparisons based on raw significance levels can be very misleading. Consider a hypothetical example. Investigator A reports a ganzfeld study with 100 trials and a hit-probability of .25. She obtains 33 hits, a conventionally-significant result ($z = 1.7$, $p = .045$). An attempted replication by investigator B yields 11 hits in 33 trials. Since B's result is not significant ($z = 0.91$, $p = .18$), he concludes that he has failed to replicate A's results. B's conclusion is incorrect; his scoring rate (33% hits) is identical to A's. Even the significance levels of the two studies are not that different: ($z = 1.70.91\sqrt{2} = 0.56$, $p = .288$).

Thus, it is useful to have a basis for comparing study outcomes that

is independent of the study sample sizes. Most parapsychological experiments, particularly those in the older literature, use the trial rather than the subject as the sampling unit. It is necessary to use a trial-based effect size estimator in such cases. In the precognition meta-analysis, for example, we use an effect size for each study that is the z-score divided by the square root of the number of trials in the study.

Overall Cumulation. As shown in the top part of Table 1, the overall results are highly significant. There is strong evidence for overall directional hitting. Thirty percent of the studies show overall significant hitting at the 5% level.

Lower bound confidence estimates of the mean z-score displayed in the bottom portion of Table 1 indicate that the mean z-score is well above zero at the 95% confidence level.

As indicated earlier, significance levels are related to sample size and it is therefore not surprising that z's correlate positively with sample size. The correlation (r) is 0.156 z's (307 df, p =.003).

The effect size analysis is presented in Table 2. The directional outcome is significantly above zero.

Replicability across Investigators. Virtually the same picture emerges when the cumulation is by investigator rather than study. The combined z is 12.31. Twenty-three investigators (37%) had directional outcomes significant at the 5% level. The mean (investigator) effect size is .028 ± .091.

These results indicate a substantial level of cross-investigator replicability and directly contradict the claim of critics such as Akers (1988) that successful parapsychological results are achieved by only a small handful of investigators.

The Filedrawer Problem. There is a well-known reporting bias throughout the behavioral sciences, favoring publication of "significant" studies (e.g., Sterling, 1959). The exteme view of this "filedrawer problem," as Robert Rosenthal describes it, "is that the journals are

TABLE 1

Precognition Significance Levels

	Z
Mean	0.65
Standard Deviation	2.68
Combined (Stouffer) z	11.41
p,	6.3×10^{-25}
Filedrawer Estimate	14,268
Lower 95% Confidence Estimate of Mean	.40

TABLE 2

Precognition Effect Sizes

	ES
Mean	.020
Standard Deviation	.100
$t(308)$	3.51
p_t	.00025
Lower 95% Confidence Limit	.011

filled with the 5% of the studies that show type I errors, while the filedrawers back at the lab are filled with the 95% of the studies that show nonsignificance . . ." (Rosenthal, 1984, p. 108). Recognizing the importance of this problem, the Parapsychological Association in 1975 adopted an official policy against selective reporting of "positive" results. Even the most cursory examination of the parapsychological literature will show that nonsignificant results are frequently published and in the precognition database, 60% to 70% of the studies reported nonsignificant results. Nevertheless, 75% of the precognition studies were published prior to 1975, when the Parapsychological Association formulated its policy, and it is necessary to ask to what extent selective publication bias could account for the cumulative effects we observe.

The central section of Table 1 uses Rosenthal's (1984) filedrawer statistic to estimate the number of unreported studies with z-scores averaging zero that would be necessary to reduce the known database to nonsignificance. The filedrawer estimate suggests that there would need to be over 46 unreported studies for each reported study in order to reduce the cumulative hitting (directional) outcomes to a nonsignificant level.

Based on this analysis, we conclude that it is implausible that the cumulative significance of the precognition studies is due to selective reporting.

Study Quality. While precognition experiments are not usually vulnerable to sensory leakage problems, there are a number of other potential threats to validity that must be taken into account. Statistical and methodological variables are defined and coded in terms of procedural descriptions (or their absence) in the research reports. One point is given (or withheld) for each of the following criteria:

Specification of Sample Size. Did the investigator preplan the number of trials to be included in the study or was the study vulnerable to the possibility of optional stopping? Credit was given to reports which explicitly specified the sample size. Studies involving group testing, in

which it was not feasible to precisely specify the sample size, were also given credit. No credit was given to studies in which the sample size was either not preplanned or not addressed in the experimental report.

Preplanned Analysis. Was the method of statistical analysis, including the outcome (dependent variable) measure, preplanned? Credit was given to studies explicitly specifying the form of analysis (and the outcome measure). No credit was given to those not explicitly stating the form of the analysis or those in which the analysis was clearly post-hoc.

Randomization Method. Credit was given for use of random number tables, random number generators, or mechanical shufflers, but not for hand shuffling, die casting, or drawing lots.

Controls. Credit was given to studies reporting randomness control checks, such as RNG control series and empirical cross-check controls.

Recording. One point was allotted for use of automated recording of targets and responses and another for duplicate recording.

Checking. One point was allotted for use of automated checking of matches between target and response and another for duplicate checking of hits.

Each study received a quality weight between zero and eight. We found no overall relationship between study quality and effect size ($r307 = -.062$, $p = .279$). Of the eight quality measures, controls and duplicate recording correlated signficantly positively with effect size and randomization correlated significantly negatively. Eighty percent of the studies (N = 247) used adequate methods of randomization and the Stouffer z's for these studies alone remain highly significant ($z = 5.49$).

It has long been believed by critics of parapsychology that psi disappears as methodological rigor increases. The precognition database provides no support whatsoever for this belief. Precognition effect sizes have remained relatively constant over a half-century of research, even though the methodological quality of the research has improved significantly. The correlation between effect size and date of publication is $-.064$. Study quality and date of publication are, however, positively and significantly correlated ($r307 = .266$, $p < 10^{-5}$).

Moderating Variables.[2] The stability of precognition study outcomes over a 50-year period is also of course bad news. It indicates that we have not yet developed sufficient understanding of the conditions un-

[2] Throughout this report, t-test comparisons involving unequal variances are computed using the separate within groups variance for the error (Wilkinson, 1988) and degrees of freedom following Brownlee (1965).

derlying the occurrence (or detection) of these effects to reliably increase their magnitude. Can meta-analysis help? I believe the answer is "Yes." Our precognition meta-analysis has identified a number of variables that appear to covary systematically with magnitude of precognitive performance. I will briefly discuss three:

1. Selected versus unselected subjects;
2. Individual versus group testing;
3. Feedback Level.

Selected vs. Unselected Subjects. Precognition studies using subjects selected on the basis of prior performance show larger effects than studies with unselected subjects. Two-thirds of the studies with selected subjects are significant at the 5% level. Indeed, the *mean* directional *z*-score for these studies is 2.37 (sd = 3.42). The basis of selecting subjects was the subject's performance in a previous experiment or in pilot tests. As shown in Table 3, the magnitude of effect size is significantly higher for selected subjects studies than for studies with unselected subjects. The t test of the difference in mean effect size is equivalent to a point-biserial correlation of .48.

Is this difference due to less stringent controls in studies with selected subjects? The answer appears to be "No." The average quality of studies with selected subjects is in fact significantly higher than studies using unselected subjects ($t52 = 2.57$, $p = .013$).

Individual versus Group Testing. Studies in which subjects are tested individually by an experimenter have a significantly larger mean effect size than studies involving group testing (Table 4). The t test of the difference is equivalent to a point-biserial correlation of .234, favoring individual testing. Forty-one percent of the studies with subjects tested individually are significant at the 5% level. The methodological quality of studies with subjects tested individually is significantly higher than in studies involving group testing ($t201 = 3.57$, $p = .00022$).

Feedback. There is a significant positive relationship between the degree of feedback subjects receive concerning their performance and precognitive effect size (Table 5).

TABLE 3

Selected vs. Unselected Subjects

Subjects	N Studies	Mean ES	SD
Selected	44	0.096	0.147
Unselected	265	0.010	0.082

$t_{47} = 3.76$, $p = .00023$

TABLE 4

Individual Versus Group Testing

Test Setting	N Studies	Mean ES	SD
Individual	134	0.045	0.111
Group	123	−0.001	0.077

$t_{255} = 3.85$, $p = .000075$

Subject feedback information is available for 139 of the studies. These studies fall into four feedback categories: No feedback, delayed feedback (usually via notification by mail), run-score feedback, and trial-by-trial feedback. For analysis purposes, these categories were given numerical values between 0 and 3. Directional precognition effect size correlates .228 with feedback level (137 *df*, $p = .0035$). Of the 67 studies involving trial-by-trial feedback, 49% were significant at the 5% level, while only 1 of the 18 studies with no subject feedback (5.6%) was significant. Degree of feedback correlates positively, though not significantly, with research quality ($r137 = .124$, $p = .145$).

Precognitive Time Span. We have examined a number of other factors that vary across studies including differences among subject populations, target variations, etc. Space (and time) limitations require that a full presentation of our findings in this area be postponed for another occasion. However, there is one other question that I will discuss now, because of its intrinsic interest, and because it illustrates an important limitation of meta-analysis.

What, if anything, can we say regarding the temporal range of precognitive functioning? We attempted to address this question through analysis of the 190 studies which provide information concerning the interval between the subject's response and the determination of the target. Since the information provided was usually not very precise, our analysis was limited to seven broad temporal categories: milliseconds, seconds, minutes, hours, days, weeks, months.

TABLE 5

Subject Feedback of Results

Feedback Level	N Studies	Stouffer z	Mean ES	SD_{ES}	SIG.5%
No Feedback	18	−1.41	−0.002	0.044	5.6%
Delayed	24	1.83	0.010	0.038	25.0%
Run score	30	13.25	0.039	0.084	46.7%
Trial-by-trial	67	12.87	0.063	0.136	49.3%

The reported intervals range from a few milliseconds to one year and we did find a marginally significant decline in precognitive effect size across these seven temporal intervals ($r188 = -.131$, $p = .036$). To complicate matters, the relationship interacts with the selected/unselected subjects difference. The negative relationship between performance and temporal distance, to the extent that it exists at all, appears to be restricted to studies involving unselected subjects. These studies show a much larger negative relationship than the database as a whole ($r149 = -.211$, $p = .009$), and studies with selected subjects show a nonsignificant *positive* correlation between performance and time interval ($r37 = 0.84$). The difference between these correlations approaches significance ($z = 1.92$, $p = .054$).

Unfortunately these findings cannot be taken very seriously because the precognitive interval is systematically related to both study quality and degree of feedback. Studies using automated testing methods, for example, generally received higher quality ratings and were much more likely to be associated with trial-by-trial feedback than studies involving longer precognitive time spans.

Such confounds are inevitable in meta-analytic investigations. Meta-analysis provides an important and valuable method of summarizing an existing research literature, but it is not a substitute for new experiments.

What the Data Tell Us. Returning to the five basic questions we asked at the beginning of this exercise, what has the meta-analysis told us about experimental precognition effects?

Is there overall evidence for accurate target identification (i. e., above chance scoring) in experimental precognition studies? Yes. The cumulative results cannot reasonably be attributed to chance fluctuation. Independently significant outcomes are shown in 30% of the studies and by 39% of the investigators and these outcomes cannot plausibly be attributed to selective reporting of positive results.

What is the magnitude of the overall precognition effect? The effects are small. While we knew that before the meta-analysis, we now have some indication of their actual magnitude. Our confidence estimates indicate, for example, that the average magnitude of significance level is at least one-third of a standard deviation.

Is the observed precognition effect related to variations in methodological quality that could pose serious threats to validity? There is no overall relationship between study quality and effect size. Of the individual quality criteria, two correlate significantly positively with outcome and only one (method of randomization) correlates significantly negatively with

outcome. Even if we discard studies with nonoptimal randomization, the results remain highly significant.

Does precognition performance vary systematically with potential moderating variables? Yes. We have identified three correlates of precognitive performance. Subjects selected on the basis of prior achievement show significantly larger effect sizes than unselected subjects, individually tested subjects perform at significantly higher levels than those tested in group settings, and precognitive achievement covaries with the degree of feedback provided to the subjects.

Should these meta-analytic findings be regarded as conclusive? Probably not, but they provide a much richer and better-informed foundation upon which to base future research than we had before. And that, in my opinion, is the value of meta-analysis.

REFERENCES

Akers, C. (1987). Parapsychology is science, but its findings are inconclusive. *Behavioral and Brain Sciences, 10,* 566–568.

Brownlee, K A. (1965). *Statistical theory and methodology in science and engineering.* New York: John Wiley & Sons.

Cooper, H. M. (1984). *The integrative research review: A social science approach.* Beverly Hills, CA: Sage.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Light, R. L., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Wilkinson, L. (1988). *SYSTAT: The system for statistics.* Evanston, IL: SYSTAT, Inc.

DISCUSSION

RAO: Chuck, I think you have a very interesting paper here and we feel very reassured that the research effort of over half a century has produced some solid robust effect attesting the reality of precognition. I would like to make a couple of comments. First of all, it is very interesting to see that in spite of the variation in test techniques and different subject groups, experimenters and laboratories, we seem to have a consistent effect. It is reassuring indeed that the past effort is greatly reinforced by the fact that much improved methodology and greater control seen in current research have essentially given the same

results as the original studies. The second point is with regard to the blocking analysis that you have done. It is possible that there may be some kind of a confounding here. Now the selected subjects did better than the unselected subjects. The individual testing was better than group testing. It is all very likely that the selected subjects were tested individually.

HONORTON: That is right.

RAO: And the unselected subjects were tested in groups so this could be a confounding variable. I can think of a number of others.

HONORTON: There are and I mentioned some in the written version of the paper.

RAO: There are other variables which could also account for some of these differences.

HONORTON: However, Ram, if you look at it in terms of a multiple regression analysis it is clear that the strongest of those three predictors is the selected subjects feedback.

RAO: It is possible that the differences are genuine and you may be right about the selected subjects feedback. But I have a lingering worry that a consideration of effect size without regard to the sample size may mislead us sometimes. Higher effect sizes obtained with smaller samples may have other explanations than the size of the sample itself, such as experimenter expectations, his effort and involvement. In any case I think we need further analysis and evidence to say one way or the other.

HONORTON: That is not true in the precognition database. In fact if you are talking about effect sizes rather than significance levels there should not be a significant relationship between effect size and sample size. In fact I do not know of any evidence at all to support . . .

RAO: That, I think, is really the paradox that I am talking about.

HONORTON: But what I am saying is that I do not think there is any empirical support for what you are saying.

RAO: Well, I think that is something we should look into seriously. I do not know if you have any data to show that it is not the case.

HONORTON: The data from this meta-analysis shows that there is no relationship between effect size and sample size. There is a significant relationship between significance level and effect size and that is what you would expect. In fact, if you look at the number of trials in a study in relation simply to whether the study was significant or not, there are almost three times as many trials in the significant studies as in the non-significant studies.

RAO: Again these are all confounded because of unselected subjects working under no feedback conditions and vice versa. With selected

subjects and individual testing it is more likely that the feedback is given right away. In group tests when you test 200 subjects in a classroom it is less likely the subjects are given immediate feedback. Again it is probable there are more subjects in group studies than in studies involving select subjects.

HONORTON: That is really irrelevant to the issue of relationship between effect size and sample size.

RAO: I do not think so. Let me explain. As you recognize, the significance of a result is a function of both effect size and the sample size. Hence both of them are important. For example, if you have an effect size of .1 with data on 200 subjects the result may be significant. But with just 10 subjects the same effect size may not be statistically significant. Therefore, the consistency of the effect size in these two samples does not lead us to have the same confidence about the genuineness of the effect in both the cases. So what I am saying then is that the reason why we would like to have probability estimates is to see whether this effect is not due to some kind of variability. Therefore, sticking only to effect size ignoring the sample size altogether is likely to leave room for making some errors. So I would suggest that we do not simply look at the effect size and throw away and disregard other factors.

HONORTON: Who has thrown away what? Who has thrown away something?

RAO: We have to keep in perspective the probability values as well. We just cannot speak about effect size alone.

HONORTON: Yes, but we can assess the probability values of the effect sizes. In other words, when you look at an effect size you are talking about the magnitude of the effect. RAO: Right.

HONORTON: Taking out, stripping out the effect of the sample size.

RAO: The magnitude of the effect is meaningful only when we have confidence in the reality and genuineness of the effect.

PALMER: I was going to make a somewhat similar point. Let me make it slightly differently.

HONORTON: I want to respond to another aspect of what Ram was saying first, if I may. Quite obviously, when you do a meta-analysis you are not doing an experiment. You take the data as they exist and you learn as much from them as you can. Obviously, there are confounding factors in any meta-analysis that can only be resolved through further new research. The whole purpose of the meta-analysis is to provide a more informed estimate of what the effect size is and what the conditions are that are most likely to be productive for your study.

PALMER: I have a bit of a problem about comparing effect sizes with

different sample sizes, and it has to do with the stability of the effect size. As an example, say you are doing an ordinary card test with a run of 25 trials. Someone gets 10 hits. That is equivalent to an effect size of 10. That happens from time to time in the lab and does not get anybody too excited. However, if you were to maintain an effect size of 7 over, say, several thousand trials, as in some of the old card guessing experiments, by any standard of evidentiality you would have something of much greater consequence, even though the effect size is smaller. I think a solution to that—and I really have to credit Jessica for this—is to report confidence intervals around the effect sizes. This would give the kind of information that would keep one from being misled about the stability of an effect size. In fact, I would almost go so far as to suggest it as a reporting requirement in our journals. We have recently had a reporting requirement that authors need to give means and standard deviations, which in a sense is the same thing. But I think confidence intervals may be even a better way of conveying the essential information, along with the effect size. You certainly cannot get a good characterization of the data simply by reporting p values.

My second point has to do with the sources of studies for meta-analyses of this type. From time to time I am amazed when I see reports of parapsychology experiments appearing in, particularly, psychology journals, places such as *Psychological Reports*. I don't think there are a large number of these, but, as you might expect, they are consistently negative, so they probably have a different mean as compared to the rest of the sample from the parapsychology journals. Particularly when you are dealing with something like precognition, which is a very broad and widely used procedure, it might be good to go into *Psychological Abstracts* and get some sources from there as well. The good news here, at least based on the studies in the psychology journals that I have seen, is that if you do a blocking in terms of the source of the study in relation to study quality, you can come up with results that I think would be rather flattering to us.

SCHOUTEN: What really amazed me was the results of the file drawer approach, where you find that on the average, you would need about 140 insignificant results for each study in your database which was significant to wipe out the overall significant effect of the meta-analysis. It might be unclear how I arrive at the number of 140. The explanation is: Chuck reports in his paper that ±14,000 non-significant studies would be needed to wipe out the significant overall results of his meta-analysis. His meta-analysis is based on the outcomes of over 300 actual experiments of which, however, only less than 100 were significant at the 5% level. So on the average it looks as if 14,000/100 = 140 non-

significant studies are needed for each significant one to make the meta-analysis non significant. I still do not understand that.

HONORTON: Averaging?

SCHOUTEN: Averaging. Now that is an amazingly high number. How can that be? I just don't understand how that is possible. My second question is just curiosity. When you report some results from these analyses, for instance, selected versus unselected subjects, what sort of importance do I have to attach to that? Would you rate it as equal as, for instance, if you had run an experimental study and had found a significant difference to that effect? Or do you consider it to be merely suggestive? I have no idea.

HONORTON: I consider it somewhere in between. I think it is very strongly suggestive, but certainly not conclusive. It is not conclusive because, as Ram pointed out, there are other variables that might interact and confound with it. We found that there is also a tendency for subjects who were tested individually to do better than those who were tested in groups. Selected subjects are usually tested individually. To give you another example which is in my paper, we were of course very interested to see whether there was any declining precognitive effect size over time, over intervals ranging from a few milliseconds to a year. And there is in fact a significant negative correlation between effect size and precognitive interval. However, this is also very strongly related both to feedback and to quality. The quality relationship, which is very strong, is opposite to what it would have to be in order to be a problem. That is that the strong results are in the shorter time intervals. Those studies also have the highest quality, but as for feedback there is simply nothing that we can really do about that. We have evidence that relatively immediate feedback is associated with much stronger results than delayed or no feedback. If somebody is doing precognition over a year, he is not getting feedback for a very long time. So I consider it to be basically an exploratory technique in terms of the process-oriented aspect. In terms of being able to say that there is overall evidence that something non-chance is going on, I would say that that is as close to conclusive as we can get. The filedrawer estimates and the overall significance levels simply are not compatible unless your basic starting point is that the likelihood of precognition is less than one in a million or something like that, which with precognition may not be that uncommon.

BROUGHTON: I have the feeling that all too often we do not accumulate our knowledge over generations of experimenters. I hope we are indeed learning, as your moderating variables analyses show, that there are patterns, and there are regularities which have direct practical

consequences for the way we design experiments. I hope these do start influencing the next generation of experimenters. I would like to take this opportunity to ask you what should the rest of us be doing to make this job easier? I mean as people who are going to be doing experiments and reporting them, what sort of things do you advise?

HONORTON: Let me say first that I agree with you. I have never liked the term parapsychology particularly, but by golly given the degree to which it has been unfairly attacked in recent years I wear the badge with great pleasure and honor. And another thing is that when you do these meta-analyses you begin to realize that we really are a community. There are half a dozen of you in this room whose research has contributed to the database. Now, the major source of frustration for anyone who does a meta-analysis in parapsychology is that, in spite of the fact that the results do not seem to correlate with methodological threats to validity in any of the meta-analyses that have been done so far, I am embarrassed by much of the literature in terms of the lack of sophistication in the way we report our findings. We call ourselves parapsychologists, but a Martian looking at our literature would think that the human aspect of this is very minimal—subjects 15 males and 14 females, college age, volunteers—the description of subjects alone is pathetic. Looking at the very vital, philosophically tremendously important issue of the reach of precognition over time, the fact that even in the more recent literature involving automated testing techniques we have not started to adopt some fairly precise way of talking about the interval between the subject's response and the selection of the target is embarrassing. So I think that the journals and the people who have been involved in meta-analyses should get together and have a conference sometime and try to come up with some generalized guidelines for reporting in different areas. When you do a meta-analysis in a particular area you find where a lot of the missing elements are. Now Ray Hyman and I in our joint communique I think did that for the ganzfeld domain. As far as I am concerned that is up to the editors of the journals now to make sure that the ganzfeld papers that are submitted include the level of descriptions that are dealt with there, for example. I would hope that Dean Radin and Roger Nelson would offer similar descriptions for the RNG area. I might add that the problem is not simply in the so-called proof-oriented research. Some of the most flagrant examples of inadequate reporting are in the more process-oriented research where the investigator has a pet theory or model which predominates with very little description of what the subjects were told to do, how were they recruited, these kinds of factors. I think we can do much better. I think one of the very totally uncontroversial

claims that can be made about meta-analysis is that it will help us to improve our reporting of our research in such a way as to increase the ability of future replicators to actually replicate what we have done and not have to read between the lines and second guess us.

UTTS: I, too, would like to share in the congratulations to you for this work and for your other meta-analyses. This is actually a comment that I was going to make this morning after Rao's paper and we did not have time for me to do it. And that is that I think one of the things that has come out of this meta-analysis is the fact that you did not find a relationship between effect size and sample size. Is that right?

HONORTON: Yes.

UTTS: I think the belief has been around for a long time that there is that relationship. In fact, that is one of the things the critics have tried to push, that when you increase your sample size suddenly you decrease your effect size. In his meta-analysis paper in the *Journal of Parapsychology*, Ray Hyman gave a statistical argument that that was the case in the ganzfeld database. It turns out when I looked at his argument that it was a statistical fluke and that indeed he had not shown that relationship. So anyway I would just like to say that I hope that we will see meta-analysis shattering some other myths that we have had around for awhile.

HONORTON: I think that, as far as I know, this is the first meta-analysis that goes beyond overall accumulation and starts to look at moderating variables. That is where we are going to find the real limitations of meta-analysis in terms of not being able to parse out how the different variables might be confounded. But at least we know a lot more now about what to look for, we have a much more informed basis for future precognition studies now than we did before. I think that is asking enough of this relatively new approach to data integration.

MORRIS: First let me just add to the compliments. I think this is really good, especially getting at the process aspects of things. However, I would like to know how you really define what a precognition study is. As you know many of them have multiple interpretations.

HONORTON: A precognition study was a study in which the subjects' task was defined for them as being precognition, as predicting the future.

MORRIS: Did you find ambiguities where that was difficult to assess?

HONORTON: The very first experimental precognition study was by Carrington and it was a dice throwing study. A few years later it would have been called a PK study. We considered it a precognition study because that is how he conceptualized it.

MORRIS: I think that is important, too.

HONORTON: Now the other thing I should mention here—and I think, Bob, you might be particularly interested in this—that an analysis on this database that we have not done yet, but one of the things that we kept coded, is in the non-automated studies that used random number tables or various shuffling methods, who was the randomizer? And did he use one of these complex calculations to try to reduce the likelihood that it was a combination of contemporaneous psi effects on the part of the experimenter? I have not looked at that yet, but I suspect that that will be informative. If there is, there should not be significant variation across randomizers or randomizing conditions, if the randomizer is not implicated in the outcome. So I think we will, for the first time on a larger scale, at least be able to begin to address that issue as well.

MORRIS: I am glad to hear that you are doing that. You might also pull in some of the studies that might be interpreted within the IDS model as well to see whether or not that shakes loose anything.