Journal of Parapsychology, Vol. 56, June 1992

RESPONSE TO HANSEN, UTTS, AND MARKWICK

Statistical and Methodological Problems of the PEAR Remote Viewing (*sic*) Experiments

By Y. H. Dobyns, B. J. Dunne, R. G. Jahn, and R. D. Nelson

ABSTRACT: Most of the issues raised by Hansen, Utts, and Markwick, including shared descriptor preferences, environmental or temporal cues, and agent encoding, have long been acknowledged, adequately addressed in our experimental designs and analytical techniques, and fully documented in our literature. The remainder of their concerns, including randomization of targets and reference score distributions, trial-by-trial feedback, stacking, and cheating are either misapplied, fundamentally incorrect, or have trivial impact. Additional calculations and derivations, supplementing those previously published, further demonstrate the insensitivity of our matrix scoring methods to target and descriptor dependence from any source. In sum, it is readily shown, both empirically and theoretically, that none of the stated complaints compromises the PEAR experimental protocols or analytical methods, which remain rigorous and effective methodologies for remote perception research. Thus, the published results and conclusions of our entire 336-trial database are fully reaffirmed.

Once the polemical generalizations and ad hominem assaults that disfigure their presentation are stripped away, the mélange of issues raised by Hansen, Utts, and Markwick (hereafter denoted HUM) seem to fall into three categories: (a) those that are indeed pertinent but have been previously recognized, thoroughly resolved, and fully presented in several publications on this topic; (b) those that are valid in general principle but are misapplied, misunderstood, or ineffective in this context; and (c) those that are fundamentally incorrect. In this rebuttal, we shall attempt to respond only to their various technical challenges to our analyses and protocols, following, insofar as possible, their order and titles.

Methodological Problems

Randomization

The Randomization section of the HUM article begins with two categorically incorrect assertions: "Randomness is the foundation

upon which all statistical inference is built. Without any source of randomness, it is impossible to assign probabilities to events."

Classical statistics and information theory are perfectly well equipped and are commonly applied to assess the degree of difference between two empirical distributions, neither of which qualifies as random (Box, Hunter, & Hunter, 1978; Snedecor & Cochran, 1980; Whalen, 1971). The basic issue of any remote perception research is whether data acquired under a given protocol differ from those acquired under another, or from some necessarily arbitrary control standard, none of which are likely to be fully random in the strict sense of the term. In our studies, for example, the primary question is whether given bodies of remote perception data more accurately represent specific characteristics of their proper targets than those of the other, irrelevant (mismatched) targets in that pool. Secondary questions include whether matched-target data acquired in a volitional protocol differ from those acquired in an instructed protocol, whether some percipient-agent pairs perform better than others, whether certain target characteristics are more perceptible than others, etc. The five analytical methods applied to these comparisons are quite competent to assess the statistical likelihood of these empirical differences without reference to any absolutely random distributions (Dunne, Jahn, & Nelson, 1983; Jahn et al., 1982; Jahn, Dunne, & Jahn, 1980).

HUM then proceed from this generic error to express assorted concerns about "subject's preferences and biases," which subsequently reappear in various forms throughout the balance of their paper. More specifically, they dispute the target selection process in the volitional trials and the coding of the descriptor lists. The former is indeed a legitimate a priori concern—so much so that a primary design consideration for establishment of the volitional protocol provided that its results be directly compared with those of the more traditional instructed protocol wherein the target pool is rigorously randomized. As noted below, no systematic scoring benefit seems to accrue from the former, more casual protocol. Specifically, the composite z score for 125 instructed trials is 5.771, with an associated p value of 4×10^{-9}, whereas the 211 volitional trials have a z score of 3.549 and a p value of 2×10^{-4}.

HUM's passing complaint about the randomization of the instructed target pool is similarly spurious; our randomization procedure has been adequately described or referenced in every publication or talk in which instructed protocol data have been presented (Bisaha & Dunne, 1979; Dunne & Bisaha, 1979; Dunne, Dobyns, &

Intner, 1989; Dunne, Jahn, & Nelson, 1983). Simply to repeat, these target pools are identified by an individual otherwise uninvolved in the program. Target locations are stored in randomized, numbered, sealed envelopes in safe-files of a second disinterested individual. Target envelope numbers are selected just prior to trial initiation by a suitable random process, usually our well-calibrated random event generator. The agent then is provided that envelope by the archivist, with no knowledge of its content by any member of the research group or percipient. The agent never opens the envelope until he has left the building and has no further contact with any participant until the trial is completed and the response forms collected.

Agent Coding

The HUM concern that shared descriptor preferences between percipients and agents may artificially inflate their matched-target scores is legitimate in principle but has previously been raised by several commentators, including ourselves, and has been the subject of an extensive analytical program that has been detailed in a comprehensive technical report (Dunne, Dobyns, & Intner, 1989) and in a number of personal communications to HUM themselves (Y. H. Dobyns, letter to Betty Markwick, Aug. 24, 1990; B. J. Dunne, letters to Betty Markwick, Nov. 9, 1989, April 2, 1990; R. D. Nelson, letters to G. Hansen, Feb. 21, 1986, Feb. 5, 1988). It turns out that most of this potential vulnerability is obviated by the basic technique of comparing matched-target scores with the complete mismatched (off-diagonal) score distributions. Nonetheless, one may still worry, as HUM suggest, that internal variations in the mismatched matrices may feed through as secondary sources of score inflation. Specifically for this reason, extensive analytical, ANOVA, and Monte Carlo calculations were performed some time ago to assess the possible magnitude of such artifacts (Dunne, Dobyns, & Intner, 1989). As one example, these analytical judging procedures were applied to a number of subsets of the database, each of which comprised only data from given percipient-agent pairs, wherein any potential shared descriptor preferences would operate equivalently in the mismatched as well as matched scores, and would thereby provide absolutely no advantage. In other studies, potential advantages from knowledge of season, time of day, geographical locale, and other potential cues were similarly assessed. Applied to our entire database, such studies found that some of these factors fed through to

trivial advantage, others to disadvantage, but none to any scale that individually or collectively altered the bottom-line conclusions. (See Dunne, Dobyns, & Intner, 1983, Section D, particularly Table D and Figure 10.) An extension of these calculations since performed to strengthen these results and to respond to other privately communicated HUM complaints is included in the following Appendix.

The HUM allusions to photographs and external judging further demonstrate their lack of comprehension of the basic purpose and strategy of the PEAR remote perception program. Agent encoding of targets was introduced deliberately and specifically to assess the role of the percipient-agent bond in the anomalous information transfer process, and to eliminate the highly variable subjective, and possible psychical, overlay of third-party judging on such experiments. Thus, we purposefully do not use analysts or photographs in the evaluation, although photographs taken by the agent at the scene may be used for secondary documentation. We have, however, assessed the possible vagaries of agent encoding by having a substantial subgroup of targets used in actual trials encoded by a third party not otherwise involved. Scores computed using these secondary encodings were found to be highly correlated with those using the agent encodings, and the mean scores were statistically indistinguishable therefrom. A similar close correlation was found in multiple-agent experiments (for example, compare Dunne, Jahn, & Nelson, 1983, Series 15a, 15b).

Shielding of Agent from Percipient: Potential Cheating by Subjects

Insinuations of cheating, however veiled or qualified, are the ultimate resort of hostile critics whose technical positions are indefensible. For members of this scholarly field to invoke such innuendo against their colleagues is repugnant and destructive; in this particular case, it is also quite illogical, for several reasons.

Throughout its 15-year history, our laboratory has concentrated on the quality of its equipment, experimental design and controls, analytical methods, theoretical models, interdisciplinary research staff, and commitment to basic understanding of the phenomena it studies. Sheer magnitude of effects has never been its principal interest or the focus of its reports. Indeed, most of our published results have tended to be more conservative than those from other laboratories; only in the overall size of the databases and variety of experiments have we claimed any precedence.

In the remote perception area in particular, our clearly stated purpose from the outset has been the development of incisive analytical techniques for the quantitative assessment of the data, rather than repetitious demonstration of an anomalous effect that we regard as already well established by many other laboratories. These techniques have been uniformly applied to all subsets of our own data, including several that have yielded no anomalous effect at all. All of the participants in our experiments are fully aware of, and interested in, this primary analytical purpose, and none is ever personally identified, financially compensated, or offered any other incentive for achievement. In this context, any carelessness of shielding or temptation to cheat is evidently profitless and self-defeating. Nonetheless, we have invariably followed well-defined protocols that could ensure reliable data, without excessive encumbrance or suffocation of the participants (Dunne, Jahn, & Nelson, 1983).

The disproportionately large component of data involving one particular long-term participant (010) is totally consistent with this overarching commitment to fuller understanding of the phenomena. Although HUM are quick to point out the reduction in overall z score when this participant's data are excluded, they either do not recognize or do not bother to acknowledge that this is totally attributable to the corresponding reduction in total database size: in fact, as shown in the Appendix, the average effect size actually *increases* slightly with these data removed. Thus, this participant's imputed "deception" or "cheating" could have succeeded only in reducing the scale of the anomalous effect.

Statistical Issues

Stacking

One of the prestated secondary purposes of our program is the assessment of correlations among responses of multiple percipients and their effects on overall performance. Such trials are a minority of the database, and in all cases the various percipients were separated from each other by long distances, and often by temporal intervals, rendering any stacking influences highly unlikely, or anomalous in their own right. The multiple-percipient trials are invariably analyzed both as separate groups as well as included in the grand concatenations. Their removal has negligible impact on the overall results.

Dependence Due to Target Selection Method

The first HUM concern here is the effect of no replacement of used targets on the statistical independence of the subsequent trials. Again, although this issue is superficially reasonable, it is rigorously demonstrated in the Appendix and confirmed in an independent analysis by Vassy (personal communication, Feb. 21, 1986) to be completely impotent within the PEAR scoring procedures. Normalization to an empirical background proves to be absolutely invulnerable to any statistical vagaries whatsoever that are internal to the sequence of targets or to the perceptions. Even without this assurance, it is well known that random selection without replacement from progressively larger pools has asymptotically zero nonindependence and, in principle, the target pools of the PEAR data are indefinitely large, because their selection process is totally unconstrained and the descriptors employed for their representation are by design sufficiently generic to be applicable anywhere.

HUM next worry about trial-by-trial feedback. Actually, most experiments (44 of 49 series) were done in series of several trials with no feedback until all trials were finished. Nevertheless, to examine this issue explicitly, we have performed relevant analytical and Monte Carlo tests, also detailed in the Appendix, that show that even the most aggressive strategy using trial-by-trial feedback, namely, inversion of the target descriptors as guesses for the subsequent perception, could substantially enhance scores only in very small series. Applying this technique to our actual data, we find only a portion of one series where a noticeable effect can thereby be obtained, and even there the net contribution is far too small to reduce the scores to nonsignificance if corrected. It should be noted that the series in question was part of the early ex post facto encoded subset which provided the basis for subsequent improvements and enhancements of the PEAR procedures. When the much larger ab initio encoded database is divided into those portions with and without trial-by-trial feedback, it is found that the subset with feedback has a *smaller* effect size than the set without feedback. We thus conclude that for our full PRP database, nonindependence and trial-by-trial feedback are nonproblems.

Target Pool Definition

Again, as shown in the Appendix and in Dunne, Dobyns, and Intner (1989), the possibility that "subjects might have some idea

about the range of targets in [a] particular pool" is completely disarmed by the use of local mismatched distributions.

PEAR's Position

HUM allude to a previous manuscript wherein one author (H) communicated some of these similar concerns. They fail to acknowledge, however, a large number of subsequent private responses wherein these issues were thoroughly discussed. Throughout all of this, it has been HUM's repeated assertion that some unspecified bias within the specific percipient-agent subsets could invalidate the basic premise of subgroup analysis. To the contrary, as shown in the Appendix, even for arbitrarily variable biases within subsets, the main diagonal remains unbiased relative to the mismatch distribution. Only if there are strong correlations in these biases could such an artifact arise, and since all reasonable sources of such correlations have already been shown to be negligible in the original analysis (Dunne, Dobyns, & Intner, 1989), the bias proposed by HUM cannot intrude. Possible shared descriptor bias associated with cueing on the weather, time of day, season, localization of target, and so forth was chosen as the focus of a HUM presentation at the 1991 Parapsychological Convention. As noted in personal communication with the authors (Jahn et al., letter to J. Utts, Oct. 14, 1991), this point can be directly dispelled by examining the composition of our database, for example:

1. All trials were in daylight hours.
2. Seasons were uniform across all subsets.
3. Most targets were hundreds or thousands of miles away from the agent's location.
4. Most trials were performed precognitively; in many cases perceptions were completed several hours or even days before the targets were even selected.

In other words, to activate the suggested mechanism for preferential cueing of descriptors, the percipient would have needed to forecast the weather for a very distant location that had only been broadly identified, and then transcribe this knowledge into the likely agent descriptor responses for an outdoor or indoor target that had not yet been selected. If the target selection was instructed from a prepared pool, such a process would have had to be even more convoluted. Conversely, if we select from our data only those trials that

could conceivably benefit from this strategy, we find that the effect size is no larger than for the major groups that could not possibly have been influenced.

Discussion and Conclusions

These sections of HUM are repetitions of previously stated, largely erroneous criticisms leading to correspondingly illegitimate generic conclusions. Curiously, in attacking this largest extant remote perception database of 125 instructed and 211 volitional trials as "too small" for adequate comparison, the authors suggest as a model for "appropriate analysis" an experimental series of 10 trials that is demonstrably irregular in all relevant analytical characteristics (Schlitz & Gruber, 1981), although they themselves have no experimental data or experience to contribute. On these grounds, they propose to discard the empirical finding that our very large, statistically viable group of instructed trials actually scores somewhat *higher* on average than the even larger group of volitional trials, thereby obviating, de facto, several of their major concerns.

The alternative experimental designs HUM suggest are not specified, let alone justified, and since the "defects" they purport to address are in the HUM interpretations, it is difficult to pursue their generic proposition. In our view, the "correct experimental design" they seem to be advocating is, in fact, pragmatically regressive. As we have detailed elsewhere (Jahn, 1983, 1988, 1991; Jahn & Dunne, 1987), it is our belief that in any research involving subtle psychological factors, imposition of unnecessarily rigid constraints that imply mistrust of the participants and inhibit their spontaneity may suppress, if not totally suffocate, the phenomena of interest. While effective precautions and controls are clearly essential in any valid protocols, they should not become so draconian that they poison the experimental ambience, or encumber the data acquisition and analysis to the point that the requisite large databases cannot reasonably be accumulated. With our more sophisticated designs and analytic strategies, it is possible to retain user-friendly protocols that allow some freedom of style, spontaneity, and enjoyment for the participants, while still deploying effective statistical checks to assess and compensate for any departures from "ideal" specifications that this flexibility entails.

Obviously, our results raise difficult epistemological questions, and less anomalous explanations should indeed be sought as vigor-

ously as those that appear to extend beyond known mechanisms. But the ultimate understanding of such phenomena will not benefit from ill-formed criticisms of necessarily imperfect basic research. It will only follow from many more good experiments and increasingly more incisive models.

The HUM Appendix

The HUM Appendix is largely a potpourri of earlier erroneous challenges, salted with demonstrably inferior alternatives. For example, at one point HUM propose an "optimal guessing strategy" that produces off-diagonal scores that are identical to the diagonal scores. In their discussion of statistical problems, the authors lump selection without replacement together with the question of whether diagonal scores should be included in the comparison matrix of off-diagonal scores. They fail to give further detail on the selection issue, but do perform a remarkably inapt assessment of the diagonal inclusion question. For an illustrative case they choose the experiment of 10 trials mentioned above, constituting a 10-by-10 matrix that not only yields an abysmally nonnormal distribution, but is far too small for their purpose (Schlitz & Gruber, 1981). Nonetheless, they then proceed via the assumption that a p value of 4.7×10^{-6}, derived by counting permutations, is valid (B. Markwick, private correspondence, Feb. 14 and June 5, 1990). In our Appendix, this permutation approach is shown to be generally incorrect for data of this type, rendering all of their subsequent arguments and recommendations invalid. Furthermore, as we have repeatedly demonstrated to HUM and again show in the Appendix, their diagonal inclusion question becomes trivial or moot in databases of the size of ours.

Beyond all this, the philosophical issue of appropriately defining any relevant "chance" distribution is ignored. In contrast, we have clearly defined our procedure as a comparison of experimental scores against a control condition established by an array of pseudo-scores derived from mismatched perceptions and targets, all based on the same set of 30 binary descriptors. As also shown in our Appendix, the resulting scores are slightly more conservative than, but closer to, the exact analytical norm than scores using the procedure recommended by HUM. We have examined these distributions for statistically required characteristics, compared the results with other appropriate procedures, and consulted extensively with experienced

senior statisticians at Princeton and elsewhere, and with several members of the critical community, none of whom has found fault with this strategy.

Concluding Comment

Critical assessment and dialogue are an indisputably essential component of good science. But to fulfill its proper purpose, such criticism must be informed, objective, and astute. When, for whatever reason, it lapses into slovenliness in substance, ad hominem attack, or ulterior motivation, it becomes a burden rather than a buttress to the scientific method. Parapsychology, for both epistemological and ontological reasons, has long labored under an excessive burden of illegitimate criticism that has seriously impeded its own scholarly progress. Much of this has obviously been directed from outside, hostile communities; but, in our opinion, an excessive traffic of this sort has also prevailed within the parapsychological family itself. This internecine carping has posed major distraction to the research it purports to abet, has provided fodder for opponents of the field to munch at their pleasure, and has discouraged uncommitted potential allies and supporters from associating with this field or with its investigators. It is our profound hope that this scholarly community will soon find better means to discipline its own critical commentary to the same high standards to which it aspires in its creative research.

Appendix

Diagonal Issues

In their own Appendix, HUM raise the issue of inclusion versus exclusion of the diagonal in constructing the background empirical "chance" distribution against which experimental trials are evaluated. The case they offer as a counterexample, originally posed in private communications from Markwick (Feb. 14 and June 5, 1990) is a 10 $\times$ 10 matrix of experimental scores (Schlitz and Gruber, 1981). Calculating a score for this matrix by the PEAR method produces an extremely low probability, and it is the contention of HUM that the true p value of this matrix is demonstrably much larger. In particular, they contend that including the diagonal values in the empirical

"chance" calculation would produce a p value much closer to the larger, "correct" value.

However, the premise of HUM that the "correct" p value of a scoring matrix is that value established by counting permutations is itself incorrect. It is true that a 10×10 matrix has $10! \approx 3.63 \times 10^6$ permutations, so that even if the matrix as given presents the best possible permutation, that is, if no other ordering of the columns with regard to the rows gives a larger total score on the main diagonal, permutation counting cannot return a p value smaller than 2.76×10^{-7} for the probability that this ordering of the scores appeared by chance. But this reasoning corresponds to a specific, inappropriate statistical model, namely, that for a given set of targets, the percipient somehow shuffles a preexisting set of perceptions in an attempt to find the best possible ordering for correspondence to the targets. In point of fact, in any encoded free-response protocol, the percipient is not choosing an ordering from some small number of available responses, but rather creates each response from an immense space of potential responses, for example, over 10^9 possible ways of filling out a perception checksheet in the PEAR protocol. The HUM scheme of permutation counting is nothing more than a ranking analysis that asks, "Does each perception match its proper target more closely than all other possible targets? Failing that, how badly does the relative ranking of targets to perceptions fall short of this ideal case?" With an analytical measure of the *degree* of match between any two responses, such as the PEAR protocol provides, one may legitimately ask a much more sensitive question, "*How closely* do the perceptions correspond to their proper targets?" a criterion invisible to simple permutation counting, and embodying far more information.

Consider some counterexamples to the premise that permutation counting gives the correct probability value for a general scoring matrix. First, let us imagine a procedure that, under the null hypothesis, produces independent random values at each position in a scoring matrix. Let us further envision that these values are normally distributed. Any given total score along the main diagonal can then be compared to an appropriate normal distribution to establish a completely unambiguous probability for its occurrence. As a specific example, consider the matrix:

47	29	35	27	33	19	33	34	32	10
30	27	28	20	41	30	16	32	29	50
31	36	53	35	38	36	31	26	39	43
23	21	44	49	41	32	36	29	39	34
28	31	31	27	40	30	23	26	23	22
25	34	37	38	24	31	26	29	24	20
31	27	28	25	37	34	46	35	24	36
22	37	24	25	30	32	30	48	38	22
27	28	29	35	28	17	40	42	45	33
19	26	21	18	25	33	40	32	30	40

Here the 90 off-diagonal elements are constructed to fit as closely as practical a normal distribution with mean 30 and variance 50 that is arbitrarily taken as the "chance" distribution of this hypothetical set. These 90 integers are then randomly assigned to off-diagonal locations in the matrix. (Examining the statistics of the off-diagonal elements will show them, in fact, to be much better behaved in their statistical moments than would be expected for a genuinely random draw from the above distribution.) The diagonal elements are randomly drawn from the same distribution, but with a uniform mean shift added to produce a total deviation of + 126 above the expected value, corresponding to a z score of 5.634 and thus a $p = 9.1 \times 10^{-9}$. This is the "real" p of the dataset under the null hypothesis. An experimenter presented with these data, knowing the expected characteristics, but not knowing the actual underlying distribution other than having the off-diagonal elements as a sample for it, would in this instance get almost exactly the correct value by applying the PEAR procedure.

Now consider permutation counting. The total score of the main diagonal is 426, and there are at least 4 permutations that produce higher values. (Interchange columns 2 and 5; interchange 2 and 6; interchange 2 and 10; interchange 2, 6, 10 → 10, 2, 6.) Therefore the p value by permutation counting cannot be higher than $5 \times 2.76 \times 10^{-7} = 1.38 \times 10^{-6}$, since the given permutation is, at most, among the best 5 possible. This p value is demonstrably incorrect by over two orders of magnitude.

This is not an artifact of the particular matrix chosen. As long as the mean shift of the diagonal elements is not too large compared to either their own standard deviation or that of the off-diagonals, as is generally the case in remote perception data, the odds are quite good that one can find permutations that bring in large off-diagonal elements in place of small diagonal elements. The assumption of independent elements, used to simplify the statistical model for this example, is in

fact not essential, as has been shown in a more detailed analysis (Dobyns, 1992).

It is not always the case that counting permutations yields an incorrectly conservative estimate, as occurs above. Consider a 10×10 matrix whose off-diagonal elements are, with equal frequency, 20 and 80, and whose diagonal elements are all 81. Since all off-diagonal elements are lower than all diagonal elements, it is already in the best possible permutation, and permutation counting thus returns a p value of 2.76×10^{-7}. However, by the PEAR scoring algorithm, the main diagonal has a net z of 3.268, yielding a far more conservative p of 5.42×10^{-4}

The overarching point of these examples is simply that the permutation-counting method, regarded by HUM as returning the "correct" p values for scoring matrices, rests on an inappropriate statistical hypothesis that does not correspond to the way scores are generated, and ignores much of the information in the experimental data.

HUM next insist that diagonal scores must be included in the calculation of the chance background, since this more conservative test does, at least with regard to certain datasets published in the parapsychological literature, produce a p value closer to that erroneously obtained from permutation counting. But such diagonal inclusion is categorically unjustifiable on experimental grounds. The experimental condition superimposes some unknown degree (possibly zero) of anomalous information transfer on a background of fluctuations driven by the degree of correspondence between an arbitrary target scene and an arbitrary imagined scene. The proper control condition is a distribution of target scenes compared to imagined scenes where no such information transfer could have occurred. To construct a chance distribution from a mixture of trial and mismatch scores is to compare experimental data with a mixture of experimental and control data, rather than the more incisive test of comparing experimental data against a control. As such, it is guaranteed artificially to deflate the impact of any real effect that may be present. Further analysis and calculations of diagonal-included versus diagonal-excluded scoring are presented later in our calculations section.

In sum, the arguments in favor of diagonal-included scoring are specious and are founded on a logical fallacy. Contrary to the opinions of HUM, permutation-counting analysis is a potentially misleading procedure whose injudicious application may have deleterious effects largely unrecognized in the parapsychological literature. Indeed, all forms of ranking analysis, when compared to analytical scoring tech-

niques, suffer from the same inferiority as permutation counting in surrendering information about the degree of similarity in transcripts.

Variable Biases

The issue of bias effects in internally uniform subsets was thoroughly discussed in Appendix C-II of Dunne et al. (1989). To recap briefly, an internally uniform subset was defined as a group of trials in which agent and percipient preferences (biases) were constant. It was noted that any such bias will equally affect the trial scores (main diagonal of scoring matrix) and the mismatch comparison scores (also called "off-diagonal" or "empirical chance" scores). As discussed earlier in the same reference, combining subsets with different biases can produce a spurious effect, but such effects can be completely precluded by restricting score calculations to uniform subsets and using standard composition rules to combine the z scores for subsets into an aggregate z score for the entire matrix.

HUM have objected to this on two related grounds, namely, that trials are not statistically independent and that biases may differ within an apparently uniform subset. They claim that statistical nonindependence of the trials may arise from choice without replacement in the instructed target pools or from the free choice of location by the agent in the volitional trials. Either of these arguments essentially amounts to contesting the existence of internally uniform subsets. (Note that nonindependence is equivalent to a variable bias because the *effective* response probability on a given trial will be a function of the innate probability and the correlations with trials already performed.) To address these issues, the following analysis will dispense with the assumption of uniformity and work in a completely general framework of trial-by-trial response probabilities.

We shall first address the PEAR Method B, since that was the primary technique used in the published report. This procedure uses an empirical agent-response frequency, α, for each descriptor. Each descriptor that matches between agent and percipient reports is awarded a score of $1/\alpha$ if the response is positive and $1/(1 - \alpha)$ if negative. The total score on the 30 descriptors is normalized by the score that would have been attained had all the descriptors matched. For simplicity, consider only a single descriptor, assuming the others to be unbiased, an assumption that produces no loss of generality, as will be shown later. In accordance with the procedure derived in Dunne et al. (1989), regard α as the actual response frequency to that descriptor in the data subset under consideration, rather than a global

parameter. Consider a subset comprising N trials, with A_i being the probability that the agent responds positively to the descriptor under consideration in the ith trial, and P_i being the corresponding probability of a positive response from the percipient. Although this representation can be used for an actual dataset by requiring all A_i and P_i to be 0 or 1, the intent of this analysis is to consider prior expectation values for an experimental sequence not yet performed. Nevertheless, it applies with equal force to real data, provided one remembers that the quantity here treated as an expectation value will for real data be a definite value, not necessarily equal to the expectation, but always an unbiased estimator of it. The notation $< \chi >$ will be used to denote the expectation value of χ.

By definition, $< A_i > = (1/N) \Sigma A_i = \alpha$. We may define $a_i = A_i - \alpha$ as the local deviations of the agent's expected behavior, for a specific trial, from the global mean; $\Sigma a_i = 0$ necessarily. We may likewise define $p_O = < P_i >$ and $p_i = P_i - p_O$ such that $\Sigma p_i = 0$. Let E_{ij} be the expected score of target i against perception j, given values for a_i and p_j. This has the form

$$E_{ij} = \frac{29 + N_{ij}}{58 + D_{ij}} \tag{1}$$

where the constants come from the other, presumed unbiased descriptors. The numerator and denominator contributions are

$$\begin{aligned} N_{ij} &= \frac{1}{\alpha} A_i P_j + \frac{1}{1 - \alpha} (1 - A_i)(1 - P_i) \\ &= \frac{1}{\alpha}(\alpha + a_i)(p_0 + p_j) + \frac{1}{1 - \alpha}(1 - \alpha - a_i)(1 - p_0 - p_j) \\ &= 1 + a_i \left(\frac{p_0 + p_j}{\alpha} + \frac{p_0 + p_j - 1}{1 - \alpha} \right) \end{aligned} \tag{2}$$

and

$$\begin{aligned} D_{ij} &= \frac{1}{\alpha} A_i + \frac{1}{1 - \alpha} (1 - A_i) \\ &= 2 + a_i \left(\frac{1}{\alpha} - \frac{1}{1 - \alpha} \right) \end{aligned} \tag{3}$$

Thus,

$$E_{ij} = \frac{30 + a_i\left(\frac{p_0 + p_j}{\alpha} + \frac{p_0 + p_j - 1}{1 - \alpha}\right)}{60 + a_i\left(\frac{1}{\alpha} - \frac{1}{1 - \alpha}\right)} \tag{4}$$

If A_i and P_j are probabilities, E_{ij} is an expectation value; if they represent actual data, E_{ij} is an actual score. Again, we are concerned with the expectation value for general data. No assumptions have been made concerning the values of a_i and p_i aside from the defining constraints that they sum to 0 over the full range of the index. They are therefore a completely general representation of any possible variation in response probabilities from trial to trial. Any kind of correlation or guessing strategy can be represented by assigning appropriate values to A_i and P_i or, correspondingly, a_i and p_i.

We are ultimately concerned with the expected difference between the mean value of trials on the main scoring diagonal versus the mean value of off-diagonal trials. Note that the denominator of E_{ij} in Formula 4 above does not depend on j. Therefore, we may directly compute T_i, the expected score of the ith trial, and O_i, the expected mean value of the ith row of off-diagonal scores:

$$T_i = E_{ii} = \frac{30 + a_i p_0\left(\frac{1}{\alpha} + \frac{1}{1 - \alpha}\right) + a_i\left(\frac{p_i}{\alpha} + \frac{p_i - 1}{1 - \alpha}\right)}{60 + a_i\left(\frac{1}{\alpha} - \frac{1}{1 - \alpha}\right)} \tag{5}$$

and

$$\begin{aligned} O_i &= \frac{1}{N - 1}\sum_{j=1;j\neq i}^{N} E_{ij} = \frac{1}{N - 1}\sum_{j=1}^{N}(1 - \delta_{ij})E_{ij} \\ &= \left(\frac{1}{N - 1}\right)\frac{1}{60 + a_i\left(\frac{1}{\alpha} - \frac{1}{1 - \alpha}\right)}\sum_{j=1}^{N}(1 - \delta_{ij}) \times \\ &\quad \left(30 + a_i\left(\frac{p_0 + p_i}{\alpha} + \frac{p_0 + p_j - 1}{1 - \alpha}\right)\right). \end{aligned} \tag{6}$$

In the numerator of O_i, for any term t that does not depend on j, $\frac{1}{n-1}\Sigma_j(1-\delta_{ij})t = t$. The only terms that do depend on j are proportional to p_j, and $\frac{1}{N-1}\Sigma_j(1-\delta_{ij})p_j = -p_i/(N-1)$ since $\Sigma_j p_j = 0$. Applying this to Formula 6 we obtain

$$O_i = \frac{30 + a_i p_0\left(\frac{1}{\alpha} + \frac{1}{1-\alpha}\right) - a_i\left(\frac{p_i/(N-1)}{\alpha} + \frac{p_i/(N-1)+1}{1-\alpha}\right)}{60 + a_i\left(\frac{1}{\alpha} - \frac{1}{1-\alpha}\right)}. \tag{7}$$

Much of this expression is identical to Formula 5 for T_i, so that the rowwise difference may, after some algebra, be rendered by the relation

$$T_i - O_i = \frac{N}{N-1}\frac{a_i p_i}{60(\alpha-\alpha^2) + a_i(1-2\alpha)}. \tag{8}$$

The ultimate quantity of interest, of course, is the expected difference between μ_D, the mean of the scores on the diagonal, and μ_O, the mean of the off-diagonal scores. Note that T_i and O_i are already expectation values:

$$< \mu_D - \mu_O > = \frac{\Sigma T_i}{N} - \frac{\Sigma O_i}{N} = \frac{1}{N}\sum_i (T_i - O_i). \tag{9}$$

Thus, this expected difference can be obtained by summing Formula 8 over all values of i. In the specific case that $\alpha = 0.5$, $< \mu_D - \mu_O > \propto \Sigma_i a_i p_i$. Since a_i and p_i both have expectation 0 by definition, this quantity can be nonzero only if there is a nonzero correlation between the a's and the p's. Any actual dataset will, due to random variations, display some actual correlation, positive or negative; however, this is simply one of the contributions to the inevitable variation about the central tendency that will be displayed by any statistical measure. The important point is that, unless there is some correlated influence that gives $\Sigma_i a_i p_i$ a nonzero expectation, the difference between on-diagonal and off-diagonal means has zero expectation. Moreover, the requisite correlation is between the set of a_i and the set of p_i; the internal cor-

relations that may appear within each set, owing to nonrandom target selection or psychological factors of the participants, are completely irrelevant.

For more general values of α, we may simplify the denominator of Formula 8 by factoring out $(1 - 2\alpha)$ so that

$$< \mu_D - \mu_O > = \frac{1}{(N - 1)(1 - 2\alpha)} \sum_{i=1}^{N} \frac{a_i p_i}{K + a_i}, \tag{10}$$

where $K = 60\dfrac{\alpha - \alpha^2}{1 - 2\alpha}$. Although this is somewhat more complicated to evaluate, Monte Carlo calculations for randomly varying a and p indicate that it has expectation zero. This zero expectation results even when the a_i and p_i generation algorithms for the Monte Carlo analysis are designed to build high degrees of internal correlation into each group. The character of Formula 10 can be observed analytically by multiplying the entire expression by $\Pi_{j=1}^{N}(K + a_j)$, resulting in

$$< \mu_D - \mu_O > \propto \sum_{i=1}^{N} a_i p_i \prod_{j \neq i}^{N} K + a_j. \tag{11}$$

The expansion of $\Pi_{j \neq i} K + a_j$ is a polynomial in powers of K and the various a_j.

$$\begin{aligned} \prod_{j \neq i} K + a_j &= K^{N-1} + K^{N-2} \sum_{j \neq i} a_j + K^{N-3} \sum_{j \neq i; k \neq i, j} a_j a_k + \ldots \\ &= K^{N-1} + K^{N-2}(-a_i) + K^{N-3}\left(\sum_{j \neq i} a_j(a_i + a_j)\right) + \ldots \\ &= K^{N-1} - K^{N-2} a_i + K^{N-3}\left(2a_i^2 - \sum_j a_j^2\right) + \ldots . \end{aligned} \tag{12}$$

We can see from the example of the leading terms that every term in the expansion must ultimately resolve to either (a) a constant, independent of i, or (b) a term proportional to a_i^{χ} for $1 \leq \chi \leq N - 1$. This entire polynomial then multiplies the factor $a_i p_i$, and the resulting family of polynomials is summed for each i. Therefore it follows that the entire summation takes on the form

$$< \mu_D - \mu^{\circ} > \propto C_1 \sum a_i p_i + C_2 \sum a_i^2 p_i + C_3 \sum a_i^3 p_i + \ldots \tag{13}$$

where the C's are coefficients involving powers of K and sums such as $\Sigma_j a_j^{\chi}$ for various powers χ. Although explicitly calculating these coefficents for the general case would be very tedious, the important point is the dependence of the summations. The first, as we have seen above, corresponds to a correlation between a and p. The second similarly corresponds to a correlation between a^2 and p, and so forth for the later terms. Now, if there is no reason to expect a correlation between p and a, there is even less reason to expect correlations between p and higher moments of a. Therefore, *in the absence of real information transfer, anomalous or otherwise, between the agent and the percipient, that causes a correlation in their respective response biases,* all of these summations have expectation 0 which is a sufficient, though not necessary, condition for $< \mu_D - \mu_O > = 0$.

Thus, if all other descriptors are unbiased, and if the biases introduced on a given descriptor by agents and percipients are not correlated with each other, the expected bias to the given descriptor is zero in Method B scoring with local α's. It then follows, as one can see by imagining adding biases to one descriptor at a time until the full set has been dealt with, that all the descriptors are unbiased, in the sense of having no expected difference between trial scores and mismatch scores. The point to emphasize is that this result is not affected by any internal correlations or dependencies *within* the agent or percipient response frequencies, but only on correlations *between* the two. Dunne et al. already make a thorough examination of possible sources of correlations in the bias, such as variations of personal preferences or knowledge of the season in which the trial occurred. These are reinforced and extended by extensive ANOVA calculations described in the calculations section of this Appendix. Issues of statistical nonindependence of targets, or selection of targets without replacement, or even of nonrandom choice of targets, are thus seen to be irrelevant because they can *only* introduce structure to the sequence of a_i describing the targets. The important issue, then, is correlation between agent and percipient responses, to which concerns of target nonindependence are an irrelevant red herring.

To this point, the discussion has centered on Method B, which has some drawbacks from an analytical standpoint, as seen in the preceding complicated derivations. Moreover, conclusions ultimately involve a generalization from one to many descriptors which, given the complexity of the calculations, may seem obscure. In Method A, on the other hand, trial scores are simple sums of descriptor scores, and the scoring calculations themselves are far simpler. Because it was found that Method A produces results statistically similar to Method B, an analysis of Method A in the same terms seems worthwhile.

Using the same notation, we find that the expected value of scoring matrix element i, j for one descriptor is

$$\begin{aligned} E_{ij} &= A_i P_j + (1 - A_i)(1 - P_j) \\ &= (\alpha + a_i)(p_o + p_j) + (1 - \alpha - a_i)(1 - p_o - p_j) \\ &= \alpha p_o + (1 - \alpha)(1 - p_o) + a_i(2p_o - 1) + p_j(2\alpha - 1) + \\ &\quad 2\alpha_i p_j. \end{aligned} \tag{14}$$

We are ignoring here the normalization 1/30 applied in the actual score since, unlike the Method B denominator, it is a constant multiplier to the entire matrix. For brevity, let us define $C = \alpha p_O + (1 - \alpha)(1 - p_O)$, which we recognize as the constant expected contribution from the average response frequencies. The expected mean of the diagonal scores is

$$< \mu_D > = \frac{1}{N} \sum_{i=1}^{N} E_{ii} = C + \frac{2}{N} \sum_{i=1}^{N} a_i p_i \tag{15}$$

since $\Sigma a_i = \Sigma p_i = 0$ causing the linear terms to drop out of the sum. The expected off-diagonal mean is

$$\begin{aligned} < \mu_O > &= \frac{1}{N^2 - N} \sum_{i,j} (1 - \delta_{ij}) E_{ij} \\ &= \frac{1}{N^2 - N} \left(\sum_{i,j} E_{ij} - \sum_i E_{ii} \right). \end{aligned} \tag{16}$$

Now consider the double sum $\Sigma_{ij} E_{ij}$ in more detail:

$$\begin{aligned} \sum_{ij} E_{ij} &= \sum_{i=1}^{N} \sum_{j=1}^{N} C + a_i(2p_0 - 1) + p_j(2\alpha - 1) + 2a_i p_j \\ &= \sum_{i=1}^{N} NC + Na_i(2p_0 - 1), \text{ since } \sum_j p_j = 0 \\ &= N^2 C, \text{ since } \sum_i a_i = 0. \end{aligned} \tag{17}$$

Therefore,

$$\begin{aligned} < \mu_O > &= \frac{1}{N^2 - N}\left(\sum_{i,j} E_{ij} - \sum_i E_{ii}\right) \\ &= \frac{1}{N^2 - N}(N^2 C - N < \mu_D >) \\ &= \frac{N}{N - 1} C - \frac{1}{N - 1} < \mu_D > . \end{aligned} \tag{18}$$

Thus, the expected difference between the two means is

$$\begin{aligned} < \mu_D - \mu_O > &= < \mu_D > - < \mu_O > \\ &= < \mu_D > - \frac{N}{N - 1} C + \frac{1}{N - 1} < \mu_D > \\ &= \frac{N}{N - 1}(\mu_D - C) = \frac{2}{N - 1} \sum_{i=1}^{N} a_i p_i . \end{aligned} \tag{19}$$

It then follows, somewhat more simply than for Method B, that the expected bias on each descriptor is 0 provided the agent and percipient response frequencies on that descriptor are mutually uncorrelated. Unlike Method B, the generalization to the full set of 30 descriptors is the simple sum of the bias computation for each descriptor, and is therefore zero when the condition of agent-percipient noncorrelation is maintained.

The previous material covers the discussion of "internally uniform" subsets in Dunne et al. as a special case. For an internally uniform subset, all of the a_i and p_i are zero, which trivially leads to the result that the bias formulae derived above are zero. Likewise, blocks of data by different agents and/or percipients with different response biases will bias the final score if and only if those local biases are correlated, and any such effect can be averted by conducting one's analysis within subsets such that, for example, the same agent and same percipient are involved with no communication prior to completion of the entire block, so that no possibility of correlated variation occurred. Most of the remote perception data were gathered under precisely such conditions.

These general results are of course much stronger than those derived in Dunne et al., since the requirement of internally uniform subsets is now replaced by the much weaker requirement of subsets whose internal variation is uncorrelated between agent and percipient.

Calculations

Analysis of variance. As noted in the text, analysis of variance can be applied to examine possible effects of secondary parameters in the PRP database, and thus it provides further perspective on some of the questions raised by HUM. The entire formal database has been subjected to a general linear model ANOVA with z scores as the dependent variable and several secondary parameters or categorical groupings as independent variables. The factors included volitional versus random target selection, single versus multiple percipients, ab initio versus ex post facto encoding, summer versus winter trials, trial-by-trial feedback versus post-series feedback, data involving Participant 10 versus all others, and a three-level factor comparing targets near Chicago, near Princeton, and elsewhere. The ANOVA main effect was found to be significant, $F\ (8,320) = 2.741, p = .0061$, but none of the seven secondary parameters was individually significant. This means that the scores are indeed shifted from their normal expectation, but that the effect is not greatly influenced by any of these factors. Two of the interactions were marginally significant, namely, target selection mode with single versus multiple percipients, $F\ (1,320) = 3.009, p = .084$, and target encoding with geographic location, $F\ (1,320) = 3.666, p = .056$. The assessment of interactions is limited to those which are not confounded (e.g., we cannot compute an interaction for geographic location with target selection because all Chicago targets were randomly selected). However, the overall effect of interactions is not significant, indicating that in general the anomalous shift of z scores is not affected differentially by particular combinations of secondary parameters.

We conclude that none of the secondary parameters are important, with the possible exception that their interactions may provide some useful insight. In particular, this analysis shows that the target selection procedure is not a significant factor; and contrary to the HUM thesis that agent selection of targets may result in artifactual score contributions, the small difference in the subgroup mean shifts favors the randomly selected targets. Furthermore, and contrary to their assertion that the PEAR sample sizes may be too small for such comparisons, the internal and pooled standard errors for the subgroup means differ by less than two percent, as might be expected from groups with well over 100 trials. Similarly, the ANOVA shows that several of the other HUM speculations about possible spurious contributions to scoring have no basis in fact. HUM suspect "stacking" in the multiple percipient trials, but these have a smaller mean score

than the single percipient subset. Seasonal differences, one of the claimed possible sources of bias, represented in the ANOVA as summer versus winter trials, do not approach significance. Addressing their suggestion that Participant 10's data ought to be excluded because they contribute unduly, this analysis yields an F ratio of 0.084 ($p = .77$) for the factor comparing these trials against those of all other participants, showing HUM's contention and their imputation of cheating to be entirely specious.

Trial-by-trial feedback is also a nonsignificant factor, but its suggestive interaction with the encoding factor raises a noteworthy point. The interaction is driven by a small subset of 28 trials with trial-by-trial feedback in the ex post facto category that displays higher scores than the other combinations. This group is special in a number of ways that may be instructive; and while we cannot precisely identify the source of their uniqueness without more research, the following features are probably relevant: They are all from the earliest part of the database and include several of the most impressive trials by any standard; they were done in a pure free-response mode without the analytical check sheet methodology; they used tape recording of rich, lengthy descriptions; and they were all produced by percipients new to this kind of experience, who had no conditioned expectations. In addition, as detailed in the Monte Carlo examination of trial-by-trial feedback, these trials could in principle have been affected, though not strongly, by such feedback as a source of spurious inflation. Finally, there is the possibility that the z scores for this ex post facto subgroup may be larger because these data served as the model for the analytical judging descriptor choices, and hence they may be more responsive to these descriptors.

In summary, the ANOVA leads to the conclusion that none of the parameters examined has an appreciable impact on the character of the data, with the possible exception of trial-by-trial feedback, which shall be examined more fully via a Monte Carlo analysis later in this section. The conclusion in turn obviates HUM's concerns about randomness in target selection, stacking, sensory cueing, and overparticipation by one individual; the next section focuses further on the last of these issues.

One participant's role. HUM criticize the extensive involvement of Participant 10 as either agent or percipient. This individual's total involvement is 244 of the 336 formal trials—as agent or percipient in all 59 of the ex post facto encoded trials and in 185 of the 277 ab initio encoded trials. The effect sizes and z scores of these datasets can be summarized thus:

Dataset	Effect size	*N* trials	Total *z*
No. 10, ex post facto	0.754	59	5.79
No. 10, ab initio	0.253	185	3.44
Others, ab initio	0.267	92	2.56

Two conclusions are readily apparent. First, using the comparable ab initio protocols, trials involving Participant 10 have a slightly smaller effect than trials not involving No. 10, although the difference is not statistically significant. The larger *z* score of the trials with Operator 10 is purely due to the larger dataset size. The value just quoted for the *z* score of the dataset exclusive of Participant 10 differs from the value 2.17 quoted in HUM. That value apparently was based on a composite *z* score subtraction from the tables of individual agent and percipient effect sizes published in Dunne et al., and therefore differs from the value computed directly by scoring within the appropriate submatrix. Second, when Participant 10's performance is compared across the two different protocols, the effect size is larger by a factor of 3 in the ex post facto data. These data, with third-party encoding throughout, and almost entirely drawn from an instructed target pool, are much closer to the protocol implicitly recommended by HUM as the standard of parapsychology: randomly chosen targets, encoding by consensus of a third-party panel, complete and unambiguously enforceable shielding between agent and percipient (in many of the trials both participants were physically accompanied by observers for the entire interval from trial initiation to final recording of transcripts; most of the exceptions involved participants on different continents), and so forth. If Participant 10's performance is to be attributed to chicanery, it seems decidedly odd that it is spectacularly better under a protocol that makes any form of deception vastly more difficult. The close consistency of Participant 10's performance in the ab initio protocol with that of a large pool of volunteers suggests instead that this person's large statistical contribution arises entirely from a willingness to generate large amounts of data.

Monte Carlo analyses. Earlier Monte Carlo runs did not address the issue of mutual nonindependence of descriptors, either within or between checksheets. (Here and in the following, "checksheet" is used to mean one full set of descriptor responses produced by either agent or percipient.) Although there are both empirical and conceptual grounds for regarding this nonindependence as unimportant, it was nonetheless considered worthwhile to examine these effects under conditions as close to actual experimental data reduc-

tion as possible. The two classes of nonindependence at issue are quite distinct. The interdependence of descriptors within a given checksheet, arising from the tendency of real or imagined scenes to contain related elements, is expected a priori to be inconsequential in the analysis because its net effect is to reduce the amount of information in a checksheet. Although this would alter the distribution relative to some ideal chance value occurring in an uncorrelated world, such effects should already be properly compensated, because the chance background used is empirically derived from rearrangement of the same responses. The nonindependence between trials can arise from the fact that instructed targets are drawn without replacement, or that agents never visit the same volitional target twice and may tend to avoid closely similar scenes.

Regardless of the source of the nonindependence, the net statistical effect is that a certain probability distribution of target checksheets exists, from which specific checksheets are drawn without replacement. The vagaries of human volition and personal response bias are thus subsumed into the hypothetical distribution of checksheets. Such representation in terms of overall checksheet probabilities is extremely powerful. Assigning a net probability to each of the 2^{30} possible checksheet response configurations fully describes the space of possible responses, subsuming in one stroke the descriptor probabilities, two-descriptor correlations, *N*-descriptor correlations, and so forth. The drawback to this approach is that comparison of some billion possible target sheets versus a like number of possible responses is not feasible at the level of individual probabilities. It is, in contrast, practical to estimate effects via Monte Carlo analysis.

To reduce the space of possible checksheets to manageable size, only 8 descriptors at a time were used in the first stage of the investigation, leaving only 256 entries for which probabilities needed to be calculated. Response frequencies for each of the possible checksheets could then be computed to reproduce with high accuracy the descriptor frequencies and two-descriptor correlations of any arbitrary set of 8 descriptors in the actual target data.

The calculation of the 8-descriptor checksheet distribution requires a set of 8 descriptor frequencies and the corresponding 8×8 correlation matrix. This was obtained for each Monte Carlo analysis by selecting 8 descriptors from the actual target data and the corresponding 8×8 submatrix of the full 30×30 correlation matrix. For each set of target probabilities so generated, the following calculation was run:

1. Generate 10 targets by drawing from the set of 256 checksheets, without replacement, in accord with the probability assigned each checksheet.

2. Generate 3 different sets of 10 "perceptions," using three different algorithms: *Biased guessing* uses the actual frequencies and correlations in the perception data of the same 8 descriptors selected to establish the target checksheet probabilities; the same exact checksheet probability calculation is performed. This procedure reproduces whatever biases exist in the actual percipient guessing strategy. *Informed guessing* uses the same checksheet probability list as the target generation process, representing a percipient who is trying to gain an advantage from knowledge of the agent's preferences. *Adaptive guessing* addresses the question of trial-by-trial feedback by adjusting the response probabilities for each trial to avoid the salient aspects revealed by earlier targets. The first response in adaptive guessing uses the informed-guess probabilities. The second response is the logical inversion of the first target. Thereafter, the agent response frequencies are observed empirically for the first N targets and are inverted to generate the probabilities for perception $N + 1$. Scoring method A is then used to create three 10×10 score matrices, each of which is scored in two ways: diagonal versus off-diagonal (diagonal-excluded, the standard PEAR approach) and diagonal versus whole matrix (diagonal-included, the approach recommended by HUM.)

The final result of these calculations is six different composite z scores for three different 10×10 scoring matrices. Since, by construction, no anomaly is present (all three of the perceptions are computed using guessing strategies, and even adaptive guessing uses information that would be available to the percipient by non-anomalous means under the conditions for which it is testing, namely, trial-by-trial feedback within a series), the computed z statistic should display its null hypothesis distribution, with mean 0 and variance 1. To estimate the actual distribution of the test statistic, the above calculation was repeated 1,000 times. This gives an estimate of the behavior of the z scores for the given set of descriptors. Different descriptor subsets were then chosen for a total of 20 repetitions of the 1,000-set run. The first three sets were, respectively, descriptors 1–8, 9–16, and 17–24; the remaining 17 repetitions were set up by randomly drawing 8 of the 30 descriptors to generate checksheet probabilities.

For biased guessing with the diagonal excluded, which is in some sense the closest approximation to PEAR data scored with the PEAR procedure, the mean z score is 0.00479 ± 0.00728. (All $\pm$ uncer-

tainty values are 1 σ limits computed from the variability across the 20 repetitions for different descriptor sets.) This is statistically indistinguishable from zero. The standard deviation is 0.956 ± 0.011, which is neither the ideal value 1 nor the increased value which might be expected for independent matrices, but instead is slightly depleted. This implies that the PEAR reporting procedure is slightly conservative, in that we are computing probability ranges based on a null hypothesis variance of 1 on a test statistic whose actual standard deviation under the null hypothesis is 0.956. Biased guessing with the diagonal included produced a mean z of 0.00350 ± 0.00651, again effectively zero, but with a standard deviation of only 0.8540 ± 0.0098. Thus, had we adopted this method, the results would have been even more overconservative than those from the diagonal-excluded test.

The informed-guessing calculation also fails to produce a spurious effect. The diagonal-excluded score develops a mean of 0.00513 ± 0.00708, with a standard deviation 0.9582 ± 0.0090. The diagonal-included statistics yield a mean of 0.00562 ± 0.00635 and a standard deviation of 0.8556 ± 0.0079. Thus, it again appears that trying to guess in accordance with the preferences of the agent is unhelpful in enhancing the score, which supports the analytical discussions of guessing strategies presented earlier.

The adaptive guessing strategy, in contrast, does produce a spurious effect, which also confirms the analytical discussion of variable guessing strategies. In that analysis it was demonstrated that a variable guessing strategy produces a bias if, and only if, there is a correlation between the variations in agent response frequencies and percipient response frequencies. The combination of target selection without replacement and trial-by-trial feedback allows the percipient to change response frequencies in a way correlated to the agent, though neither element alone permits this. In the diagonal-excluded calculation a mean of 0.317 ± 0.029 with standard deviation 0.757 ± 0.012 is found. In the diagonal-included scoring, the mean becomes 0.284 ± 0.026, with standard deviation 0.682 ± 0.011. These results, which are mean z scores for 10-trial sets, correspond to an effect size per trial of ≈ 0.1, or about one-third to one-half of the typical PRP effect size in various subsets. They therefore merit closer examination.

First, note that in the diagonal-excluded case, if we take 1,000 samples from a population with standard deviation 0.757 we should expect the resulting mean to have a standard error of $0.757/\sqrt{1{,}000} = 0.024$, which is also the expected standard deviation of the mean from group to group if we take such samples repeatedly.

Instead, the standard deviation of the 20 means is found to be 0.117, leading to the statistical uncertainty quoted above after a further division by $\sqrt{20}$. This shows that the samples are taken from different populations, in that the mean of one group of 1,000 can be statistically distinguished from the mean of another group. Examination of the statistics of individual groups bears out the fact that the mean z score for adaptive guessing is strongly dependent on the descriptor set in use; indeed, for one set (the descriptors 9–16), it is actually slightly negative. Also note that there is danger in extrapolating the size of the bias from 8 descriptors to 30, that most of the PRP data were collected in well-defined series with trial-by-trial feedback *not* available to the percipient, and that the dependence of the effect on dataset size is not established by this first-phase analysis, which uses 10×10 subsets exclusively.

To address these issues, a second phase of the Monte Carlo analysis was conducted, wherein the full set of 30 descriptors was used. Adherence to agent statistical behavior was achieved simply by drawing, without replacement, from the pool of actual targets. "Pseudo-perception" data were generated using the adaptive guessing strategy throughout, where the relatively unimportant first trial in a set, which contains no opportunity to respond to prior trials, was generated using the agent α's without regard to higher-order correlations.

Calculations were performed for 3×3, 5×5, 10×10, and 20×20 trial sets. It was judged unnecessary to calculate for larger datasets for two reasons. First, adaptive guessing cannot be used effectively between two percipients because percipients are not given feedback about other percipients' trials. That portion of the database in which trial-by-trial feedback was given must therefore be broken down yet further into individual percipient datasets, where the effect of large datasets on adaptive guessing is irrelevant. Second, the Monte Carlo results with the dataset sizes above were found to adhere with remarkable accuracy to a log-linear relation:

$$E = (0.858 \pm 0.013) - (0.219 \pm 0.006)\ln N, \qquad (1)$$

where E is the effect size per trial and N is the number of trials. The uncertainties on the regression parameters are those emerging from the regression calculation itself. Note that this is a decreasing function; the larger the database, the less benefit is gained from adaptive guessing. This can be explained in two related ways.

1. As the number of previously observed trials grows, the impact of the latest trial on the observed agent response frequencies must

lessen. Therefore the adaptive guessing algorithm converges toward a uniform version as the α's stabilize, and we have already seen that uniform guessing algorithms are useless.

2. Regardless of the details of the guessing algorithm, adaptive guessing gets its power from the nonrepetition of targets. Whether trials are volitional or instructed, they may be regarded as being choices from a large pool of "conceivable targets" spanning the space of every possible observable scene. The main difference is that the preparer of the instructed pool preselects a finite set of discrete points in that space, which are then drawn uniformly rather than with the possible biases of the agent. However, since the preparer of an instructed pool is likely to choose a wide variety of potential targets, the finite set will nevertheless do a fair job of spanning the full range of the configuration space. Potential biases on the part of the selector simply restrict the subspace spanned by the prepared pool and therefore need not be considered further. The first trial defines a point in this space of possible scenes, and no other trial is likely to be very close to it in this metric; the instructed pool is necessarily a very sparse sampling of the configuration space, whereas the volitional agent is highly unlikely to choose another target extremely similar to a prior one. Thus, in either protocol, by confining his choices to regions of "scene space" distant from the first target, the percipient can slightly enhance his probability of a good match on the second. However, once several targets have been seen, the set of previous targets will also span the configuration space, so that it is no longer possible to find a "preferred" region of higher probability on the basis of previous trials.

It might be suggested that, if the accumulation of previous information hinders the guessing, then an even simpler guessing strategy that only takes into account a few of the most recent trials should do better. However, the point of item 2 is that, when only a single point in the target space has been established, the probability distribution for the next target is distorted by the tendency to avoid this point; this is no longer the case when such points-to-avoid are scattered randomly through the target space. In fact, such "forgetful" algorithms do not perform as well as the one that keeps track of all previously revealed data.

The figures for adaptive guessing quoted throughout this analysis are based on an algorithm that outperforms any other strategy we have been able to define, and thus they present a worst case assessment of the potential spurious effect. A very important point is that these figures definitely overestimate the amount a given perci-

pient could benefit from adaptive guessing because of the limited information a percipient can gain. In the sequences of data potentially subject to adaptive guessing, most percipients were involved intermittently with several trials intervening between successive trials by a given percipient. Thus, a given percipient might often have several extraneous targets withdrawn from the pool between any two of his own responses. Since the stabilization of target characteristics progresses as the number of targets withdrawn increases, this diminishes the size of the fluctuation that the percipient may try to take advantage of. Furthermore, the targets viewed by other percipients may skew the remaining pool in ways unanticipated by a given percipient.

Given the dependence of adaptive-guessing effect size on subset size, it is clear that the 69 trials potentially susceptible to adaptive guessing cannot be treated *en bloc*; rather, they must be broken down by percipients. That is, one must identify the number of trials contributed by each percipient, compute the effect size potentially due to adaptive guessing for a block of that size, and then average the potential adaptive-guessing effect over the percipients. When this is done it is found that the total potential adaptive-guessing contribution to the final z score is $\Delta Z = 1.273$.

Further insight may be gained by breaking down the data into the ab initio (277 trials) and ex post facto (59 trials) as well as by vulnerability to adaptive guessing. The actual effect size for the observed anomalous yield shows interesting patterns:

Ab initio data for which adaptive guessing was possible: 41 trials, effect size 0.224, $Z = 1.433$.

Ab initio data with no opportunity for adaptive guessing: 236 trials, effect size 0.234, $Z = 3.600$.

Ex post facto data with possible adaptive guessing: 28 trials, effect size 1.035, $Z = 5.477$.

Ex post facto with no opportunity for adaptive guessing: 31 trials, effect size .422, $Z = 2.349$.

In other words, in the ab initio data, the dataset most vulnerable to the suspected bias shows a *smaller* anomalous effect than the remainder. In the ex post facto data, if the 28 trials where adaptive guessing was possible are broken down into percipient subsets and the expected effect sizes are calculated, the resulting effect size potentially due to adaptive guessing is 0.503, corresponding to $Z = 2.66$. This leaves an effect size of 0.532, corresponding to a Z of 2.82, unaccounted for. In other words: The worst-case potential for adaptive guessing based on trial-by-trial feedback in the part of the

ex post facto database where such feedback was given corresponds to about half the effect size; this leaves, even in the vulnerable dataset, an effect size comparable to that in the rest of the data, though still somewhat stronger. This remaining anomaly, unexplainable by trial-by-trial feedback, produces the z score of 2.82 mentioned above for the 28 potentially susceptible trials alone. This feeds through to a final z score of 3.64 (instead of 5.79) for the ex post facto data as a whole and thus of 5.59, instead of 6.36, for the entire database.

It should be reemphasized that the above discussion is very much a worst-case analysis. Extending it further by assuming that adaptive guessing did occur using the most effective strategy in the potentially vulnerable ab intio data as well, the overall z score would only be reduced to 5.082 ($p = 1.87 \times 10^{-7}$). But again, the ab initio data refute such an assumption; if adaptive guessing is present, it is evidently ineffective; in fact, it is worse than useless. Likewise, while we have learned that adaptive guessing strategies could have some spurious impact in the ex post facto data, the remote perception database is nonetheless sufficiently robust to retain $p \approx 10^{-8}$ after the largest possible correction for this.

References

Bisaha, J. P., & Dunne, B. J. (1979). Multiple subject and long-distance precognitive remote viewing of geographical locations. In C. T. Tart, H. E. Puthoff, & R. Targ (Eds.), *Mind at large* (pp. 107–124). New York, Praeger.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters, an introduction to design, data analysis, and model building*. New York: John Wiley & Sons.

Dobyns, Y. H. (1992). The permutation counting fallacy. (Technical Note PEAR 92001). Princeton, NJ: Princeton Engineering Anomalies Research, School of Engineering and Applied Science, Princeton University.

Dunne, B. J., & Bisaha, J. P. (1979). Precognitive remote viewing in the Chicago area: A replication of the Stanford experiment. *Journal of Parapsychology*, **43**, 17–30.

Dunne, B. J., Dobyns, Y. H. & Intner, S. M. (1989). Precognitive remote perception. III: Complete binary data base with analytical refinements. (Technical Note PEAR 89002). Princeton, NJ: Princeton Engineering Anomalies Research, School of Engineering and Applied Science, Princeton University.

Dunne, B. J., Jahn, R. G., & Nelson, R. D. (1983). Precognitive remote perception. (Technical Note PEAR 83003). Princeton, NJ: Princeton Engineering Anomalies Research, School of Engineering and Applied Science, Princeton University.

JAHN, R. G. (1983). On the representation of psychic research to the community of established sciences. (Technical Note PEAR 83004). Princeton, NJ: Princeton Engineering Anomalies Research, School of Engineering and Applied Science, Princeton University.

JAHN, R. G. (1989). Anomalies: Analysis and aesthetics. *Journal of Scientific Exploration*, **3** (1), 15–26.

JAHN, R. G. (1991). Psi at the Savoy. *Journal of Indian Psychology*, **9,** 14–23.

JAHN, R. G., & DUNNE, B. J. (1987). *Margins of reality: The role of consciousness in the physical world.* San Diego: Harcourt Brace Jovanovich.

JAHN, R. G., DUNNE, B. J., & JAHN, E. G. (1980). Analytical judging procedure for remote perception experiments. *Journal of Parapsychology*, **44,** 207–231.

JAHN, R. G., DUNNE, B. J., NELSON, R. D., JAHN, E. G., CURTIS, T. A., & COOK, I. A. (1982). Analytical judging procedure for remote perception experiments. II: Ternary coding and generalized descriptors. (Technical Note PEAR 82002). Princeton, NJ: Princeton Engineering Anomalies Research, School of Engineering and Applied Science, Princeton University.

SCHLITZ, M., & GRUBER, E. (1981). Transcontinental remote viewing: A rejudging. *Journal of Parapsychology*, **45,** 233–237.

SNEDECOR, G. W., & COCHRAN, W. G. (1980). *Statistical methods, seventh edition.* Ames: Iowa State University Press.

WHALEN, A. D. (1971). *Detection of signals in noise.* New York: Academic Press.

Princeton Engineering Anomalies Research
C-131, School of Engineering and Applied Science
Princeton University
Princeton, NJ 08544-5263