*Journal of Scientific Exploration,* Vol. 14, No. 4, pp. 499–555, 2000 0892-3310/00


# Mind/Machine Interaction Consortium: PortREG Replication Experiments

R. JAHN, B. DUNNE, G. BRADISH, Y. DOBYNS, A. LETTIERI, AND R. NELSON

*Princeton Engineering Anomalies Research*
*Princeton University, Princeton, NJ*

J. MISCHO, E. BOLLER, AND H. BÖSCH

*Freiburg Anomalous Mind/Machine Interactions*
*Institut für Grenzgebiete der Psychologie und Psychohygiene e.V., Freiburg, Germany*

D. VAITL, J. HOUTKOOPER, AND B. WALTER

*Giessen Anomalies Research Project*
*Justus-Liebig-Universität Giessen, Giessen, Germany*

**Abstract—**A consortium of research groups at Freiburg, Giessen, and Princeton was formed in 1996 to pursue multidisciplinary studies of mind/machine interaction anomalies. The first collaborative project undertaken was an attempted replication of prior Princeton experiments that had demonstrated anomalous deviations of the outputs of electronic random event generators in correlation with prestated intentions of human operators. For this replication, each of the three participating laboratories collected data from 250 × 3000–trial × 200 binary-sample experimental sessions, generated by 227 human operators. Identical noise-source equipment was used throughout, and essentially similar protocols and data analysis procedures were followed. Data were binned in terms of operator intention to increase the mean of the 200-binary-sample distributions (HI); to decrease the mean (LO); or not to attempt any influence (BL). Contiguous unattended calibrations were carried forward throughout. The agreed upon primary criterion for the anomalous effect was the magnitude of the HI–LO data separation, but data also were collected on a number of secondary correlates. The primary result of this replication effort was that whereas the overall HI–LO mean separations proceeded in the intended direction at all three laboratories, the overall sizes of these deviations failed by an order of magnitude to attain that of the prior experiments, or to achieve any persuasive level of statistical significance. However, various portions of the data displayed a substantial number of interior structural anomalies in such features as a reduction in trial-level standard deviations; irregular series-position patterns; and differential dependencies on various secondary parameters, such as feedback type or experimental run length, to a composite extent well beyond chance expectation. The change from the systematic, intention-correlated mean shifts found in the prior studies, to this polyglot pattern of structural distortions, testifies to inadequate understanding of the basic phenomena involved and suggests a need for more sophisticated experiments and theoretical models for their further elucidation.

*Keywords*: random event generator (REG) — random event experiments — human/machine anomalies — mind/machine interactions

## Table of Contents



## List of Tables

MAIN TEXT

## List of Figures



## I. Context and Background

### *A. History and Organization*

Electronic random event generators (REGs) have long been used in a wide range of laboratory experiments designed to test the hypothesis that human consciousness may interact directly with random physical systems (Radin & Nelson, 1989; Schmidt, 1970). The results have provided strong statistical evidence that the mean outputs of these devices can deviate from chance expectation in direct correlation with prestated intentions of the participants and that aberrations in various other features of the output count distributions may reflect subtler aspects of the human/machine interactions. Over the past two decades, the Princeton Engineering Anomalies Research Laboratory (PEAR) has produced very large databases in REG experiments of this class (Nelson, Bradish, & Dobyns, 1989), which have further confirmed the existence of these types of human/machine anomalies, and have indicated some of their physical and psychological characteristics (Jahn, Dobyns, & Dunne, 1991; Jahn et al., 1997).

While these PEAR experiments have constituted extensive conceptual replications of earlier work elsewhere (Bierman & Houtkooper, 1975; Radin & Nelson, 1989; Rhine & Humphrey, 1944; Schmidt, Morris, & Rudolph, 1986) and also have included many internal replications within themselves, it was

felt that more might be learned from further, more broadly based studies of similar character and comparable controls, conducted in collaboration with other researchers having complementary professional interests and experience. For this purpose, a consortium of laboratories was assembled in 1996, comprising the Freiburg Anomalous Mind/Machine Interactions group (FAMMI) at the Institut für Grenzgebiete der Psychologie und Psychohygiene (IGPP) in Freiburg, the Giessen Anomalies Research Project (GARP) in the Center for Psychobiology and Behavioral Medicine at Justus-Liebig-Universität Giessen, and the PEAR Laboratory at Princeton University. The primary agenda of this "Mind/Machine Interaction Consortium" was a program of professional interaction and shared technology that would broaden and deepen our collective understanding of these consciousness-related anomalous phenomena.

As an initial effort to establish sound and effective strategies for long-term collaboration, it was agreed that the first project to be addressed would be an extensive, commensurate repetition of prior PEAR REG experiments, conducted contemporaneously in all three locations. The first phase of this project was to be as strict a replication as feasible, given the essential differences of structure and style of the three laboratories. At the same time, it was to provide a platform for developing and deploying effective shared technologies, protocols, database acquisition and management techniques, and interlaboratory and interpersonal communications that would enable productive longer-term collaborations. A second phase of the project also was planned that would accommodate the three laboratories' specialized interests and capabilities in psychological, psychophysiological, and engineering investigations, respectively, but this article shall deal only with Phase I.

### *B. Prior PEAR Experience*

*1. Equipment.* Over its many years of mind/machine experimentation, the PEAR program has developed several versions of electronic random event generators, utilizing different primary sources of noise but maintaining important common features of design. An original "benchmark" experiment employed a commercial random source sold by Elgenco, Inc. The core of this module is proprietary, but Elgenco's engineering staff describe it as "solid state junctions with precision preamplifiers," implying processes that rely on quantum tunneling to produce unpredictable, broad-spectrum noise in the form of low-amplitude voltage fluctuations. A much simpler and more compact REG, termed "PortREG," was developed subsequently, based on thermal noise in resistors, which also produces a well-behaved, broad-spectrum fluctuation. A yet later-generation device, called "MicroREG," uses a field effect transistor for the primary noise source, again relying on quantum tunneling to provide uncorrelated fundamental events that compound to an unpredictable voltage fluctuation.

In all cases, the electronic process begins with a white-noise frequency dis-

tribution. For example, the benchmark REG, on which most of the prior data were acquired, presents a flat spectrum, +/– 1 dB, from 50 Hz to 20 kHz. A subsequent 1000-Hz low-end cutoff attenuates frequencies below the data-sampling rate. This filtering, followed by appropriate amplification and clipping, produces an approximately rectangular wave train with unpredictable temporal spacing. Gated sampling, typically at 1-kHz, then yields a regularly spaced sequence of randomly alternating +/– bits, suitable for rapid counting. To eliminate biases from such environmental stresses as temperature change or component aging, "exclusive or" (XOR) masks are applied to the digital data streams in regularly alternating +/– patterns. In the experiments, output data are presented and recorded in "trials" that are the sum of $N$ samples (typically 200 bits) from the primary sequence, thus mitigating any residual short-lag autocorrelations. The final output of the benchmark REG thus is a sequence of conditioned bits and, in the later devices, of bytes, presented to the computer's serial port, which then are collected into a sequence of trials, usually presented at approximately one trial per second. Calibrations of all of the devices conform to statistical chance expectations for the mean, standard deviation, skewness, and kurtosis of the accumulated trial-score distributions, and for time-series of independent events (cf. Appendix I).

2. *Experimental design and results.* The basic experimental designs embody further protocol-level protections against artifacts. Using a "tripolar" protocol, participants generate data under three conditions of prespecified intention, namely to achieve high (HI) or low (LO) output distribution mean values, or to generate baseline (BL) data. With the exception of these expressed intentions, which are immutably prerecorded in the experiments' computer files, all other potentially influential protocol variables are maintained constant within an experimental session.

In addition to the primary variable of tripolar intention, a number of secondary parameters are available as options that can be explored in separate sessions and assessed as factors that may contribute to the experimental outcomes. These include *human variables*, such as the identities of the individual operators, their gender, the number co-operating in the effort, and whether they are "prolific," i.e., have accumulated sufficient data to permit robust internal comparisons of their results; *technical variables*, such as the different noise sources, including not only the physical random sources described but also various hardwired and algorithmic pseudorandom generators, designated as nondeterministic and deterministic sources, respectively; *operational variables*, including information density (bits per second); the number of trials in automatically sequenced "runs;" the instruction mode (volitional or instructed); the type of feedback provided to the operator, etc.; and *physical variables*, including the spatial separation of the operator from the machine (up to thousands of miles) and temporal separations between operator attempts and actual operation of the devices (up to several hours or even a few days).

For the purposes of the replication studies reported here, we shall refer mainly to that segment of previous PEAR data provided by individual opera-

tors adjacent to "benchmark" REG equipment. These "local, single-operator" experiments, contributed over 12 years by 91 participants, constituted 522 replications at the "series" or "session" level, comprising nearly two and a half million, 200-bit trials. The primary results of this segment are summarized as "Prior PEAR Data" in Table 0. This database also was subjected to a broad range of subordinate analytical tests, including specific searches for indicative structural details and broad-based analyses of variance, all of which have been extensively reported in the archival literature and supporting technical reports (Jahn et al., 1997; Dunne, 1991; Dunne et al., 1994; Nelson et al., 2000) and will be reviewed as appropriate in the following text.

In passing, it might be noted that these particular experiments were complemented by an array of studies that used many other forms of random generator equipment and protocols (Jahn, Dunne, & Nelson, 1987), including the much more compact "PortREG" devices chosen for the replication program to follow, several macroscopic mechanical analogs (Dunne, Nelson, & Jahn, 1988; Nelson et al., 1994), various pseudorandom devices (Jahn et al., 1997), "remote" and "off-time" protocols (Dunne & Jahn, 1992), and nonintentional "FieldREG" experiments (Nelson et al., 1996, 1998). Results of these studies were generally consistent and collectively extended the statistical significance of the entire program by several orders of magnitude (Jahn et al., 1997).

Many human/machine experiments of this sort have been conducted at other laboratories, and most of these have yielded commensurate anomalous results (Radin, 1997). Related studies have also demonstrated responses from biological substances or living organisms employed as the random targets of the operators' intentions (Braud, 1993; Braud & Dennis, 1989; Grad, 1963). In some cases, the role of the operators has been played by other than human species, e.g., by chicks, rabbits, and mice, many of whom seem capable of eliciting anomalous correlations of machine behavior with their biological or emotional needs (Peoc'h, 1995). From this array of empirical studies, it appears that operator desire is capable of establishing observable relationships to the outputs of such random physical systems, by some unknown means that is largely independent of the nature of the device and also independent of the intervening distance and time. The ubiquitous character of these anomalies bespeaks broad potential importance to contemporary scientific understanding and to individual and cultural welfare.

## II. Consortium Replication

### *A. Experimental Design*

At a planning meeting held shortly after the inception of the Mind/Machine Consortium, the members decided to undertake yet another replication of this class of REG experiments. Second-generation PortREG technology was selected for the random source because of its simplicity, portability, and relatively low cost, with confidence of its efficacy based on various indications from

preceding PEAR research and that of others that these anomalous effects are independent of the source of randomness (Jahn et al., 1997; Schmidt & Pantas, 1972). All three laboratories would employ identical protocols and data-processing techniques, to the extent feasible given the differing languages, disciplinary backgrounds, and skills. Although the primary hypothesis to be tested was confirmation of the earlier PEAR results on a simple HI–LO mean-shift criterion, secondary investigations were to provide structural data on the characteristics and correlates of the phenomena. Specifically, it was agreed that each laboratory would use large pools of operators to accumulate 250 experimental "sessions" or "series," each series consisting of 1000 × 200-sample trials in each of the HI, LO, and BL intentions and, in addition, would extract whatever structural aspects of the data befit its capabilities, such as separate HI, LO, and BL performances, gender effects, serial position effects, standard deviations, feedback correlations, experimenter effects, etc.

### *B. Experimental Results*

This section presents all of the pertinent data generated by the three laboratories in as commensurate and complete a format as possible here. We begin with a table key and brief explanatory text regarding the tabular formats. Then follows a sequence of tables that summarize the overall results of each laboratory with respect to the primary HI–LO mean-shift hypothesis, followed by a concatenation of all three databases. For comparison, these tables are preceded by similar representations of the earlier PEAR data and of the contemporaneous calibration data described in more detail in Appendix I. Following these summary tables, we then display a large array of explorations into data distribution structures and secondary parameter correlations attempted by each laboratory, both individually and collectively.

*1. Tabular key and comments.* All of the following tables use a common statistical notation:

$\Delta\mu$ = Shift in empirical trial-level mean from chance expectation of 100 (also called "effect size")

$\sigma$ = Empirical trial-level standard deviation

$Z$ = Standardized *Z-score* of the mean-shift, calculated as:

$$Z = \frac{\Delta\mu}{\sigma_0}\sqrt{N_t}$$

where

$N_t$ = Number of experimental trials

$\sigma_0$ = Theoretical chance standard deviation for 200-sample trials (7.071)

$Z_{diff}$ = *Z-score* for differences of any two indicated data subsets (see text below)

$\chi^2$ = Chi-squared statistic to test for goodness-of-fit of empirical data values to a comparison standard

The first set of tables presents the overall results of the entire PortREG-replication database for all three intentions and for the HI–LO, or $\Delta$, criterion which is regarded as the primary variable. The mean shift for the $\Delta$ column is calculated by inverting the LO data with respect to the mean and concatenating them with the HI, so it is the average, rather than the sum, of the mean shifts in the intended directions. This is intended to make statistical comparisons easier by preventing intrinsic differences of scale between the separate intentions and the $\Delta$ values. This representation makes the $\Delta$ data effectively a single large pool of trials which also have a theoretically expected mean of 100 and standard deviation of $(50)^{1/2} = 7.071$, and in which a positive mean shift corresponds to success in the direction of intention. For all of the replication series, $N_t$ comprises 1000 trials per intention. (The earliest prior PEAR data were taken in larger series of 5000 trials per intention; the series size subsequently was reduced in stages to a standard of 1000 trials per intention, which allowed the experiment to be completed in a single session. Thus, the total number of prior PEAR trials is considerably greater than one would infer from the counts of series listed in the table; e.g., the prior PEAR database is approximately equivalent to 834 PortREG series.)

For the sequence of structural tables that follow, $Z_{diff}$ refers to the differences in mean shifts, computed as follows: Given two populations, $N_1$ and $N_2$, having $Z$-scores $Z_1$ and $Z_2$, we may compute a normalized effect size for each as $\varepsilon_i = Z_i/(N_t)^{1/2}$, which is related to $\Delta\mu$ by a multiplicative constant, e.g., $\varepsilon_i = \Delta\mu_i/(50)^{1/2}$. The uncertainty associated with a $Z$-score is always 1 by construction, so the $\varepsilon_i$ have measurement uncertainties $\sigma_i = 1/(N_i)^{1/2}$. The standard normal deviate, or $Z$-score, for a difference between sets 1 and 2 is therefore the difference $\varepsilon_1 - \varepsilon_2$ divided by the uncertainty of this difference, $\sigma_d$, which is simply the sum in quadrature of the individual uncertainties: $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$. This can be reduced to an expression in the original $N$s and $Z$s:

$$Z_{diff} = \frac{\varepsilon_1 - \varepsilon_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{Z_1/\sqrt{N_1} - Z_2\sqrt{N_2}}{\sqrt{1/N_1 + 1/N_2}} = \frac{Z_1\sqrt{N_2} - Z_2\sqrt{N_1}}{\sqrt{N_1 + N_2}} \qquad (1)$$

For most of these presentations and the associated discussions, we have chosen to use only $Z$-scores without associated tail-probability ($p$) values, on the grounds that the former are completely unambiguous, depending only on the statistical character of the data used, whereas $p$-values require subjective and occasionally contentious decisions regarding the appropriateness of one- or two-tailed statistics, primary or secondary analyses, Bonferroni corrections for multiple or prospective analyses, and so forth. The direct correspondence of the $Z$ values to particular "tail-probabilities" is, of course, well tabulated, e.g.:

| $Z$ | 1.6449 | 1.9600 | 2.3263 | 2.5758 | 3.0902 |
|---|---|---|---|---|---|
| $P_z$ (1-tail) | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
| $P_z$ (2-tail) | 0.10 | 0.05 | 0.02 | 0.01 | 0.002 |

*2. Primary data summary.* The mean-shift results and their associated standard deviations, obtained by each of the laboratories in each direction of intention, are summarized in Tables 0, 00, F.1, G.1, P.1, and C.1.

It is immediately clear from these summary data that although the mean HI–LO separations found by each of the laboratories all proceed in the intended direction, they fail by an order of magnitude to reach the level of the prior PEAR data or any persuasive level of statistical significance. The implications of this result are discussed at length in Section III. Nonetheless, subtler structural anomalies, such as the almost universal depression of the trial-level standard deviations below the theoretical and calibration values, are already evident in the summary tables above, and invite more detailed searches for other secondary correlates. The following tables display the results of such examinations (cf. section III.B.5).

*3. Structural data.*

*(a) Gender effects.* In the following sequence of tables, we display breakdowns of the laboratory data in terms of various subordinate secondary parameters that proved instructive in the prior PEAR studies. As a first example, the distinctions between male and female operator performance that were studied extensively in the early work (Dunne, 1998) are broken down here by laboratory and intention (Tables F.2, G.2, P.2, and C.2). Other than the differ-

TABLE 0
Prior PEAR Laboratory Data (522 Series, 91 Operators)

| Measure | BL | LO | HI | $\Delta$ |
|---|---|---|---|---|
| $\Delta\mu$ | 0.013372 | -0.015586 | 0.025994 | 0.020800 |
| $\sigma$ | 7.074 | 7.069 | 7.070 | 7.070 |
| $Z$ | 1.7132 | -2.0161 | 3.3688 | 3.8087 |

TABLE 00
Concurrent Calibrations (1049 Series)

| Measure | Theory | FAMMI | GARP | PEAR |
|---|---|---|---|---|
| $\Delta\mu$ | 0.000000 | -0.000901 | 0.000166 | -0.000207 |
| $\sigma$ | 7.0711 | 7.0753 | 7.0691 | 7.0697 |
| $Z$ | 0.0000 | -0.1175 | 0.0253 | -0.0305 |

TABLE F.1
All FAMMI Data (250 Series, 80 Operators)

| Measure | BL | LO | HI | $\Delta$ |
|---|---|---|---|---|
| $\Delta\mu$ | -0.002308 | -0.006496 | 0.006336 | 0.006416 |
| $\sigma$ | 7.0550 | 7.0642 | 7.0713 | 7.0678 |
| $Z$ | -0.1632 | -0.4593 | 0.4480 | 0.6416 |

TABLE G.1
All GARP Data (250 Series, 69 Operators)

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| $\Delta\mu$ | 0.004116 | -0.012596 | -0.00808 | 0.002258 |
| $\sigma$ | 7.0559 | 7.0418 | 7.0713 | 7.0566 |
| $Z$ | 0.2910 | -0.8907 | -0.5713 | 0.2258 |

TABLE P.1
All PEAR Data (250 Series, 78 Operators)

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| $\Delta\mu$ | 0.001216 | 0.004836 | 0.008148 | 0.001656 |
| $\sigma$ | 7.0617 | 7.0608 | 7.0622 | 7.0615 |
| $Z$ | 0.0860 | 0.3420 | 0.5762 | 0.1656 |

TABLE C.1
Concatenation Across All Laboratories (750 Series, 227 Operators)

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| $\Delta\mu$ | 0.001008 | -0.004752 | 0.002135 | 0.003443 |
| $\sigma$ | 7.0575 | 7.0556 | 7.0683 | 7.0619 |
| $Z$ | 0.1235 | -0.5820 | 0.2614 | 0.5964 |

ence between single-operator and dual-operator performance, which was explored only by the PEAR group, the only remarkable gender differences evident in the concatenated data are in the baseline results. Most of this effect is contributed by the FAMMI and GARP operators, with little assistance from PEAR, despite the prominence of such a disparity, albeit with opposite sign, in the prior PEAR experience (Dunne, 1998). Also possibly worth noting are the almost uniformly higher standard deviations of the female operators.

*(b) Assignment effects.* As one element in a broad search for subjective or psychological correlates, the data have been divided into those trials wherein the directional intention of the operator was assigned by an auxiliary random process of some sort (Instructed), and those for which the operator selected the direction (Volitional), within the constraints of balanced numbers of HI, LO, and BL trials (Tables F.3, G.3, P.3, and C.3). Here, one subset of data emerges as disparate; the GARP experiments in the Instructed mode show very significant anticorrelation with intention, in contrast to the corresponding Volitional data, which correlate positively. The difference $Z$-score is highly significant by any reasonable criterion. However, similar effects are not found in the FAMMI or PEAR data, leaving the concatenated data less impressive in this distinction.

*(c) Feedback effects.* The feedback presented to the operator is another subjective correlate previously examined at PEAR. The alternatives here are (a) a

TABLE F.2
Gender Effects in FAMMI Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Male operators (150 series, 40 operators) | | | | |
| $\Delta\mu$ | 0.029607 | 0.001320 | 0.002567 | 0.000623 |
| $\sigma$ | 7.0542 | 7.0616 | 7.0673 | 7.0644 |
| $Z$ | 1.6216 | 0.0723 | 0.1406 | 0.0483 |
| Female operators (100 series, 40 operators) | | | | |
| $\Delta\mu$ | -0.050180 | -0.018220 | 0.011990 | 0.015105 |
| $\sigma$ | 7.0559 | 7.0680 | 7.0775 | 7.0727 |
| $Z$ | -2.2441 | -0.8148 | 0.5362 | 0.9553 |
| Differences | | | | |
| $Z_{diff}$ (F - M)$\mu$ | -2.7639 | -0.6769 | 0.3264 | 0.7095 |

TABLE G.2
Gender Effects in GARP Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Male operators (124 series, 35 operators) | | | | |
| $\Delta\mu$ | 0.023645 | -0.003234 | -0.005306 | -0.001036 |
| $\sigma$ | 7.0493 | 7.0414 | 7.0776 | 7.0595 |
| $Z$ | 1.1775 | -0.1610 | -0.2643 | -0.0730 |
| Female operators (126 series, 34 operators) | | | | |
| $\Delta\mu$ | -0.015103 | -0.021810 | -0.010810 | 0.005500 |
| $\sigma$ | 7.0624 | 7.0422 | 7.0651 | 7.0536 |
| $Z$ | -0.7582 | -1.0948 | -0.5426 | 0.3905 |
| Differences | | | | |
| $Z_{diff}$ (F - M)$\mu$ | -1.3699 | -0.6567 | -0.1946 | -0.3268 |

graphic display showing the cumulative deviation; (b) a digital display with large numbers showing the current trial and running mean; and (c) no feedback at all, with results reported only at the end of the experimental run (Tables F.4, G.4, P.4, and C.4). Here we find several noteworthy entries in the GARP data, two in the digital subset ($\Delta$ and BL) and one in the no-feedback subset (HI), and two in the FAMMI data, in the HI-digital and HI no-feedback subsets, all of which feed through to their difference values, and are sufficient to drive several significant excursions in the concatenated data.

*(d) Runlength effects.* Since the duration of the experimental runs that require steady attention of the operators conceivably might introduce subjective factors such as boredom, distraction, and anxiety, alternatives of 100-trial (1.5-minute) and 1000-trial (15-minute) runs were admitted into the protocols (Tables F.5, G.5, P.5, and C.5). Noteworthy here are the differences between the two run lengths in the LO-intention GARP data, which, supported modestly by the corresponding PEAR data, feed through to a marginally interesting concatenation value.

TABLE P.2
Gender Effects in PEAR Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Male operators (126 series, 36 operators) | | | | |
| $\Delta\mu$ | 0.005063 | 0.017333 | 0.018103 | 0.000385 |
| $\sigma$ | 7.0591 | 7.0578 | 7.0536 | 7.0557 |
| $Z$ | 0.2542 | 0.8701 | 0.9088 | 0.0273 |
| Female operators (76 series, 22 operators) | | | | |
| $\Delta\mu$ | -0.013539 | 0.006382 | -0.039263 | -0.022822 |
| $\sigma$ | 7.0680 | 7.0673 | 7.0733 | 7.0703 |
| $Z$ | -0.5279 | 0.2488 | -1.5308 | -1.2583 |
| Multiple operators (48series, 20 operators) | | | | |
| $\Delta\mu$ | 0.014479 | -0.030417 | 0.057083 | 0.043750 |
| $\sigma$ | 7.0584 | 7.0581 | 7.0673 | 7.0627 |
| $Z$ | 0.4486 | -0.9424 | 1.7687 | 1.9170 |
| Differences | | | | |
| $Z_{diff}$ (F - M)$\mu$ | -0.5728 | -0.3372 | -1.7664 | -1.0106 |
| $Z_{diff}$ (1 - 2)$\mu$ | -0.4572 | 1.2151 | -1.6868 | -2.0519 |

Note: The Gender parameter at PEAR is treated as three-valued rather than two-valued, since operator pairs also contributed to the replication database. Rather than doing three $Z_{diff}$ comparisons, one set of comparisons between males and females, and separate comparisons between combined results of individual operators and the multi-operator database, are presented.

TABLE C.2
Gender Differences in Concatenated Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Male operators (400 series, 111 operators) | | | | |
| $\Delta\mu$ | 0.020028 | 0.004953 | 0.005020 | 0.000034 |
| $\sigma$ | 7.0542 | 7.0542 | 7.0662 | 7.0602 |
| $Z$ | 1.7913 | 0.4430 | 0.4490 | 0.0043 |
| Female operators (302 series, 96 operators) | | | | |
| $\Delta\mu$ | -0.026325 | -0.013526 | -0.010421 | 0.001553 |
| $\sigma$ | 7.0617 | 7.0570 | 7.0713 | 7.0642 |
| $Z$ | -2.0459 | -1.0512 | -0.8099 | 0.1707 |
| Differences | | | | |
| $Z_{diff}$ (F - M)$\mu$ | -2.7192 | -1.0841 | -0.9058 | 0.1260 |

*(e) Series-position effects.* Since subjective issues of boredom, anxiety, overconfidence, and learning also might manifest in the operator's performance over more major blocks of experimental effort, data also have been processed on a series-by-series basis, in a search for some definitive series-position pattern, such as that found in the prior PEAR studies (Dunne et al., 1994). In the following tables, the column labeled *N* lists the number of operators completing that number of series, and the notation 5+ denotes the combined results of all series numbered 5 and higher. Those PEAR and GARP operators who had previously performed five or more series or their equivalent on any similar REG experiments were regarded as contributing replication se-

TABLE F.3
Assignment Effects in FAMMI Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Instructed (58 series, 23 operators) | | | | |
| $\Delta\mu$ | 0.027466 | -0.015397 | 0.027017 | 0.021207 |
| $\sigma$ | 7.0534 | 7.0656 | 7.1087 | 7.0872 |
| $Z$ | 0.9354 | -0.5244 | 0.9202 | 1.0215 |
| Volitional (192 series, 80 operators) | | | | |
| $\Delta\mu$ | -0.011302 | -0.003807 | 0.000089 | 0.001948 |
| $\sigma$ | 7.0554 | 7.0637 | 7.0600 | 7.0619 |
| $Z$ | -0.7004 | -0.2359 | 0.0055 | 0.1707 |
| Differences | | | | |
| $Z_{diff}$ (I - V)$\mu$ | 1.1571 | -0.3459 | 0.8038 | 0.8129 |

TABLE G.3
Assignment Effects in GARP Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Instructed (26 series, 17 operators) | | | | |
| $\Delta\mu$ | -0.024269 | 0.082192 | -0.100192 | -0.091192 |
| $\sigma$ | 6.9907 | 7.0870 | 7.0559 | 7.0714 |
| $Z$ | -0.5534 | 1.8743 | -2.2847 | -2.9409 |
| Volitional (224 series, 69 operators) | | | | |
| $\Delta\mu$ | 0.007411 | -0.023598 | 0.002612 | 0.013105 |
| $\sigma$ | 7.0635 | 7.0365 | 7.0730 | 7.0548 |
| $Z$ | 0.4960 | -1.5795 | 0.1748 | 1.2405 |
| Differences | | | | |
| $Z_{diff}$ (I - V)$\mu$ | -0.6838 | 2.2835 | -2.2190 | -3.1838 |

ries only in the 5+ category (Tables F.6, G.6, P.6, and C.6). Interpretation of this disparate array of results is deferred until section III.3.

*(f) Individual laboratory explorations.* Some parameter or protocol options were explored by only one of the three laboratories having a particular interest in that factor, leaving no possibilities of interlaboratory concatenations. For example, Table F.7 lists the FAMMI data acquired under supervision of various experimenters. Numbers 1, 2, and 3 refer to three particular individuals; Group 4 subsumes several incidental experimenters. No remarkable individual scores appear, and a $\chi^2$ statistic computed by summing the squares of the $Z$-scores in each subset shows no evidence of significant differences in behavior. In fact, the $\chi^2$ for the HI intention is so small as to suggest anomalous consistency ($p = 0.980$).

In Table G.7 are listed the results of a GARP investigation of the importance of the control of the REG trials by an automatic sequencer *vs.* allowing the operator to initiate each trial ad libidum. No sensitivity to this option appears in these data.

TABLE P.3
Assignment Effects in PEAR Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Instructed (133 series, 45 operators) | | | | |
| $\Delta\mu$ | 0.007241 | 0.008346 | -0.002594 | -0.005470 |
| $\sigma$ | 7.0617 | 7.0586 | 7.0614 | 7.0600 |
| $Z$ | 0.3734 | 0.4304 | -0.1338 | -0.3990 |
| Volitional (117 series, 52 operators) | | | | |
| $\Delta\mu$ | -0.005632 | 0.000846 | 0.020359 | 0.009756 |
| $\sigma$ | 7.0617 | 7.0633 | 7.0632 | 7.0632 |
| $Z$ | -0.2725 | 0.0409 | 0.9848 | 0.6674 |
| Differences | | | | |
| $Z_{diff}$ (I - V)$\mu$ | 0.4542 | 0.2646 | -0.8098 | -0.7598 |

TABLE C.3
Assignment Effects in Concatenated Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Instructed (217 series, 85 operators) | | | | |
| $\Delta\mu$ | 0.008871 | 0.010848 | -0.006373 | -0.008611 |
| $\sigma$ | 7.0510 | 7.0639 | 7.0735 | 7.0687 |
| $Z$ | 0.5844 | 0.7146 | -0.4199 | -0.8022 |
| Volitional (533 series, 201 operators) | | | | |
| $\Delta\mu$ | -0.002193 | -0.011103 | 0.005598 | 0.008351 |
| $\sigma$ | 7.0602 | 7.0522 | 7.0662 | 7.0592 |
| $Z$ | -0.2264 | -1.1464 | 0.5780 | 1.2193 |
| Differences | | | | |
| $Z_{diff}$ (I - V)$\mu$ | 0.6145 | 1.2191 | -0.6649 | -1.3322 |

Finally, in Table G.8 are presented GARP results for four classes of operators: those selected and processed in a formal fashion; members of the research staff; students in the laboratory; and casual visitors. Here the only striking disparity is contributed by the visitor category in the BL intention, leading to a slightly elevated $\chi^2$ indicator for that condition.

*(g) Temporal evolution of effect sizes.* As an alternative representation of the full replication databases, Figures 1 through 3 present sets of cumulative deviation graphs that summarize the historical evolution of each laboratory's compounding results for the mean shifts under HI, LO, and BL intentions. For comparison, Figure 4 shows similar plots of the prior PEAR results. Figure 5 compares cumulative deviations of the HI–LO separations for each of the three laboratories. In all of these figures, the dotted parabolic envelopes are the loci of cumulative deviations corresponding to one-tailed chance probabilities of .05 at the given abscissa.

TABLE F.4
Feedback Effects in FAMMI Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Digital (7 series, 3 operators) | | | | |
| $\Delta\mu$ | 0.055571 | 0.087000 | 0.173286 | 0.043143 |
| $\sigma$ | 7.0474 | 7.0079 | 7.1947 | 7.1029 |
| $Z$ | 0.6575 | 1.0294 | 2.0503 | 0.7219 |
| Graphic (229 series, 80 operators) | | | | |
| $\Delta\mu$ | 0.000642 | -0.012694 | -0.005419 | 0.003638 |
| $\sigma$ | 7.0531 | 7.0629 | 7.0640 | 7.0635 |
| $Z$ | 0.0434 | -0.8591 | -0.3667 | 0.3481 |
| None (14 series, 8 operators) | | | | |
| $\Delta\mu$ | -0.079500 | 0.048143 | 0.115143 | 0.033500 |
| $\sigma$ | 7.0900 | 7.1121 | 7.1275 | 7.1202 |
| $Z$ | -1.3303 | 0.8056 | 1.9267 | 0.7928 |
| Differences | | | | |
| $Z_{diff}$(D - G)$\mu$ | 0.6402 | 1.1620 | 2.0829 | 0.6512 |
| $Z_{diff}$(D - N)$\mu$ | 1.3049 | 0.3754 | 0.5617 | 0.1317 |
| $Z_{diff}$(G - N)$\mu$ | 1.3018 | -0.9882 | -1.9584 | -0.6861 |

TABLE G.4
Feedback Effects in GARP Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Digital (50 series, 37 operators) | | | | |
| $\Delta\mu$ | 0.058980 | -0.042820 | 0.045300 | 0.044060 |
| $\sigma$ | 7.0518 | 7.0416 | 7.0625 | 7.0520 |
| $Z$ | 1.8651 | -1.3541 | 1.4325 | 1.9704 |
| Graphic (189 series, 69 operators) | | | | |
| $\Delta\mu$ | -0.003709 | -0.001127 | -0.015365 | -0.007119 |
| $\sigma$ | 7.0572 | 7.0425 | 7.0735 | 7.0580 |
| $Z$ | -0.2280 | -0.0693 | -0.9447 | -0.6190 |
| None (11 series, 10 operators) | | | | |
| $\Delta\mu$ | -0.110818 | -0.072273 | -0.125546 | -0.026636 |
| $\sigma$ | 7.0518 | 7.0300 | 7.0723 | 7.0517 |
| $Z$ | -1.6437 | -1.0720 | -1.8621 | -0.5587 |
| Differences | | | | |
| $Z_{diff}$(D - G)$\mu$ | 1.7629 | -1.1725 | 1.7060 | 2.0354 |
| $Z_{diff}$(D - N)$\mu$ | 2.2802 | 0.3955 | 2.2942 | 1.3426 |
| $Z_{diff}$(G - N)$\mu$ | 1.5444 | 1.0258 | 1.5887 | 0.3980 |

## III. Structural Analyses and Their Interpretation

### *A. Primary Results*

The formal hypothesis with which this ensemble of mind/machine experiments was undertaken was that the prior PEAR database, as represented in Table 0 and Figure 4, would be statistically replicated in scale and character. From the summary Tables F.1, G.1, P.1, and C.1 and from the cumulative devi-

TABLE P.4
Feedback Effects in PEAR data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Digital (37 series, 12 operators) | | | | |
| $\Delta\mu$ | 0.018811 | -0.032811 | -0.024811 | 0.004000 |
| $\sigma$ | 7.0885 | 7.0232 | 7.0444 | 7.0338 |
| $Z$ | 0.5117 | -0.8926 | -0.6749 | 0.1539 |
| Graphic (195 series, 71 operators) | | | | |
| $\Delta\mu$ | 0.001908 | 0.012677 | 0.008015 | -0.002331 |
| $\sigma$ | 7.0582 | 7.0615 | 7.0667 | 7.0641 |
| $Z$ | 0.1191 | 0.7917 | 0.5006 | -0.2058 |
| None (18 series, 8 operators) | | | | |
| $\Delta\mu$ | -0.042444 | -0.002722 | 0.077333 | 0.040028 |
| $\sigma$ | 7.0442 | 7.1300 | 7.0509 | 7.0906 |
| $Z$ | -0.8053 | -0.0517 | 1.4673 | 1.0741 |
| Differences | | | | |
| $Z_{diff}$ (D - G)$\mu$ | 0.4216 | -1.1344 | -0.8187 | 0.2233 |
| $Z_{diff}$ (D - N)$\mu$ | 0.9533 | -0.4682 | -1.5896 | -0.7929 |
| $Z_{diff}$ (G - N)$\mu$ | 0.8052 | 0.2796 | -1.2584 | -1.0875 |

TABLE C.4
Feedback Effects in Concatenated Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Digital (94 series, 52 operators) | | | | |
| $\Delta\mu$ | 0.042915 | -0.029213 | 0.027234 | 0.028223 |
| $\sigma$ | 7.0659 | 7.0319 | 7.0654 | 7.0486 |
| $Z$ | 1.8607 | -1.2666 | 1.1808 | 1.7306 |
| Graphic (613 series, 220 operators) | | | | |
| $\Delta\mu$ | -0.000297 | -0.001057 | -0.004212 | -0.001577 |
| $\sigma$ | 7.0560 | 7.0562 | 7.0678 | 7.0620 |
| $Z$ | -0.0329 | -0.1170 | -0.4664 | -0.2470 |
| None (43 series, 26 operators) | | | | |
| $\Delta\mu$ | -0.072000 | -0.003953 | 0.037744 | 0.020849 |
| $\sigma$ | 7.0610 | 7.0987 | 7.0819 | 7.0903 |
| $Z$ | -2.1115 | -0.1159 | 1.1069 | 0.8647 |
| Differences | | | | |
| $Z_{diff}$ (D - G)$\mu$ | 1.7446 | -1.1368 | 1.2696 | 1.7015 |
| $Z_{diff}$ (D - N)$\mu$ | 2.7914 | -0.6136 | -0.2553 | 0.2533 |
| $Z_{diff}$ (G - N)$\mu$ | 2.0327 | 0.0821 | -1.1894 | -0.8991 |

ation graphs of Figures 1, 2, 3, and 5, we conclude that this hypothesis has not been confirmed. Although the agreed upon primary indicators of effect, the HI–LO ($\Delta$) mean shifts and their corresponding $Z$-scores, progress in the intended directions in all three laboratory results and in their cross-laboratory combinations, the effect size is essentially one order of magnitude smaller than for the prior data (.0034 versus .0208) and thus falls well below any credible statistical significance ($Z$ = 0.596 versus 3.809). Alternatively stated, if

TABLE F.5
Runlength Effects in FAMMI Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| 100-trial runs (198 series, 80 operators) | | | | |
| $\Delta\mu$ | -0.006313 | 0.001283 | 0.007687 | 0.003202 |
| $\sigma$ | 7.0515 | 7.0593 | 7.0659 | 7.0626 |
| $Z$ | -0.3973 | 0.0807 | 0.4837 | 0.2850 |
| 1,000-trial runs (52 series, 22 operators) | | | | |
| $\Delta\mu$ | 0.012942 | -0.036115 | 0.001192 | 0.018654 |
| $\sigma$ | 7.0682 | 7.0825 | 7.0921 | 7.0873 |
| $Z$ | 0.4174 | -1.1647 | 0.0385 | 0.8507 |
| Differences | | | | |
| $Z_{diff}$ (H - T)$\mu$ | -0.5527 | 1.0733 | 0.1864 | -0.6271 |

TABLE G.5
Run-Length Effects in GARP Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| 100-trial runs (173 series, 68 operators) | | | | |
| $\Delta\mu$ | 0.005590 | -0.035046 | -0.008850 | 0.013098 |
| $\sigma$ | 7.0566 | 7.0503 | 7.0736 | 7.0620 |
| $Z$ | 0.3288 | -2.0615 | -0.5206 | 1.0896 |
| 1,000-trial runs (77 series, 33 operators) | | | | |
| $\Delta\mu$ | 0.000805 | 0.037844 | -0.006351 | -0.022097 |
| $\sigma$ | 7.0545 | 7.0224 | 7.0661 | 7.0443 |
| $Z$ | 0.0316 | 1.4851 | -0.2492 | -1.2264 |
| Differences | | | | |
| $Z_{diff}$ (H - T)$\mu$ | 0.1562 | -2.3795 | -0.0816 | 1.6249 |

the prior PEAR results are used as the standard of replication, this prediction is refuted at a $Z = -2.87$ level.

Given the sophistication and scope of the experimental and analytical procedures followed in both these contemporary studies and in the prior PEAR work, and given the many examples of both "successful" and "unsuccessful" high-quality research performed elsewhere over the past several decades (Radin & Nelson, 1989), this stark failure to replicate reaffirms an enduring and ubiquitous "reproducibility problem" that has long characterized mind/machine interaction experiments of this class (Bierman & Houtkooper, 1981; Shapin & Coly, 1985). Some resolution of this replication paradox would seem to be essential to sustained progress in this field. To this purpose, various categorical possibilities need to be acknowledged and assessed:

1. Some physical or technical conditions, essential to generation of the anomalies, were not properly recognized and/or incorporated in the replication program. The primary and secondary parameters so far in-

TABLE P.5
Runlength Effects in PEAR Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| 100-trial runs (139 series, 49 operators) | | | | |
| $\Delta\mu$ | 0.009540 | -0.013626 | 0.001468 | 0.007547 |
| $\sigma$ | 7.0661 | 7.0690 | 7.0621 | 7.0655 |
| $Z$ | 0.5030 | -0.7184 | 0.0774 | 0.5627 |
| 1,000-trial runs (111 series, 42 operators) | | | | |
| $\Delta\mu$ | -0.009207 | 0.027955 | 0.016514 | -0.005721 |
| $\sigma$ | 7.0561 | 7.0504 | 7.0625 | 7.0565 |
| $Z$ | -0.4338 | 1.3172 | 0.7781 | -0.3812 |
| Differences | | | | |
| $Z_{diff}$ (H - T)$\mu$ | 0.6586 | -1.4609 | -0.5286 | 0.6592 |

TABLE C.5
Runlength Effects in Concatenated Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| 100-trial runs (510 series, 197 operators) | | | | |
| $\Delta\mu$ | 0.002045 | -0.015104 | 0.000382 | 0.007743 |
| $\sigma$ | 7.0572 | 7.0589 | 7.0675 | 7.0632 |
| $Z$ | 0.2065 | -1.5254 | 0.0386 | 1.1059 |
| 1000-Trial Runs (240 series, 97 operators) | | | | |
| $\Delta\mu$ | -0.001196 | 0.017246 | 0.005858 | -0.005694 |
| $\sigma$ | 7.0582 | 7.0484 | 7.0701 | 7.0593 |
| $Z$ | -0.0828 | 1.1948 | 0.4059 | -0.5579 |
| Differences | | | | |
| $Z_{diff}$ (H - T)$\mu$ | 0.1852 | -1.8482 | -0.3129 | 1.0856 |

vestigated are not crucial to these phenomena and thus yield marginal results contaminated by artifact and obscured by random flux.

2. Certain subjective psychological conditions, essential to generation of the anomalies, were not properly recognized and/or incorporated in the replication.
3. The statistical analyses and/or their theoretical foundations deployed to distinguish anomalous and normal behavior are inadequate for the task.
4. The basic assumptions underlying the conceptual framework within which these experiments were designed are incorrect or inadequate to encompass the phenomena involved.
5. The phenomena underlying the anomalies are intrinsically irreplicable and unpredictable, even on a statistical basis and even with all objective and subjective parameters closely controlled, and thus are inaccessible to definitive scientific study.

The last, most radical possibility surely should be deferred until all other op-

TABLE F.6
Series-Position *Z*-Scores in FAMMI Data

| Series no. | *N* | BL | LO | HI | Δ |
|---|---|---|---|---|---|
| 1 | 79 | -0.6335 | -1.5824 | -0.1781 | 0.9930 |
| 2 | 42 | -0.6873 | 0.2850 | 0.2091 | -0.0537 |
| 3 | 28 | 0.0651 | 0.1944 | 0.9221 | 0.5145 |
| 4 | 22 | -1.4731 | 0.7094 | -0.3814 | -0.7713 |
| 5+ | 79 | 1.5829 | 0.0674 | 0.4750 | 0.2882 |

TABLE G.6
Series-Position *Z*-Scores in GARP Data

| Series no. | *N* | BL | LO | HI | Δ |
|---|---|---|---|---|---|
| 1 | 66 | -0.5879 | -0.0517 | -1.2969 | -0.8805 |
| 2 | 42 | 1.1227 | -2.2662 | 2.1226 | 3.1034 |
| 3 | 34 | -0.0476 | -1.2264 | -1.9581 | -0.5174 |
| 4 | 24 | 0.1433 | 2.5579 | 0.1652 | -1.6918 |
| 5+ | 84 | 0.1225 | -0.4753 | -0.1796 | 0.2091 |

TABLE P.6
Series-Position *Z*-Scores in PEAR Data

| Series no. | *N* | BL | LO | HI | Δ |
|---|---|---|---|---|---|
| 1 | 66 | 1.0674 | -0.1349 | 0.8830 | 0.7197 |
| 2 | 23 | 1.7913 | 0.3870 | -0.2173 | -0.4273 |
| 3 | 14 | -0.1888 | 2.1980 | 1.3912 | -0.5705 |
| 4 | 13 | 0.0447 | -0.2642 | 0.2679 | 0.3763 |
| 5+ | 134 | -1.3267 | -0.2268 | -0.2758 | -0.0347 |

TABLE C.6
Series-Position *Z*-Scores in Concatenated Data

| Series no. | *N* | BL | LO | HI | Δ |
|---|---|---|---|---|---|
| 1 | 211 | -0.1195 | -1.0726 | -0.3405 | 0.5177 |
| 2 | 107 | 1.1033 | -1.0618 | 1.3601 | 1.7126 |
| 3 | 76 | -0.0097 | 0.2411 | -0.1529 | -0.2786 |
| 4 | 59 | -0.7872 | 1.9405 | -0.0017 | -1.3734 |
| 5+ | 297 | -0.0096 | -0.3703 | -0.0358 | 0.2365 |

tions are exhausted. Selection among the remaining categories may possibly be informed by the internal structure of the experimental databases, e.g., from the secondary parameter breakdowns of the previous section, the higher moments of the distributions, or the sequential correlations in the data streams. In the prior PEAR studies, such attention to structural details of the data distributions proved instructive in analysis and interpretation of the experimental

TABLE F.7
Experimenter Effects in FAMMI Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Experimenter number 1 (37 series, 8 operators) | | | | |
| $\Delta\mu$ | 0.025351 | 0.044757 | 0.001324 | - 0.021716 |
| σ | 7.0567 | 7.0840 | 7.0601 | 7.0721 |
| $Z$ | 0.6896 | 1.2175 | 0.0360 | - 0.8354 |
| Experimenter number 2 (109 series, 15 operators) | | | | |
| $\Delta\mu$ | - 0.010220 | - 0.023670 | 0.013651 | 0.018661 |
| σ | 7.0673 | 7.0583 | 7.0876 | 7.0729 |
| $Z$ | - 0.4772 | - 1.1052 | 0.6374 | 1.2322 |
| Experimenter number 3 (80 series, 50 operators) | | | | |
| $\Delta\mu$ | 0.015900 | - 0.008825 | - 0.001388 | 0.003719 |
| σ | 7.0316 | 7.0780 | 7.0534 | 7.0657 |
| $Z$ | 0.6360 | - 0.3530 | - 0.0555 | 0.2104 |
| Experimenter group 4 (24 series, 46 operators) | | | | |
| $\Delta\mu$ | - 0.069708 | 0.000250 | 0.006583 | 0.003167 |
| σ | 7.0741 | 7.0138 | 7.0751 | 7.0444 |
| $Z$ | - 1.5272 | 0.0055 | 0.1442 | 0.0981 |
| Chi-squared on Zs with 4 *df* (90% CE: 0.71–9.49) | | | | |
| $\chi^2$ | 3.4402 | 2.8283 | 0.4314 | 2.2701 |

Note: *df* = degrees of freedom.

TABLE G.7
Control Mode Effects in GARP Data

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Auto (193 series, 68 operators) | | | | |
| $\Delta\mu$ | 0.011927 | - 0.015824 | - 0.012176 | 0.001824 |
| σ | 7.0529 | 7.0363 | 7.0685 | 7.0524 |
| $Z$ | 0.7410 | - 0.9831 | - 0.7565 | 0.1602 |
| Manual (57 series, 25 operators) | | | | |
| $\Delta\mu$ | - 0.022333 | - 0.001667 | 0.005789 | 0.003728 |
| σ | 7.0661 | 7.0603 | 7.0808 | 7.0705 |
| $Z$ | - 0.7541 | - 0.0563 | 0.1955 | 0.1780 |
| Differences | | | | |
| $Z_{diff}$ (A - M)$\mu$ | 1.0164 | - 0.4200 | - 0.5330 | - 0.0799 |

databases, which in that case contained strong primary results. Indeed, most of the salient features of these prior results devolved from such structural assessments, and much of our admittedly tentative and incomplete understanding of the basic nature of the phenomena is based on them. It behooves us, therefore, to establish whether the contemporary replication database, despite its minimal primary yield, nonetheless also embodies internal structural aspects that depart significantly from chance expectation. If so, these could uncover some other form and degree of anomalous effect, or indicate flaws in the experimental design that reduced the overall yield.

TABLE G.8
Effects by GARP Operator Types

| Measure | BL | LO | HI | Δ |
|---|---|---|---|---|
| Formal operators (169 series, 41 operators) | | | | |
| $\Delta\mu$ | -0.005089 | -0.028266 | -0.002923 | 0.012672 |
| $\sigma$ | 7.0621 | 7.0330 | 7.0770 | 7.0550 |
| $Z$ | -0.2958 | -1.6433 | -0.1699 | 1.0418 |
| Staff operators (30 series, 6 operators) | | | | |
| $\Delta\mu$ | 0.053967 | 0.008167 | -0.043100 | -0.025633 |
| $\sigma$ | 7.0115 | 7.0383 | 7.0532 | 7.0457 |
| $Z$ | 1.3219 | 0.2000 | -1.0557 | -0.8880 |
| Student operators (41 series, 17 operators) | | | | |
| $\Delta\mu$ | -0.033732 | 0.044415 | -0.019000 | -0.031707 |
| $\sigma$ | 7.0448 | 7.0622 | 7.0741 | 7.0681 |
| $Z$ | -0.9659 | 1.2718 | -0.5441 | -1.2840 |
| Visitor operators (10 series, 5 operators) | | | | |
| $\Delta\mu$ | 0.165300 | -0.043800 | 0.054600 | 0.049200 |
| $\sigma$ | 7.1276 | 7.1166 | 7.0174 | 7.0670 |
| $Z$ | 2.3377 | -0.6194 | 0.7722 | 0.9840 |
| Chi-squared on Zs with 4 *df* (90% CE: 0.71–9.49) | | | | |
| $\chi^2$ | 8.2328 | 4.7418 | 2.0357 | 4.4910 |

*B. Structural Anomalies*

*1. Structural parameters.* The data tables presented in section II.B.3 summarize our attempt to collate the results of the three laboratories, individually and collectively, with various experimental parameters, in the hope that any significantly deviant subsets or disparities between alternative modalities might illuminate the most important objective or subjective correlates. Specifically studied, to varying degrees, have been the following structural cells:

Operator Parameters
- Gender: Male; Female; Multiple
- Types: Formal; Staff; Student; Visitor

Protocol Parameters
- Assignment of intention: Instructed; Volitional
- Feedback modalities: Digital; Graphic; None
- Machine control: Automatic; Manual
- Run lengths: 100 trials; 1000 trials

Sequential Effects
- Series-position

Experimenter Effects
- Individuals by code number

As already noted, a substantial number of suggestive disparities have indeed appeared in the data subsets. However, because of the number of cases exam-

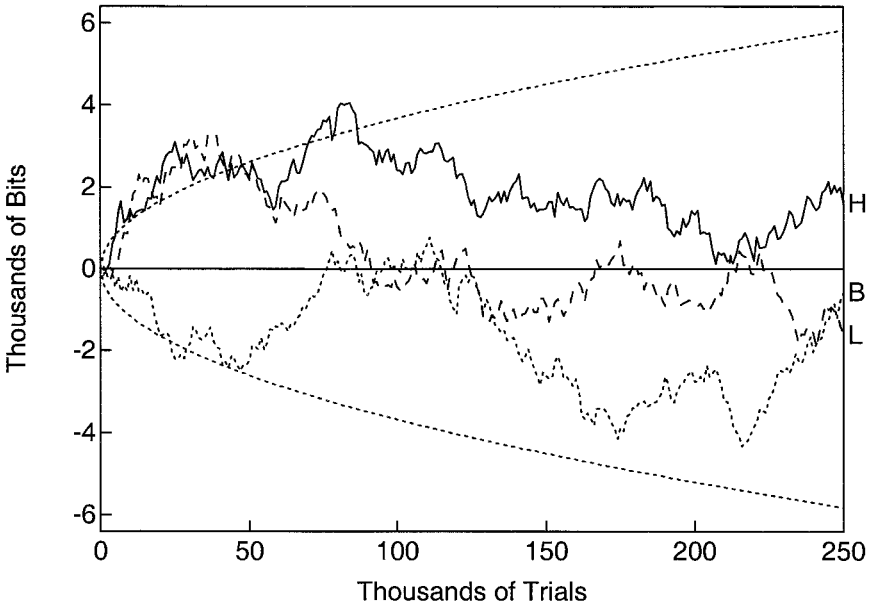


Fig. 1. FAMMI cumulative deviations.

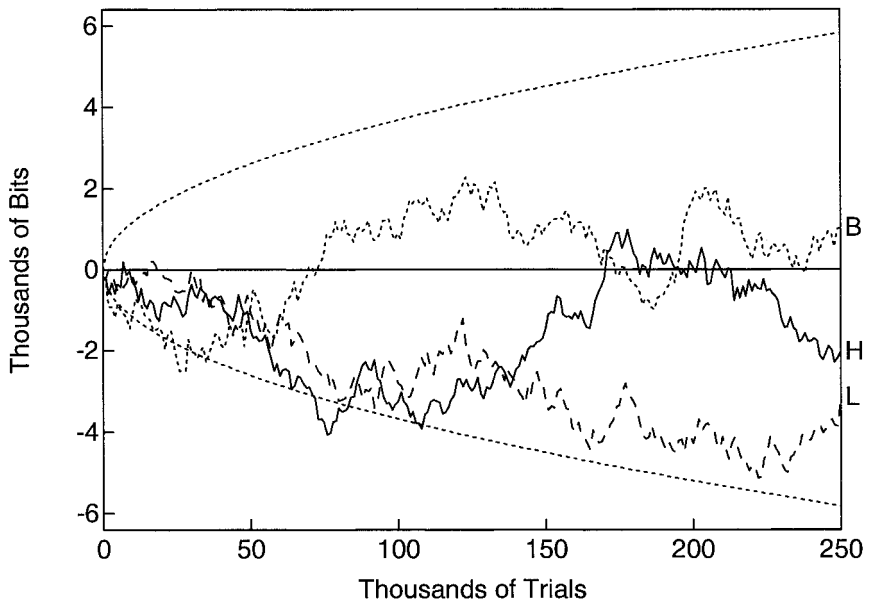


Fig. 2. GARP cumulative deviations.

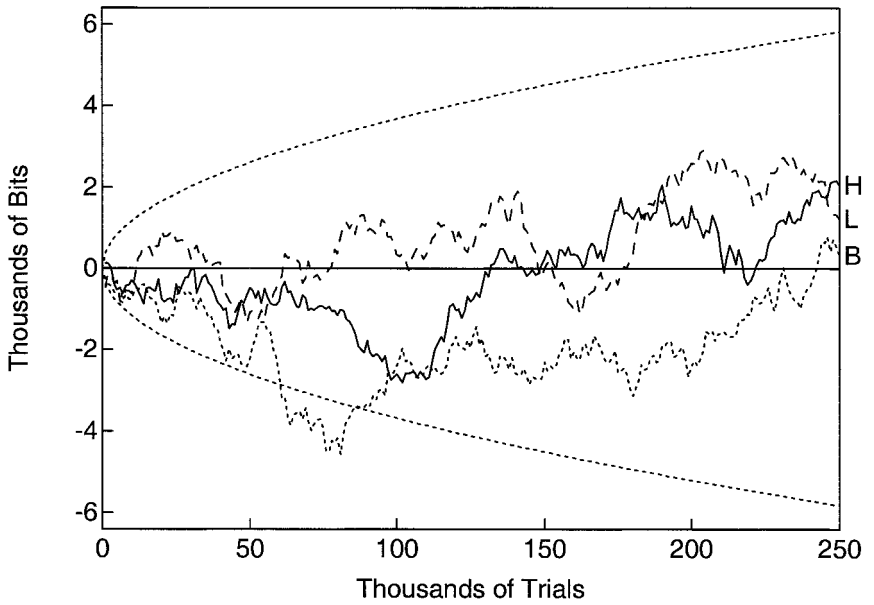


Fig. 3. PEAR Laboratory cumulative deviations.

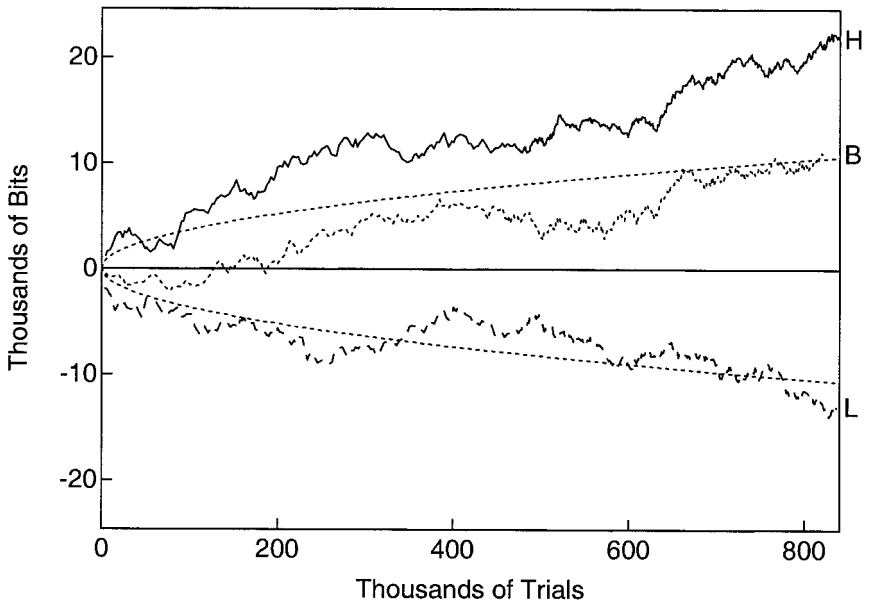


Fig. 4. Prior PEAR cumulative deviations.

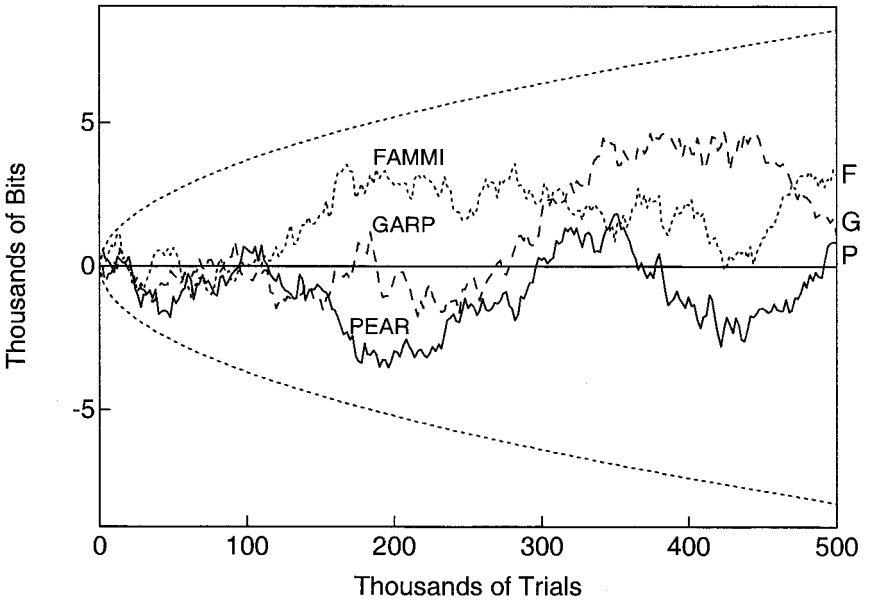


Fig. 5. Cumulative HI–LO differences for all three labs.

ined, some seemingly meaningful distinctions may appear by chance, so we cannot interpret the several large *Z*-scores in the structural tables until we have somehow corrected for the multiplicity of tests, to learn whether these are indeed larger or more numerous than would be expected by chance for the number of analyses that have been generated.

The discussion of sequential and experimenter effects will be deferred to a later section. For the moment, we will consider only the operator and protocol parameters, as they are broken down in Tables F.2 through F.5, G.2 through G.5, and P.2 through P.5. These tables report a total of 124 mean-shift *Z*-scores for the various intentional condition subsets. More importantly, 76 $Z_{diff}$ scores for differences between parameter conditions are presented. Since any structural anomalies in these parameters would appear as differences of performance between different parameter conditions, the 76 $Z_{diff}$ scores are obviously the crucial population to test. We may also check the population of mean-shift *Z*-scores, but this test is less central to the examination of structure, first because the statistical resolution is relatively weak since each *Z* involves only one half of a parameter comparison, and second because the absence of an overall intentional effect makes significant mean shifts in these full subsets much less likely.

We might naively suppose that we can perform the requisite multiple-tests correction simply by comparing the large population of $Z_{diff}$ scores to the theoretical *Z* distribution. For example, since the subset comparisons are not di-

rected, i.e., we do not have a prior hypothesis regarding the sign of any $Z_{diff}$, the presence of structure might be expected to inflate the absolute magnitude of some $Z_{diff}$ scores, and therefore the standard deviation of the $Z_{diff}$ score distribution. And, indeed, when we examine the standard deviations in these populations, we find that the 76 $Z_{diff}$ values have a standard deviation of 1.258, rather than the theoretically expected value of 1, a result unlikely with $p = 0.00098$.

At face value, this might seem strong evidence for structure in the $Z_{diff}$ population. The flaw in such a conclusion is that the analysis presupposes that the scores comprising the population are mutually independent, which they are not. To begin with, each score in the $\Delta$ column of the data tables is strongly correlated with the scores in the HI and LO columns. The breakdown in the feedback parameter, having as it does three levels, produces a set of three parameter differences, each strongly correlated with the other two. Worse, there are additional correlations between $Z$-scores in different parameter comparisons, because the populations are not in uniform proportion. For example, the fraction of instructed-assignment series generated by females is not necessarily the same as the fraction of volitional-assignment series generated by females, because of the freedom of operators to choose secondary parameters. When these proportions are not equal, $Z_{diff}$(I - V) will acquire an intrinsic correlation, positive or negative, with $Z_{diff}$ (F - M). Similar considerations apply among almost all of their parameter sets.

The presence of these correlations, of variable magnitude and sign between different $Z_{diff}$ scores, complicates the comparison with theory immensely, so much so that the attempt was abandoned. Instead, it was decided to determine the theoretical values of the population-summary parameters empirically through a Monte Carlo procedure, the details of which are given in the next section.

*2. Monte Carlo simulations.*

*(a) General treatment.* We wish to determine whether the populations of $Z$-scores, especially the population of 76 $Z_{diff}$ scores, emerging from Tables F.2 through F.5, G.2 through G.5, and P.2 through P.5, depart from the expected chance distribution for this array of tests when applied to random data. To determine this chance distribution, we employ a Monte Carlo procedure which in essence involves repeatedly performing the analysis on data that are guaranteed to be random.

The analysis programs that were used to process the empirical data for the above tables take, as input, the indicial information describing the parameters for each series, and the actual data generated in the series. For the Monte Carlo process, we submit to those programs exactly the same indicial information, along with ersatz data constructed with a numerical pseudorandom algorithm to match the null-hypothesis distribution for these experiments. The fact that we are using the indicial information from the actual experiments guarantees that we reproduce the correct correlation structure in the output $Z$ population.

(We use simulated data rather than simply reordering the actual data, because if structure does exist in the actual data, the statistics of the raw data must necessarily be distorted to some extent. Randomly reordering the raw data, as is often done in Monte Carlo applications, does not serve our purpose in the current case. A random reordering breaks the connection between the data and the indicial information but leaves intact—merely relocated—the shifted values that constitute the structural anomaly and therefore does not give a reliable measure of the null-hypothesis distribution.)

Thus, each iteration of the Monte Carlo process produces its own population of 76 $Z_{diff}$ scores. (It also produces a population of 124 mean-shift $Z$-scores, which are also analyzed and reported for the sake of completeness.) This process is then repeated a total of 5000 times to ensure that the distribution parameters are well estimated. Any measure—e.g., the standard deviation described above—that characterizes the population of $Z_{diff}$ scores produced by the actual data thus can be compared with 5000 samples from its null-hypothesis distribution produced by the Monte Carlo procedure.

Table M.1 presents the results of this comparison with the Monte Carlo populations for several such summary measures. Each of these measures is a slightly different quantification of the qualitative hypothesis that the population of $Z_{diff}$ scores in the actual data has larger absolute values than predicted under the null hypothesis. The measures presented are the standard deviation, discussed above; the largest absolute value of any $Z_{diff}$ in the population; and the number of $Z_{diff}$ scores in the population exceeding each of three thresholds. The "population" referred to here is always the population of 76 $Z_{diff}$ values (or in Table M.1a, 124 mean-shift $Z$-scores) produced by a single instance of the analysis, real or simulated (not the population of 5000 simulated instances).

The columns of Table M.1 present, first, the value of the named measure in the actual data; next, the mean and standard deviation of the named measure across the 5000 Monte Carlo iterations; and next, the number of Monte Carlo iterations where the value of this measure exceeds the value in the actual data. (The number in this column, when divided by 5000, is a form of empirical upper-tail $p$-value describing the position of the actual data in the Monte Carlo distribution.) A final column presents measure values obtained when the actual data are replaced, not by simulated data but by calibration data from the experimental apparatus. This is included as a precaution against the possibility that differences between real and simulated data might derive from properties of the physical data source, rather than from an experimental effect. The actual calibrations from Freiburg, Giessen, and Princeton were used to replace the experimental data for their respective laboratories, in this calculation.

From Table M.1, we note that, as expected, the population of 124 mean-shift $Z$-scores is indistinguishable from the null hypothesis distribution as constructed by the Monte Carlo process. The $Z_{diff}$ Table M.1b, however, is much more interesting. For example, the standard deviation of the $Z_{diff}$ population now yields a $p$-value of .014, quite different from the erroneous calculation

mentioned above, but clearly indicative of anomalous structure. Although conceptually the standard deviation increase is the primary indicator of a modified $Z_{diff}$ distribution, the other measures can provide additional information about the nature of the modification. However, the introduction of these different measures might suggest that the question of multiple analysis has appeared yet again, requiring some form of Bonferroni correction. This multiplicity is an unavoidable consequence of the initial exploratory decision to examine several specific ways in which the actual population of $Z_{diff}$ scores might depart from the null hypothesis prediction. It is possible, however, to render irrelevant all issues of multiple testing by calculating a single summary statistic encompassing all five measures presented in Table M.1.

As the table shows, each of the five measures has a mean and standard deviation determined from the Monte Carlo population. A normalized score can be calculated for each parameter relative to this distribution by subtracting the distribution mean from the observed value and dividing the difference by the standard deviation. (We do not call this normalized score a *Z*-score because some of the measures are not normally distributed.) The sum of these normalized scores is a single statistic that weights equally the departure from Monte Carlo norms in each of the five measures. This sum can be calculated not only for the actual data but also for each individual iteration of the Monte Carlo simulation. Comparing this combined-measures summary statistic in the real data with the distribution of values in the 5000 Monte Carlo iterations gives us a single, definitive *p*-value for the degree to which the real data stand out from the null hypothesis: There are 109 iterations that exceed the real data in the summary statistic, and 0 exact ties, leading to a *p*-value of .022. Since this is a single-test result requiring no correction, we may safely conclude that the population of $Z_{diff}$ scores in the PortREG database can be distinguished from the null hypothesis at a $p = .022$ level. Thus the apparent structural anomalies noted in Tables F.2 through F.5, G.2 through G.5, and P.2 through P.5 are, to this same level of confidence, real differences rather than statistical artifacts.

Figures 6, 7, and 7a represent these results in an instructive graphical form. Figure 6 shows the positions of the full subset empirical data *Z*-scores on the Monte Carlo calculated distributions. As expected, there is little departure from chance behavior here, save a slight positive shift of the largest *Z*-value. In Figure 7, however, substantial displacements of the empirical $Z_{diff}$ values with respect to the Monte Carlo background are clear by each of the five criteria, reaffirming the numerical values mentioned above. Figure 7a, shows similar major displacement of experimental value of the composite statistic just described, with respect to the Monte Carlo distribution.

While this analysis cannot guarantee that any *particular* subcells are aberrant, it can identify a hierarchy of such disparities that are most likely to represent legitimate structural anomalies. For example, Table M.2 lists the ten most prominent departures of the subcell difference *Z*-score from their corresponding Monte Carlo simulations, indexed by direction of intention and laboratory.

The secondary parameters are given in the order that makes the $Z_{diff}$ positive; thus, the first entry lists "V - I," denoting that the volitional data have a larger $\Delta$-effect than the instructed. From Table M.1b, we know that the number of $Z_{diffs}$ in the range above 2.0 to be affected by chance is about 3.5; hence, it is likely that some six or seven of the entries in Table M.2 correspond to real, nonrandom differences in operator achievements.

*(b) Most favorable cells.* While such Monte Carlo treatments provide no guarantees that any given one of these categories in fact entails anomalous results, they can provide guidelines for the most profitable cells to study more directly, leading to identification of the more important secondary parameters, and hence possibly to superior further experiments. As just one example, the data subset comprising all of the trials performed at the GARP laboratory using volitional assignment of direction of intention, nongraphic feedback, automatic machine control, and 100-trial runs shows a significant yield in the HI–LO separation of $\mu = 0.488 \pm 0.0241$ ($Z = 2.02$), whereas the subset of all data delineated by instructed assignment, graphic feedback, automatic control, and 1000-trial runs shows a strong negative yield of $\mu = -0.2308 \pm 0.0913$ ($Z = -2.53$). The source of this disparity may be further localized by noting that the combination of all GARP instructed, graphic subsets yields $\mu = 0.1010 \pm 0.0323$ ($Z = -3.13$), suggesting that the subjective parameters of volitional/instructed assignment and graphic/nongraphic feedback were particularly pertinent to GARP operator performance. Such observations then prompt examination of the corresponding subsets in the FAMMI and PEAR databases to see if such effects appear in these venues, as well.

To facilitate such interlaboratory cell comparisons, it is necessary to devise a standard procedure for dividing all of the PortREG databases into commen-

TABLE M.1
Comparison of All Laboratory Data with 5000 Monte Carlo Simulations

| Measure | Data | 5000 Monte Carlos | No. M. C. > data | Calib. data |
|---|---|---|---|---|
| (a) Distributions of 124 mean-shift *Z*-scores | | | | |
| SD of *Z* | 0.961 | 0.980 ± 0.129 | 2659 | 0.888 |
| Largest \|*Z*\| | 2.941 | 2.691 ± 0.437 | 1289 | 2.705 |
| No. (of 124): \|*Z*\| > 1.5 | 16 | 16.702 ± 6.932 | 2518.5[a] | 13 |
| No. (of 124): \|*Z*\| > 2.0 | 5 | 5.725 ± 4.037 | 2443[a] | 5 |
| No. (of 124): \|*Z*\| > 2.5 | 1 | 1.572 ± 1.950 | 2532[a] | 1 |
| (b) Distributions of 76 $Z_{diff}$ scores | | | | |
| SD of $Z_{diff}$ | 1.258 | 0.995 ± 0.114 | 68 | 0.937 |
| Largest $\lvert Z_{diff} \rvert$ | 3.184 | 2.597 ± 0.432 | 452 | 2.901 |
| No. (of 76): $\lvert Z_{diff} \rvert > 1.5$ | 19 | 10.206 ± 3.834 | 91.5[a] | 7 |
| No. (of 76): $\lvert Z_{diff} \rvert > 2.0$ | 10 | 3.540 ± 2.299 | 49.5[a] | 4 |
| No. (of 76): $\lvert Z_{diff} \rvert > 2.5$ | 2 | 1.003 ± 1.189 | 961[a] | 1 |

[a] Since these parameters are discrete, an exact match can occur between the value in the actual data and the value in a Monte Carlo iteration. Therefore, the number reported here is the number of Monte Carlo values strictly greater than the data plus one half the number of exact matches; this is a standard approach to calculating tail populations with discrete data.

surate subsets that control for various possible confounds. As already noted, many of the subset parameters are mutually confounded due to unequal subset sizes. For example, the GARP data appear to show differences between intentional assignment modes and also between feedback modes. Since the proportions of a given assignment mode are not guaranteed to be the same in all feedback modes, when we only dissect the data according to one parameter at a time we cannot know whether (a) a real difference between assignment modes drives an apparent difference between feedback types, (b) a real difference between feedback types drives an apparent difference between assignment modes, (c) both parameters are independently important, or (d) both parameters are *interdependently* important; i.e., that the difference in performance might not be associated with either parameter in isolation but only appears when they jointly take on appropriate values. To distinguish these cases, we need to decompose the data according to several parameters at once, creating "cells" that are consistent according to several secondary parameters. This has two benefits. First, we can distinguish among cases (a) through (c), by making unconfounded tests for each parameter. Second, we can identify case (d) if the differences between cells contain information not explicable in terms of the unconfounded effects of isolated parameters.

Ideally, one should break down the data according to *all* secondary parameters. Unfortunately, there are so many of these that to make such a complete subdivision would result in very small data subsets with correspondingly poor statistical resolution. Moreover, there is a significant risk that some cells in such a complete breakdown would be entirely empty, appreciably complicating the interpretation. As a balance between rigor and practicality, the following compromises are made:

1. Only "optional" parameters subject to operator choice are considered. Gender, fixed for each operator, is ignored. Series position, also not optional, and in any case showing hard-to-interpret variations, also is ignored.
2. Only parameters for which all three laboratories examined the parameter are considered. This reduces the selection to assignment mode, run length, and feedback.
3. Because each laboratory has a huge majority of its data in the graphic feedback condition, the other two modes are collapsed into a single "nongraphic" feedback category.

The result of these compromises is the eight-cell ($2 \times 2 \times 2$) breakdown used in Table C.7 and Figures 8–11. In these, a three-letter code is used to indicate the parameter values: the first letter, I or V, refers to instructed or volitional assignment; the second, G or N, to graphic or non-graphic feedback; the third, H or T, to 100-trial or 1000-trial runs. The values plotted on the figures are absolute mean shifts *in direction of intention* for HI, LO, BL, and $\Delta$. The $Z$-scores tabulated in Table C.7 are based only on the $\Delta$-effect, but all four intentional conditions are plotted in the figures.

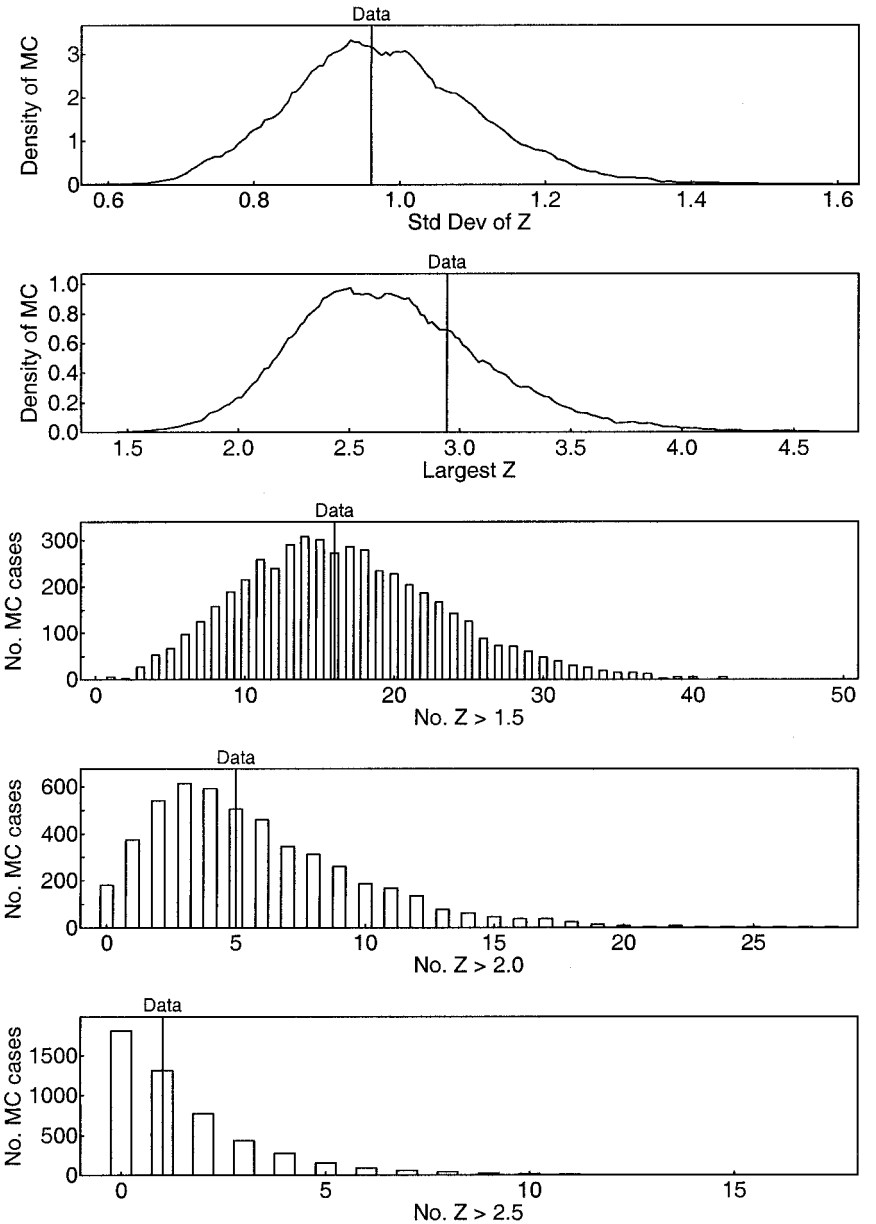


Fig. 6. Mean-shift *Z*-scores vs. Monte Carlo populations.

Returning to our particular example, the comparisons of performance under the volitional, nongraphic, 100-trial protocol (VNH), and the instructed, graphic, 1000-trial protocol (IGT), are seen to be particularly inconsistent across the three laboratories. This has encouraged further, ad hoc experimen-

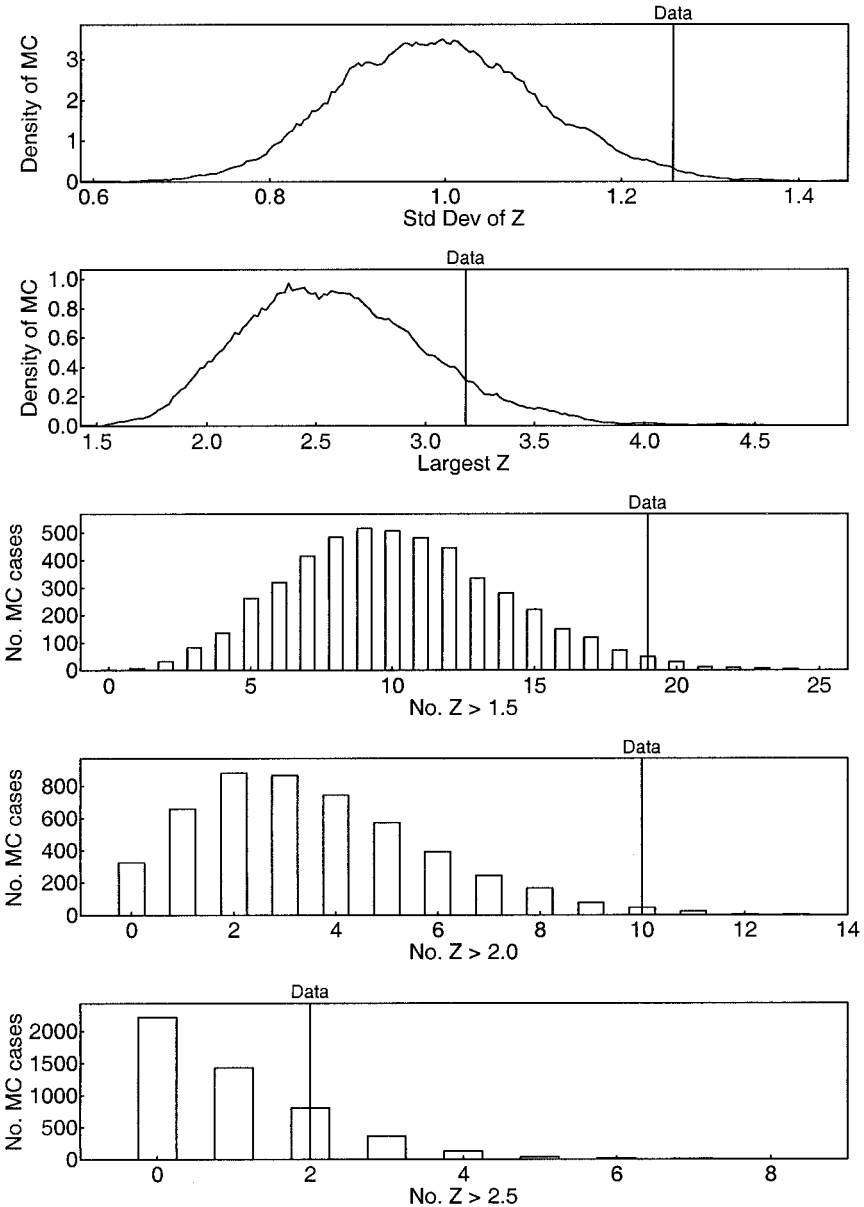


Fig. 7. Difference Z-scores vs. Monte Carlo populations.

tation, which is now in progress, and has prompted some new initiatives in theoretical modeling, which cannot be detailed here.

Similar structural exercises can be attempted in terms of other discriminators suggested by the Monte Carlo "most prominent" list above, such as opera-

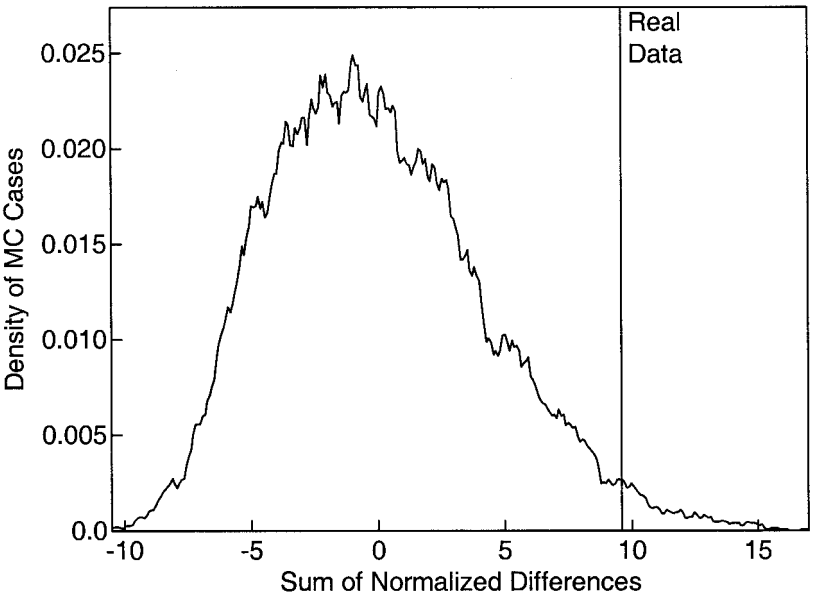


Fig. 7a. Composite statistic for difference *Z* vs. Monte Carlo.

tor gender, or single vs. multiple operators, both of which revealed striking disparities in the prior PEAR studies. In the replication studies, however, these effects are not so clearly evident. With reference to Tables F.2, G.2, and P.2, the only suggestive disparities appear in the PEAR data alone, and here most prominently in the single- vs. multiple-operator comparison, which was not explored by the other laboratories. Nonetheless, Table C.7 also presents a set of rudimentary correlation coefficients that indicate a much closer correspondence of the cell-by-cell result patterns between GARP and PEAR than between FAMMI and either other laboratory.

Obviously, it would be most desirable if it were possible by some means to extract from these structural cell results a completely unconfounded set of correlations with individual secondary parameters. Some form of analysis of variance (ANOVA) suggests itself, and indeed such has been employed twice in analyzing the prior PEAR data (Nelson et al., 1991, 2000), but even with the much higher overall yield of that database, the insights gained thereby did not vastly exceed those acquired from more directed ad hoc analyses. Nonetheless, once one has the cell scores, it is straightforward, although tedious, to construct the unconfounded secondary parameter effects. For example, to assess the effect of assignment mode, one must first compare the four pairs of cells that differ only in this parameter, i.e., IGH vs. VGH, IGT vs. VGT, INH vs. VNH, and INT vs. VNT. Each of these comparisons can be reduced to a difference *Z*-score using the formula at the end of section II.1. The four *Z*-scores

TABLE M.2
Most Prominent Z-Score Differences from Monte Carlo Comparisons

| Parameter | Intention | Lab | $Z_{diff}$ |
|---|---|---|---|
| ASG V–I | Δ | GARP | 3.184 |
| GEND M–F | BL | FAMMI | 2.764 |
| RUNL T–H | LO | GARP | 2.380 |
| FDB D–N | HI | GARP | 2.294 |
| ASG I–V | LO | GARP | 2.284 |
| FDB D–N | BL | GARP | 2.280 |
| ASG V–I | HI | GARP | 2.219 |
| FDB D–G | HI | FAMMI | 2.083 |
| MULT 2–1 | Δ | PEAR | 2.052 |
| FDB D–G | Δ | GARP | 2.035 |

so produced then can be combined into a single $Z$ giving the overall effect of that parameter, according to the composition rule:

$$Z_c = \left( \sum_{i=1}^{N} Z_i \sqrt{n_i} \right) \Big/ \sqrt{\sum_{i=1}^{N} n_i} \qquad (2)$$

where $Z_c$ denotes the composite $Z$ for a set of scores, $Z_i$, all measuring the same effect on databases of sizes $n_i$, $i = 1,\ldots,N$. In this manner it is possible to extract unconfounded correlations with certain specific secondary parameters, with the results displayed in Table C.8.

Particular further examples could be cited, but the broader point at issue is that the combination of the Monte Carlo simulations of the cellular data subsets with subsequent specific analyses of the most suggestive cells may help to localize the most pertinent objective and subjective parameters, and to refine future experiments to optimize these factors. We feel that only through such a detailed and disciplined process, tedious as it may be, is there hope for more effective and replicable experimentation, leading to better understanding of the phenomena.

*3. Series-position effects.* One possible structural indicator not explicitly explored in the Monte Carlo comparisons but readily accessed within the various laboratory databases, commonly termed "series-position effects," relates to the evolution of operator performance as a function of the number of experimental series performed. The prior PEAR data displayed a remarkably ubiquitous and consistent trend for scores to be highest for the first series attempted, then to deteriorate for the next two series, then to return to higher performance on the fourth, fifth, and subsequent series (Dunne et al., 1994). With reference to Tables F.6, G.6, and P.6, some such serial oscillations of performance are apparent, particularly in the GARP and PEAR data, but these are far from consistent across the three laboratories. Nonetheless, the composite data (Table C.6) also show some series-position pattern, but quite different from that of the prior PEAR results.

As a supplementary indicator, standard $\chi^2$ tests applied to these patterns, computed relative to chance expectation and relative to their respective empirical mean values, are displayed in Table C.9, along with their corresponding probabilities of chance occurrence. The last line presents the same analysis of the prior PEAR data. Clearly, only the GARP data exhibit a credible series-position pattern, albeit quite different in form than the prior PEAR results. Namely, the highest scoring in that replication is occurring in the second series, rather than in the first, and the lowest scoring in the fourth, rather than the third. In other words, the series pattern has shifted by one series.

*4. Operator-specific features.* Another structural anomaly identified in the prior PEAR data was the persistence of individual operator accomplishment features or "signatures," apparent over several series of effort, or over entire databases. Since few of the operators involved in the replication studies produced sufficient data for us to pursue this tendency solely in that context, we have modified the question to query whether those five operators who have appreciable databases in both the prior PEAR experiments and the replication study show similarities of performance between the two applications. For each of these operators, we calculate a *Z*-score for the difference in their HI–LO performance between the old and new experiments, using the $Z_{diff}$ formula in Equation 1. We use the same formula to calculate differences between their performances in the three individual intentions, HI, LO, and BL. The sum of the squares of those $Z_{diffs}$ becomes, for each operator, a $\chi^2$ with 3 *df* measuring the overall change in performance across all three intentions between the original experiment and the replication. The results, along with the associated chance probabilities, are presented in Table P.7. Two potentially instructive features are apparent. On the one hand, the first four operators, both individu-

TABLE C.7
*Z*-Scores in Secondary Parameter Cells, by Laboratory

| Parameter[a] | FAMMI | GARP | PEAR | All 3 labs |
|---|---|---|---|---|
| IGH | -0.1786 | -2.2266 | -0.0242 | -0.9242 |
| IGT | 0.9816 | -2.2378 | -0.3176 | -0.5417 |
| INH | 1.0979 | -0.7526 | -0.9888 | -0.3014 |
| INT | 0.5841 | 1.0941 | 0.4109 | 0.9127 |
| VGH | 0.0521 | 0.9335 | 0.8102 | 0.8740 |
| VGT | 0.2443 | -0.4092 | -0.7436 | -0.6242 |
| VNH | 0.4738 | 2.0454 | 1.3287 | 2.4720 |
| VNT | -0.4919 | -0.5920 | 0.6725 | -0.1659 |

Correlation coefficients of these response patterns

| | FAMMI–GARP | FAMMI–PEAR | GARP–PEAR |
|---|---|---|---|
| $\rho$ | -0.0061 | -0.4500 | 0.6501 |
| $Z(\rho)$ | -0.0143 | -1.1188 | 1.7452 |

Note: $\rho$ = correlation coefficient; $\rho = 1$ = perfect correlation; $\rho = -1$ = perfect anticorrelation; $Z(\rho)$ = standard normal deviate corresponding to value of $\rho$.

[a] I = instructed protocol; V = volitional protocol; G = graphic feedback; N = no feedback; H = 100-trial runs; T = 1000-trial runs.

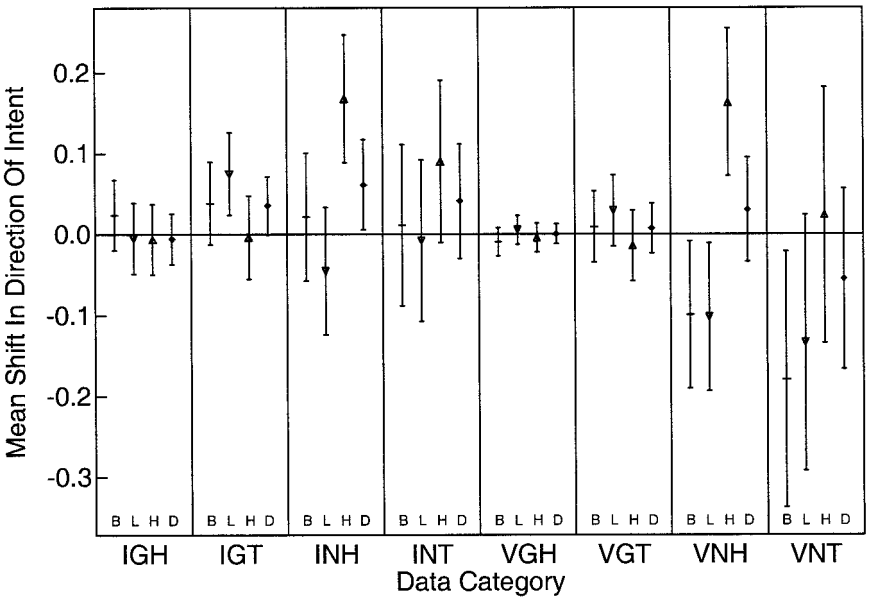


Fig. 8. FAMMI group data split by assignment (I,V), feedback (G,N), and run length (H,T).

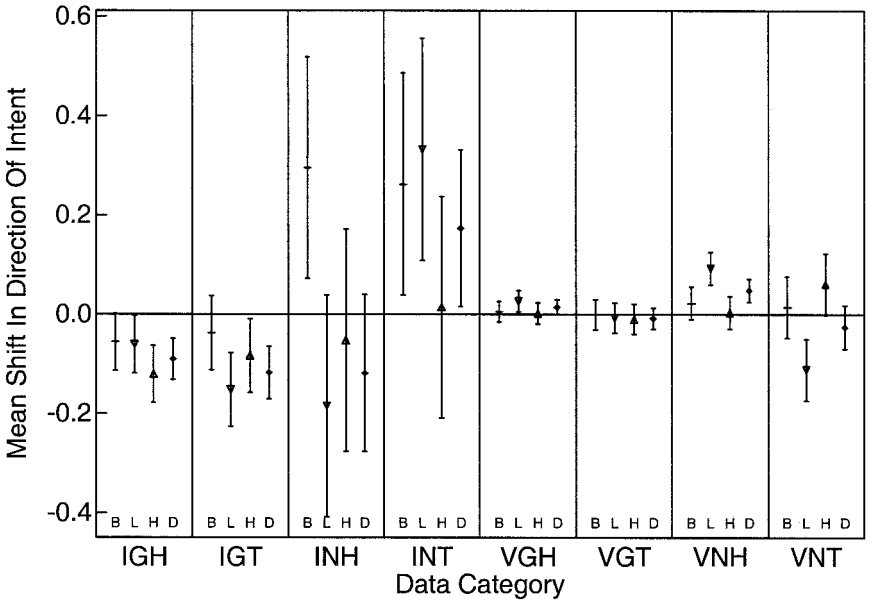


Fig. 9. GARP data split by assignment (I,V), feedback (G,N), and run length (H,T).

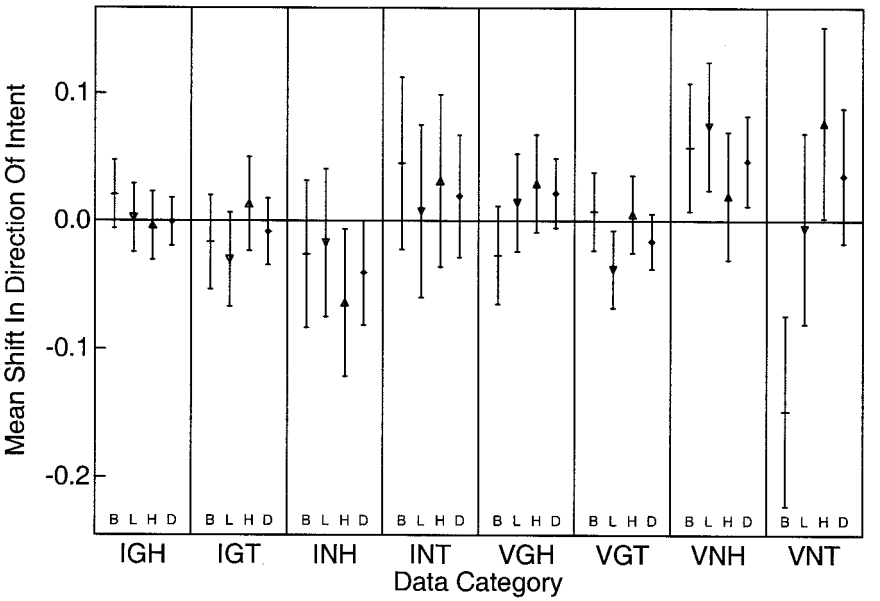


Fig. 10. PEAR data split by assignment (I,V), feedback (G,N), and run length (H,T).

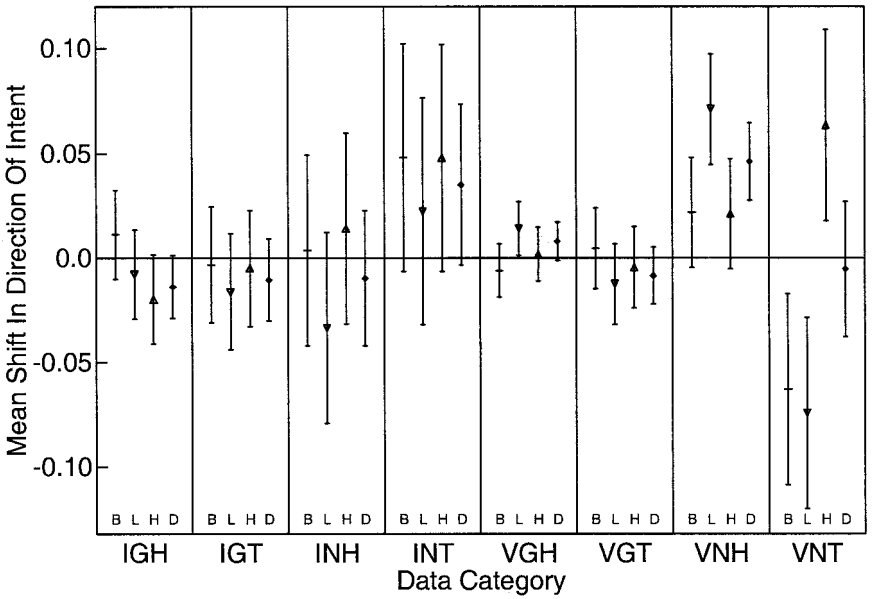


Fig. 11. All data split by assignment (I,V), feedback (G,N), and run length (H,T).

ally and collectively, performed remarkably similarly on the two experiments. On the other hand, Operator E displays a stark difference in performance between the prior PEAR and replication efforts that is virtually an inversion or "antireplication" of the prior "signature." (It may be worth noting that this operator repeatedly expressed strong resistance to being asked to validate a prior achievement through replication.) Clearly, these contradictory results cannot be resolved further without considerably more operator-specific data, but the subjective issue raised could ultimately prove important.

Other aspects of operator-specific structural anomalies have also been explored by similar $\chi^2$ techniques. For example, the possibility that a mixture of strong performances in the intended directions and in the directions opposite to intentions among the individual operators may cancel one another in the overall yield, thus obscuring the operator-level effects in the database, can be checked by a $\chi^2$ calculation encompassing all operators at the three laboratories, under all intentions. Specifically, by squaring the individual operator *Z*-scores (thus obtaining a sign-independent quantity) and adding these across all operators, we construct a $\chi^2$ with degrees of freedom equal to the number of operators. Table C.10 presents such results for the three laboratories along with their associated chance probabilities (in parentheses). Because the prior PEAR experiments indicated a gender difference in the tendency toward idiosyncratic performance, the databases are subdivided by gender as well as by laboratory.

Since the $\chi^2$ tests on the three individual intentions are mutually independent, they can be collected in a combined value indicative of the overall departure from chance behavior in all three intentions (last column). No elevated values that would suggest idiosyncratic operator performance appear. To the contrary, the PEAR female operators show a strikingly depressed $\chi^2$, especially in the LO intention, that compounds to an extraordinarily diminished value across all three intentions (39.33 on 66 *df;* $p = .996$). Considered as an improbably *small* $\chi^2$, this corresponds to $p = .004$, which we must immediately correct to .008 since we are willing to consider both unusually large and unusually small $\chi^2$. Bonferroni adjustment of this value for the seven independent subsets (two genders each at GARP and FAMMI, three at PEAR) still leaves a suggestive $p = .051$. Thus, there are moderate grounds for suspecting that this particular operator population is somehow producing performances that cluster too tightly about zero yield. Such calculations have been repeated on a series-by-series basis. Again, only the PEAR females show significant anomalies that survive the multiple-testing adjustments. It also may be worth noting that the data collected on series-position effects (Tables F.6, G.6, P.6, and C.9) and on operator-specific features (Tables P.7 and C.10) show a polyglot nature of above-chance occurrences similar to those covered in the Monte Carlo treatment (Table C.7).

*5. Standard deviations.* A different form of structural irregularity that may have indicative value can be detected in the individual laboratory and compos-

TABLE C.8
Difference Z-Scores of Unconfounded Secondary Parameters

| Parameter test | FAMMI | GARP | PEAR | All 3 labs |
|---|---|---|---|---|
| Assignment (I–V) | 0.4224 | -2.9204 | -0.9689 | -1.2611 |
| Feedback (G–N) | -0.7343 | -1.0871 | -0.4485 | -1.7823 |
| Runlength (H–T) | -0.4275 | 1.3792 | 0.5599 | 0.9331 |

ite databases. Even cursory examination of the tables of section II.B reveals many instances where the trial-level standard deviations are less than the theoretical value of 7.071. This, of course, might be an artifactual result of a flaw in the random noise sources, so these standard deviation figures should be compared not with the theoretical value, but with an empirical value derived from the concurrent calibrations of the instruments (cf. Appendix I). Since the three calibration datasets have consistent means and standard deviations, a pooled estimate of the latter may be constructed, yielding $\sigma = 7.0710$ with an empirical uncertainty of ±0.0028. Table C.11 reports *Z*-scores for the difference between the trial-level standard deviations of the active experimental data and this calibration estimate.

This method of comparison to an empirical standard technically makes them Student's *t*-scores rather than *Z*-scores. However, since there are well over 10,000 degrees of freedom in even the smallest datasets examined, the difference between the *Z* and *t* distributions safely may be neglected. By either standard, we find a statistically robust difference between the active experimental data and the calibration data in the composite across all three laboratories that is driven by substantial depressions in the LO and BL conditions. The prior PEAR finding of significantly higher experimental standard deviations for female operators compared to males (Dunne, 1998) is not sustained in magnitude by the replication data, although virtually all of the individual laboratory results show slight separations in this direction.

*6. Counts of successful operators and series.* In addition to the trial-score distribution criteria on which all of the preceding tabulations and discussions are predicated, the data also have been examined in terms of the fraction of experimental series and the fraction of operators, whose results conform to any extent with the direction of intention. Although those perspectives had proven

TABLE C.9
$\chi^2$ Tests for Series-Position *Z*-Scores

| Laboratory | $\chi^2$ (vs. theory); 5 *df*[a] | *p* of $\chi^2$ | $\chi^2$ (vs. empirical mean); 4 *df* | *p* of $\chi^2$ |
|---|---|---|---|---|
| FAMMI | 1.9316 | .859 | 1.7431 | .783 |
| GARP | 13.5799 | .019 | 13.5699 | .009 |
| PEAR | 1.1688 | .948 | 1.1680 | .883 |
| All 3 labs | 5.2207 | .390 | 5.0880 | .278 |
| Prior PEAR | 27.3385 | .00005 | 18.2453 | .001 |

instructive in some of the prior work, they clearly are not independent of the mean-shift values and in this replication study have added little new insight. Nonetheless, full tabulations of these quantities are available on request.

## IV. Summary Comments

As described in the introductory section, this coordinated replication study was the first collaborative research project attempted by the Freiburg, Giessen, and Princeton laboratories, as much to test the viability of the consortium concept, structure, management, and operations strategy as to create a major new database in mind/machine anomalies. By the former criterion, the project has been undeniably successful in that methods for provision of common experimental equipment, acquisition and reduction of experimental data, and analysis and interpretation of results have been well established and are available for deployment in subsequent research endeavors. Visitation and exchange of personnel among the laboratories at both the staff and management levels occur frequently, and the electronic communication channels that enable sharing of data and ideas function on a regular basis. In short, this first project has demonstrated that this ambitious consortium can function productively on such collaborative research enterprises.

As far as the replication results themselves are concerned, we are left with an empirical paradox. Whereas the prior PEAR experiments clearly displayed anomalous secular trends in REG output distribution means in correlation with operator intention, the three-laboratory replications, which employed essentially similar equipment and protocols, failed by an order of magnitude to replicate the primary correlations. Yet, these replication studies presented instead a substantial pattern of structural anomalies related to various secondary parameters, to a degree well beyond chance expectation and totally absent from the calibration data. To borrow a fluid mechanical metaphor, it is as if the influence of operator intention now was manifesting itself as a structural "turbulence" in the output data of the replication, rather than in a more orderly displacement of the data streams as was found in the prior PEAR studies.

With the various ad hoc examinations of these structural details described in sections II and III in hand, our search for some understanding of this substantial change in the character of the anomalous responses of the machines to operator

TABLE P.7
Consistency of Operators Between Prior PEAR and Replication Experiments

| Operator | $\chi^2$ ($p$) | $Z$ ($p$ [2-tail]) |
|---|---|---|
| A | 1.564 (.67) | 1.049 (.29) |
| B | 0.200 (.98) | 0.125 (.90) |
| C | 0.934 (.82) | -0.868 (.39) |
| D | 0.460 (.92) | 0.158 (.87) |
| E | 14.035 (.003) | 3.255 (.001) |

intention may be aided by systematic reconsideration of certain explicit and implicit assumptions with which the replication studies were undertaken:

1. ***Source independence**: The anomalous effects would manifest in the same form and scale on the PortREG sources as they had on the original PEAR benchmark machine.*

The prior PEAR data reported in Table 0 had been generated using a far more expensive and complex REG device that was replete with an array of failsafe controls, interior checkpoints, and other protections against short- and long-term deviations from strictly random behavior, that would unequivocally guarantee the integrity of the experimental results. The shift to the much simpler, less expensive, and more portable PortREG equipment seemed justified on the basis of its earlier successful deployments in other PEAR-based experiments, most notably our FieldREG studies (Nelson et al., 1996, 1998), and an extensive body of past evidence that comparable anomalous results could be obtained utilizing categorically different random physical sources (Jahn et al., 1997; Schmidt & Pantas, 1972). Yet, since that time certain other applications of PortREG equipment also have failed to produce results comparable with the prior benchmark findings, raising some questions about its consistency of sensitivity to operator intention (Jahn et al., 2000).

It has been suggested by one of PEAR's long-term operators that this reduction in effect may not be attributable to physical differences in the noise sources, per se, but to the shift of the REG unit from its original central focus in the experimental configuration to one where it appears to play only a peripheral supporting role to the computer that now dominates the operator's attention. Specifically, in the prior PEAR experiments digital feedback was presented as an LED display on the face of the REG device itself, with the

TABLE C.10
Operator Performance $\chi^2$ Values (with Associated Probabilities)

| Dataset | *df*[a] | BL | LO | HI | Δ | Combined[a] |
|---|---|---|---|---|---|---|
| FAMMI | | | | | | |
| Female | 40 | 49.05(.15) | 35.62(.67) | 46.43(.22) | 32.95(.78) | 131.09(.23) |
| Male | 40 | 32.75(.79) | 45.56(.25) | 32.55(.79) | 34.26(.73) | 110.86(.71) |
| All | 80 | 81.80(.42) | 81.18(.44) | 78.98(.51) | 67.21(.85) | 241.96(.45) |
| GARP | | | | | | |
| Female | 34 | 34.16(.46) | 31.82(.57) | 31.34(.60) | 28.30(.74) | 97.33(.61) |
| Male | 35 | 43.45(.15) | 34.80(.48) | 38.71(.31) | 40.37(.25) | 116.96(.20) |
| All | 69 | 77.61(.22) | 66.62(.56) | 70.05(.44) | 68.67(.49) | 214.28(.35) |
| PEAR | | | | | | |
| Female | 22 | 12.50(.95) | 7.55(.998) | 19.28(.63) | 13.07(.93) | 39.33(.996) |
| Male | 36 | 33.74(.58) | 41.42(.25) | 32.07(.66) | 40.53(.28) | 107.23(.50) |
| Co-operator | 20 | 19.85(.47) | 23.54(.26) | 12.42(.90) | 18.38(.56) | 55.81(.63) |
| All | 78 | 66.09(.83) | 72.51(.65) | 63.77(.88) | 71.98(.67) | 202.37(.93) |

[a] The degrees of freedom for the "Combined" column,which sums up the mutually independent contributions of BL, LO, and HI, are triple the number listed in the "*df*" column.

computer playing a more passive data-recording role, and the redundant archival data hardcopy was produced contemporaneously with the generation of the experimental data, rather than in a deferred printout. In the PortREG experiments, however, the noise source is housed in a small, unobtrusive gray box that is a far less evident component of the experimental system. Operator feedback, both digital and graphic, is produced on a computer display, rather than on the noise unit itself, and data printout is under computer control on a separate printer facility that operates only at the end of the run. Thus, the subjective experience of an operator generating data differs appreciably between the two experiments, so that while it is possible that the PortREG devices are still inherently sensitive to operator intention, their less prominent role in the experimental configuration may compromise their patterns of response. Another operator has suggested that the vast proliferation of interactive, visually engaging computer displays into public and personal applications over the past decade may have eroded much of the novelty of this format of human/machine interaction, rendering the experimental task less challenging and enjoyable. In either case, the role of feedback, rather than the noise source itself, may be the more pertinent concern, as further discussed in items 3 and 4 below.

2. ***Operator pool equivalence****: The overall performance of the pool of operators performing the replication experiments would be similar to that of the pool of operators that produced the prior PEAR results.*

This presumption seemed soundly based on extensive earlier results that these anomalous effects invariably appeared as broadly distributed, marginal shifts over the full operator population, rather than being dominated by a few exceptional operators (Jahn et al., 1997). The fact that PEAR, continuing its policy of using only uncompensated, anonymous volunteers, many of whom had participated in the prior experiments, achieved no better replication than

TABLE C.11
*Z*-Scores for Trial-Level Standard Deviations, by Laboratory and Gender

| Data | BL | LO | HI | Δ |
|---|---|---|---|---|
| All FAMMI | -1.5449 | -0.6599 | 0.0309 | -0.4301 |
| Male | -1.2757 | -0.7104 | -0.2837 | -0.6886 |
| Female | -0.9388 | -0.1895 | 0.4019 | 0.1466 |
| All GARP | -1.4539 | -2.8142 | 0.0257 | -1.9009 |
| Male | -1.4974 | -2.0454 | 0.4544 | -1.1032 |
| Female | -0.6035 | -2.0089 | -0.4123 | -1.6786 |
| All PEAR | -0.8997 | -0.9880 | -0.8451 | -1.2515 |
| Male | -0.8280 | -0.9173 | -1.2155 | -1.4791 |
| Female | -0.1623 | -0.2035 | 0.1221 | -0.0572 |
| Co-Op | -0.5499 | -0.5625 | -0.1638 | -0.5111 |
| Composite | -2.1027 | -2.4051 | -0.4257 | -1.8329 |

Note: Z-scores calculated from normal approximation to the distribution of standard deviations, which is accurate for these large datasets.

GARP or FAMMI, who followed more structured handling of operators, continues to suggest that the composition of the operator pool, per se, is not likely to be a major factor. Yet, some of the structural evidence from this present study, as discussed in items 4 and 7, may indicate otherwise.

3. ***Insensitivity to secondary parameters****: The overall results would be insensitive to minor alterations in the secondary experimental parameters.*

The prior PEAR data generated with digital feedback or no feedback were statistically indistinguishable from the graphic-feedback data, leading to the assumption that feedback was a matter of indifference or at most of individual operator aesthetic preference. Both of the ANOVA studies of the prior PEAR data also failed to uncover any overall feedback sensitivity. Yet, the differences in replication results related to this parameter indicate that it may have been a mistake to choose graphic feedback as the introductory default, even though it seemed to be the most popular choice of the operators. Similar considerations apply to the run-length option. Indeed, the breakdown by secondary parameter cells in Table C.7 indicates that data generated solely in the most conducive secondary conditions had effect sizes comparable to those seen in the prior PEAR experiments. While none of this explains why the relative insensitivity to these parameters observed previously should have changed, this presumption also now must be questioned.

4. ***Insensitivity to operator attitudes****: Various psychological or subjective parameters pertinent to operators' attitudes in addressing the experimental task, such as their prevailing emotional state, their sense of purpose or enjoyment, the laboratory ambience, the experimenter's expectations, and other environmental factors, would be adequately preserved in the aggregate by the operator selection and handling procedures exercised in the replication.*

Prior PEAR experience (Jahn & Dunne, 1988, 1997), supplemented by extant psychological and parapsychological literature (Rosenthal, 1963; Schlitz, 1986), suggested that certain aspects of the experimental ambience may be conducive to generation of anomalous effects. Examples include a friendly, relaxed, even playful atmosphere; a supportive attitude summarized as "permission to succeed;" a lack of pressure or urgency for success; an "unfocused" or "long-wavelength" state of thought and attention; etc. Given the nonreplication, however, it now appears that either these psychosocial factors are not so important or we failed to instill a propitious balance of them into our operators' experiences.

Possibly supportive of the importance and difficulty of maintaining these attitudinal factors is some mild evidence for an "epochal" segmentation of the chronological results from each laboratory. For example, with reference to the cumulative deviation graphs of Figure 5, we can identify in each laboratory's

full record long spans of HI–LO yield (FAMMI: trials 60,000–195,000; GARP: trials 245,000–345,000; PEAR: trials 195,000–350,000) that were quite comparable to those of the prior PEAR studies. The reality of such bimodal inhomogeneities in these databases, vis-à-vis chance excursions of binary random walks, cannot be confirmed statistically for this amount of data, but it is interesting to recall that the larger body of prior PEAR results also displayed a bimodal epochal character that took a statistically more convincing form. Specifically, there we found three virtually equal-length epochs, having strong performance over the first, chance performance over the second, and strong performance over the last (cf. Figure 12). While it is difficult to establish a Bonferroni-type correction factor for this sort of retrospective reexamination of an extant database, taken at face value the distinction between the three epochs is quite significant ($\chi^2 = 7.566$ on 2 $df$, $p = .0228$). The second epoch is a "nonreplication" of the first quite as stark as the overall PortREG nonreplication and is of a comparable scale. It was earlier noted that taking the overall prior PEAR database as a standard, the replication effort refuted the prediction at a level of $Z = -2.87$. Yet, Figure 12 shows us that when PEAR itself, employing a known, productive experiment with the same protocols and operator pools, generated an REG database of the scale of PortREG three times in succession, it failed to show anomalous yield *one time in three*. In this view, the joint failure of three laboratories to replicate is an event with $p = .037$, rather than the $p = .004$ one would infer from the above $Z$-score.

In both the prior PEAR and replication cases, the strong epochal results are diluted by the remainder of their respective databases. Nevertheless the presence of extended segments of high yield, and of negligible yield, in both the prior PEAR and in all three replication databases, raise valid questions concerning what subjective factors bearing on the operators or, for that matter, on the experimenters, prevailed during these lengthy periods of apparently successful replications, and did not in the other, nonproductive major segments.

5. ***Intention as primary correlate****: The specification and control of operator "intention" is adequate to designate this property as the primary correlate of the anomalous effects.*

While there is no doubt that the stipulation of an operator intention as BL, HI, or LO, irrevocably specified and recorded prior to initiation of an experimental run, qualifies as an objective index for the subsequent data, it is equally clear that the processes by which the operator assumes and deploys that intention are inherently subjective in character, and hence potentially vulnerable to any influences that alter that subjectivity. We need look no further than the substantial aberrations in *baseline* behaviors, or the ubiquitous constrictions of trial-level standard deviations, or the epochal successions just mentioned, to infer that subtle subconscious as well as conscious mental and emotional processes may be at work in conditioning the operator's expression of intention. How these processes react to the perceived "success" or "failure" of an

ongoing experimental run or of a previously completed series; to the operator's sense of "resonance" with the experiment; to the sense of importance of the achievement; or to the temporal variations in the operator's mood or state of health are not really illuminated by these experiments, and remain far from our grasp. What does emerge, however, is a legitimate question as to whether intention is the best primary correlate for such anomalies or, as suggested by the FieldREG experiments (Nelson et al., 1996, 1998), some subtler criterion for the requisite mind/machine "resonance" would be more fundamental, or at least complementary to it.

6. ***Replication criterion****: Successful replication validates the phenomenon; failure to replicate disqualifies it.*

The concept of objective replication or falsification is crucial to the exact sciences. Yet examples abound where varying degrees of compromise with rigorous replicability have been tolerated out of pragmatic necessity. For example, the essential indeterminancy of quantum events forced physicists to acknowledge that for some experimental configurations, no degree of control over the apparatus will allow the exact prediction of a single observation. Instead, exact prediction and measurement are reserved for ensembles and distributions, rather than for individual events, i.e., the definition of "replication" has been subtly changed to accommodate the intrinsic indeterminancy. Similar modifications are routinely applied in the study of dynamical chaos and complex systems, e.g., in fluid mechanical turbulence, granular media, fracture and fatigue processes, etc. Indeed, in any systems sufficiently complex that the validity of statistical limit theorems must be questioned, the concept of empirical replication may need to be modified. In our case, the potential indeterminancy of various physical outcomes is overlaid with a plethora of potentially relevant biological and psychological variables associated with the human operators and experimenters that may exceed our ability to specify, measure, or detect, let alone to control. To expect that these hypercomplex systems will submit to classical expectations of causality, determinism, and replicability may be overly presumptive.

Many attempts to address such mind/matter replication problems have been advanced in the recent literature. One of the authors (J.H.) previously proposed that failures to replicate frequently occur if a sequence of experiments is interrupted by an overall analysis of the results up to that point. He has termed this the "Meta-Analysis Demolition Effect" and has discussed its psychological and pragmatic implications (Houtkooper, 1994). Others have suggested that better understanding of the limitations on the dynamical replicability of unstable physical systems could benefit mind/matter interaction research, as well (Atmanspacher, 1997; Atmanspacher & Scheingraber, 2000). It has also been proposed that the lack of dependable reproducibility might be intrinsically related to the appearance of the anomalies, and thus constitutive of our understanding of them (Atmanspacher et al., 1999). Yet another approach has

treated all mind/matter interactions as inherently quantum mechanical in character, and thus prone to the intrinsic quantum uncertainties (Jahn & Dunne, 1986). Many other rigorous and speculative propositions could be cited, but the replication problem remains a central conundrum in this class of research.

7. ***Anomaly indicators****: Composite "bottom-line" mean shifts in directions of intention would be the primary indicators of anomalous effect; any structural anomalies would simply be embellishment thereon.*

While the overall mean-shift criterion is undoubtedly the simplest to specify, evaluate, and promulgate, it is not a particularly informative source for comprehension of any subtle psychophysical processes underlying the phenomenon. In prior work, whether "successful" by the overall mean-shift criterion or not, much more has been inferred from the structural details of the databases, than from their gross characteristics (Dunne, 1998; Dunne et al., 1994; Jahn, Dobyns, & Dunne, 1991; Jahn et al., 1997; Nelson et al., 2000). In this PortREG replication program as well, having acknowledged the bemusing failure to replicate the prior scale of "bottom-line" results, we are presented with an impressively deep reservoir of structural features that in their striking internal disparities may testify equally emphatically to a broad variety of operator influences. Just as those studies in human behavior that encompass many heterogeneous groups of people rarely yield results that are universally valid for all participating population subsets, so the broad range of personal characteristics of the operators of these experiments, if relevant at all, could be expected to express themselves in less-than-consistent, variably incoherent forms. In this view, the polyglot nature of the results is not so much paradoxical as it is consistent with, and even supportive of, the hypothesis that some human behavioral characteristic is indeed interacting with the machines.

Nor should we ignore the magnitude of this constellation of structural anomalies. Recall that those components encompassed by the Monte Carlo treatment stood out from chance at about the $p = .02$ level. But the other structural features uncovered in the data, which necessarily required alternative evaluations, contribute further to an overall chance unlikelihood that extends well beyond that. Specifically, Appendix II outlines conservative meta-analytical computations that place the composite structural anomalies at a level of chance expectation in the range of 0.001 to 0.002 (two-tailed). This approaches the level of significance that would have been achieved had the overall mean-shift replication been successful. That is, if the average prior PEAR H–LO mean shift, $\Delta$, had been sustained over the replication database, the corresponding $Z$-score would have been about 3.60 ($p = .0002$, one-tailed). In comparison, the equivalent $Z$-scores for the structural anomalies in the replication database range from 3.10 to 3.30, depending on the particular analysis base employed (cf. Appendix II).

While these reexaminations of presumptions and retrospective arguments clearly do not resolve our replication paradox, in some respects they may help

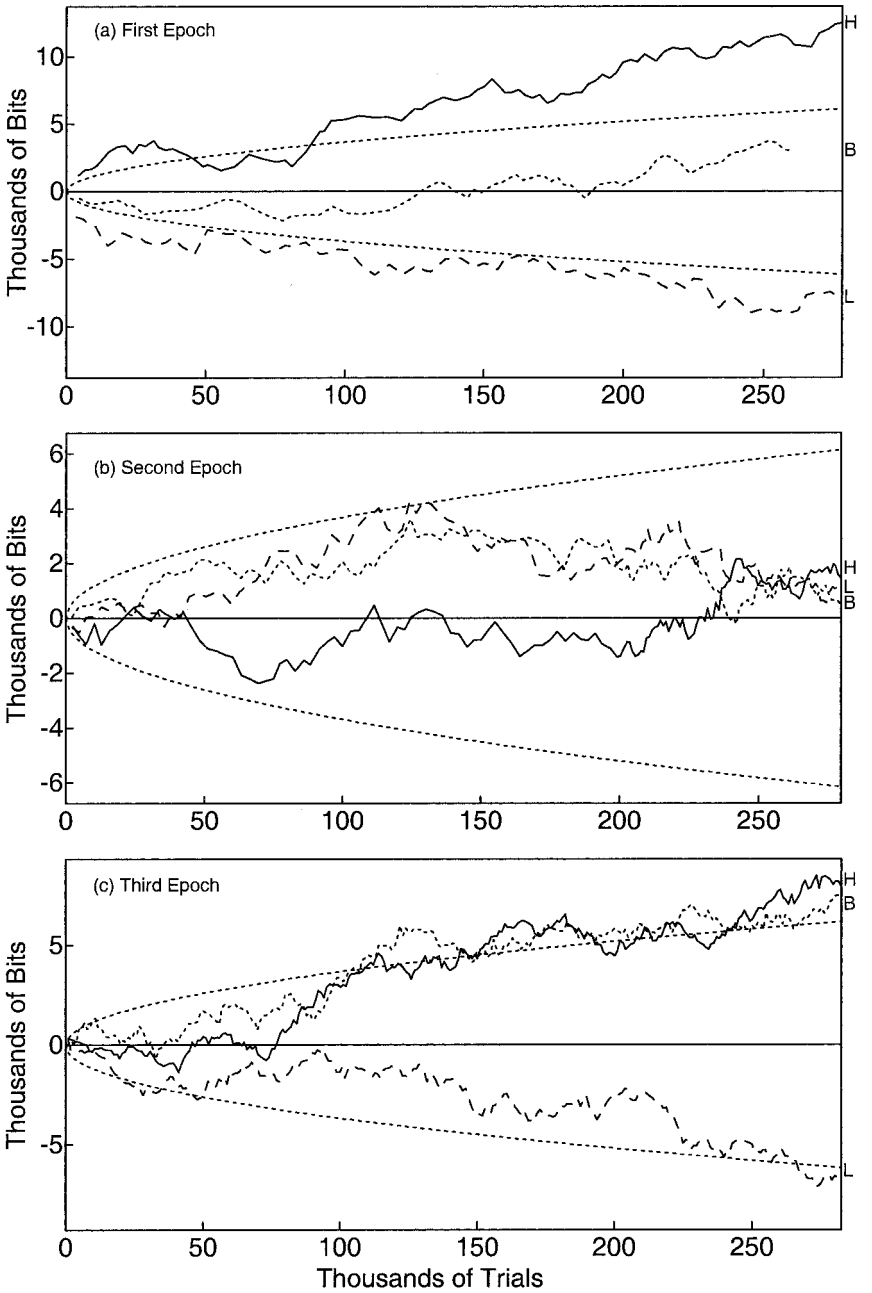


Fig. 12. Prior PEAR cumulative deviations in three epochs.

to focus suggestions for future research. The change from systematic, intention-correlated deviations to a comparably anomalous, albeit less orderly pattern of structural distortions testifies to our incomplete understanding of the basic phenomena, and warns that future empirical and conceptual efforts must proceed at a more sophisticated level. The next round of experiments and analyses will need to identify and address the implicit as well as the explicit assumptions, both in the initial designs and in the assessment of empirical results, and delve more deeply into the relationship between the anomalous manifestations and the underlying psychological and physical sources from which they emerge. No simpler conceptual route seems likely to prevail, but vigorous and insightful pursuit of this more difficult one not only may ultimately illuminate the particular mind/machine anomalies under study here but also may provide a much broader view of the relationship of the human mind to all physical reality.

## Appendix I: PortREG Equipment Calibrations

The protocol for the PortREG replication specified that concurrent calibrations be generated at each laboratory to correspond to each experimental session, using the same acquisition software but modified to run automatically. Beyond these, many other ad hoc calibration efforts were undertaken to establish that the REG devices were performing according to specifications and to characterize their performance in finer detail. Typically, the concurrent calibrations were generated following one or more experimental sessions, in blocks consisting of $3000 \times 200$-bit trials. Most were taken as 1000-trial runs, but some also were collected in 100-trial runs. Each of the laboratories collected more than the specified number of concurrent calibration trials from their respective REG sources. Specifically, GARP and PEAR generated over one million trials and FAMMI more than 850,000 trials. The results are displayed in Tables A1.F, A1.G, and A1.P. The first column of the tables lists the parameters computed in the standard suite of statistical tests for calibrations. Included are the first four moments of the statistical distribution, i.e., the Mean, SD (standard deviation), Skewness, and Kurtosis. The distribution of trial outcomes is compared with theoretical expectation by the standard $\chi^2$ calculations ($\chi^2$ Bins), and the standard deviation is calculated for blocks of 100 and 1000 trials (100-tr Sigma and 1000-tr Sigma, respectively). The distribution of runs of consecutive trials scoring greater than 100, and trials scoring 100 or less is compared with theoretical expectation ($\chi^2$ Runs), and a similar comparison against theory is made of the proportion of runs of length 50 remaining on one side of the origin (Arcsine). Finally, two autocorrelation functions are computed, for the raw trial sequences and for blocks of 50 trials (Autocorr Raw and Autocorr 50). The probability values are computed from the appropriate statistical indicators ($Z$-scores, F values, and $\chi^2$s).

In general, the consistency of the data and the deviations of parameter estimates are in accord with theoretical expectations for independent random bits

having binary probability of precisely .5, and hence these calibrations confirm the nominal statistical distribution of the overall data. However, a few specific departures from the theoretical distribution, and their implications for analysis of the experimental data, should be noted:

1. One of the most consistent structural departures from expectation in the experimental data occurs in the trial-level standard deviations shown in Table C.11. Thus, it is particularly important to examine the corresponding behavior of the calibration data. None of the three calibration databases shows a significant deviation from the nominal trial-level standard deviation of the appropriate theoretical binomial distribution. Specifically, there is only a slight increase ($p = .215$) in the FAMMI calibrations, and a slight decrease in the GARP and PEAR calibrations ($p = .66$ and $p = .61$, respectively). Therefore, it is valid to pool these values to an empirical standard of comparison for the experimental data, as described in the main text, section III.B.5.
2. The FAMMI calibrations show a marginally significant elevation in the trial-level goodness-of-fit $\chi^2$ test ($p = .045$), even though all four parameters of the trial-level distribution are nominal. Of greater concern is the fact that the standard deviations of both 100-trial and 1000-trial blocks are significantly elevated ($p = .001$, $p = .012$). Since the trial-level standard deviation is nominal, this indicates a nonindependence between trials, which produces increased average deviations at the block lengths used in the actual experiments. Taken at face value, this would suggest that the mean-shift $Z$-scores emerging from the FAMMI data are exaggerated by as much as 5.5%. (This is obtained by comparing the observed standard deviation of 1000-trial blocks, 235.871, to the theoretical value of 223.607; the ratio, 1.0548, is the factor by which $Z$-scores would be inflated by this departure from theoretical standard deviation.) The presence of intertrial dependence is confirmed by a significant autocorrelation ($p = .005$) at the trial level, driven by a succession of large, positive correlations at various lags, especially lags 5, 6, 10, and 12. A breakdown of the FAMMI calibrations into four roughly chronological sections shows that the amplified standard deviation of blocks is primarily in the first half of the data, particularly in the second quarter (series 50 to 99), which show a standard deviation increase as severe as 11.4% in the worst case. [The FAMMI team observed these deviant early calibrations and replaced the original device with a new one. No deeper examination was made, but the difference between the first and second half of the FAMMI calibrations suggests the source of the problem was some subtle malfunction of their REG device.] By any reasonable criterion, these aberrations should have no consequential impact on the primary or secondary FAMMI data, or the interpretation thereof.
3. The GARP calibrations fail of perfection only in being too good, with a $\chi^2$ for the deviation of the trial distribution so small that 97.6% of ran-

dom samples would be expected to show greater departures from the theoretical populations.

4. The PEAR calibrations show an elevated skewness, $\gamma_3 = 0.0057$, corresponding to $Z = 2.48$, in the trial distribution. The reasons for this are obscure; a chronological breakdown into 10 segments shows marginally significant positive skewness in Blocks 4, 6, and 7, with an overall bias toward positive skewness. The distribution among the blocks suggests that a small positive skewness is present throughout, with the increased population of significant outliers being a consequence of normal variation about this shifted mean. Since trial-level skewness is a departure from normality, which will be suppressed rapidly in calculations involving large numbers of independent trials, this is not considered a damaging aberration so long as the trials are independent. All of the PEAR results relating to intertrial structure are nominal, suggesting that the trials are indeed mutually independent, despite their distributional oddity.
5. The chronological breakdown of PEAR calibration data suggests the existence of a brief epoch (May through June 1998) during which trial-level standard deviation may have been suppressed. (In this segment, $\sigma = 7.0302$ and $p = .9970$. This result remains a $p = .03$ suppression even after Bonferroni correction for the examination of 10 subsets. Whether further correction for the many other parameters under scrutiny is appropriate here may be left to the individual analyst.) Since this epoch, even if it represents a genuine local suppression of standard deviation, corresponds to a concomitantly small proportion of the experimental data, and since the overall trial-level standard deviation of the calibrations is nominal, the previous remarks and conclusions concerning the *Z*-scores of Table C.11 do not need revision.

As a supplement to the concurrent trial-level calibrations, GARP also collected bit-level calibration data, to examine the behavior of the REG source at this finer scale. In contrast to the "quality control" approach of the concurrent calibrations, the GARP procedure is a "device properties" approach (Houtkooper, 1998), which examines short-term dependencies as characterized by Markov-chain transition probabilities. These are in straightforward relationship to traditional parameter-based tests, but this alternative allows more specific deviations from randomness to be scrutinized and permits calculation of standard deviations between sections of data and, hence, sensitive detection of episodic deviations from ideal randomness.

These bit-level data reveal an expected effect, namely a slight excess of the bit sequences 01 and 10 over 00 and 11. The source of the effect is the design of the REG, which includes an XOR alternating template to eliminate actual physical bias in the threshold setting of the comparator that defines voltage levels as bits. The size of this excess of alternations is on the order of a few parts in 10,000 and is detectable if data sets are accumulated over a few days. (The tests require on the order of 100 million bits.) Of course, the standard de-

TABLE A1.F
FAMMI Concurrent Calibrations (852,000 Trials)

| Parameter | Theory | Actual | Probability |
|---|---|---|---|
| Mean | 100.0000 | 99.9991 | .453 |
| Std.Dev. | 7.0711 | 7.0753 | .215 |
| Skewness | 0.0000 | -0.0001 | .479 |
| Kurtosis | -0.0100 | -0.0018 | .062 |
| $\chi^2$ Bins | 62 | 82.0572 | .045 |
| 100-tr Sigma | 70.7106 | 72.3696 | .001 |
| 1000-tr Sigma | 223.6068 | 235.8710 | .012 |
| $\chi^2$ Runs | 32 | 21.2935 | .925 |
| Arcsine | 50 | 46.8844 | .599 |
| Autocorr Raw | 25 | 46.9447 | .005 |
| Autocorr 50 | 25 | 22.3352 | .616 |

TABLE A1.G
GARP Concurrent Calibrations (1,165,000 Trials)

| Parameter | Theory | Actual | Probability |
|---|---|---|---|
| Mean | 100.0000 | 100.0002 | .490 |
| Std.Dev. | 7.0711 | 7.0691 | .661 |
| Skewness | 0.0000 | -0.0010 | .333 |
| Kurtosis | -0.0100 | -0.0134 | .227 |
| $\chi^2$ Bins | 62 | 41.9505 | .976 |
| 100-tr Sigma | 70.7106 | 70.9264 | .321 |
| 1000-tr Sigma | 223.6068 | 229.2080 | .113 |
| $\chi^2$ Runs | 32 | 22.4490 | .917 |
| Arcsine | 50 | 48.8943 | .518 |
| Autocorr Raw | 25 | 36.6390 | .062 |
| Autocorr 50 | 25 | 14.3219 | .956 |

TABLE A1.P
PEAR Concurrent Calibrations (1,130,000 Trials)

| Parameter | Theory | Actual | Probability |
|---|---|---|---|
| Mean | 100.0000 | 99.9998 | .488 |
| Std.Dev. | 7.0711 | 7.0697 | .613 |
| Skewness | 0.0000 | 0.0057 | .007 |
| Kurtosis | -0.0100 | -0.0143 | .175 |
| $\chi^2$ Bins | 62 | 64.9570 | .408 |
| 100-tr Sigma | 70.7106 | 70.8991 | .351 |
| 1000-tr Sigma | 223.6068 | 219.1441 | .829 |
| $\chi^2$ Runs | 32 | 28.0941 | .665 |
| Arcsine | 50 | 40.1359 | .839 |
| Autocorr Raw | 25 | 20.9494 | .695 |
| Autocorr 50 | 25 | 19.5021 | .772 |

viation of 200-bit trials is affected by interbit structural behavior on scales up to 199-bit sequence length and cannot be predicted reliably from this alternation excess. It is for this reason that the empirical standard deviation estimate

from the calibrations, including the empirical uncertainty thereof, was used as the standard of comparison for the statistical measures in Table C.11.

## Appendix II: Structural Meta-Analysis

The main text introduces, analyzes, and discusses many different structural features of the database, some of which prove to be individually significant, others not. The question to be addressed here is how to compound all such structural evidence into an overall statistical figure of merit. Specifically, the general problem of evaluating a number of distinct analyses on a collective basis is addressed by a meta-analytic technique.

It should be noted at the outset that not all of the participating structural analyses enter on an equal footing. Some of them are consequences of other analyses, i.e., they are re-examinations or more detailed investigations of effects that have already been evaluated in the other formats. Also, certain analyses were preplanned while others were retrospective. Moreover, while most of the analyses are based on the entire three-laboratory database, others are restricted to only single-laboratory data. The following numbered list introduces each of the structural analyses in the order they are encountered in the main text, describing its status in terms of the foregoing factors and providing any additional information required to specify how the conclusion of that specific analysis is reached. A probability value ($p$) is quoted for each analysis, to facilitate meta-analytic combination via the method of adding logarithms (Rosenthal, 1984).

1. The breakdowns by secondary parameters presented in Tables F.2 through F.5, G.2 through G.5, and P.2 through P.5 comprise a preplanned structural analysis, i.e., examination of these parameters was part of the original experimental design. While this is a complex calculation with many subparts, it has been collectively evaluated against the null hypothesis by the Monte Carlo analysis of section III.B.2, resulting in $p = .022$.
2. The series-position results, presented in Tables F.6, G.6, and P.6, constitute another preplanned analysis. The $\chi^2$ summaries in Table C.9 result in $p = .026$ after Bonferroni correction for including separate results from each of the three laboratories. (Only the rightmost column of Table C.9 is relevant, since the raw $\chi^2$ would respond to overall mean shifts, if any.)
3. Table F.7 reports a preplanned exploration of experimenter effects conducted only at FAMMI; combining the independent $\chi^2$ values results in $p = .887$.
4. Table G.7 reports a preplanned examination of control mode conducted only at GARP; constructing a $\chi^2$ from the independent $Z$-scores yields $p = .684$.

5. Table G.8 reports a preplanned examination of operator types conducted only at GARP; the composite $\chi^2$ corresponds to $p = .241$.
6. Following Table G.8, a few summary figures, and the discussion of the Monte Carlo analysis noted in Item 1 above, the next analysis in the text is the discussion of "favored cells" in section III.B.2.b. This retrospective analysis examines internal features of the structural qualities which have already been evaluated against the null hypothesis in Item 1. Although a $p$-value of .274 can be computed for this (by applying a Bonferroni correction to the most striking $Z$-score reported), it cannot properly be included in the meta-analysis.
7. In contrast, the correlation coefficients reported in the second half of Table C.7 are a retrospective examination of a different phenomenon. Such correlations between laboratories are independent of the Monte Carlo evaluation. After Bonferroni correction this yields $p = .243$.
8. The retrospective examination of unconfounded secondary parameters (Table C.8), like Item 6, is a direct consequence of the structural elements analyzed in Item 1. It produces $p = .031$ after Bonferroni correction but cannot properly be included in the meta-analysis.
9. Table P.7, presenting the evaluation of individual operator consistency between experiments, is an independent retrospective analysis, albeit one limited to a single laboratory (PEAR). This yields $p = .011$ after Bonferroni correction.
10. The summary of PEAR operator-specific performances presented in Table C.10 also qualifies as a preplanned analysis requested by certain of the authors. It yields $p = .051$ after Bonferroni correction.
11. The discussion following Table C.10 mentions, but does not report, a similar analysis based on a $\chi^2$ calculation for the series-level, rather than operator-level, data. This was a retrospective analysis that detects the same structural properties as the operator-specific analysis and must therefore be regarded as a derivative of Item 10; its $p$-value of .021 must therefore be excluded.
12. The trial-level standard deviation results in Table C.11 follow from a retrospective analysis that is independent of all previous analyses, with $p = .049$ after Bonferroni correction.
13. The counts of successful operators and series, mentioned in the last subsection of section III, are consequent to and dependent on the mean shifts. An earlier version of the Monte Carlo analysis incorporated these along with the mean-shifts and proved statistically consistent with the results of Item 1; thus, we may quote $p = .022$ for this but must consider it a consequent analysis and exclude it from the meta-analytic combination.
14. The previous 13 items cover all of the analyses presented in sections II and III, but for completeness, we must note one other independent retrospective analysis that was not included in the text. From earlier PEAR experience, it was speculated that the trial-level variance might be re-

duced in runs that were successful in the direction of intention, relative to its value in those runs contrary to intention. The calculated $p = .234$.

Table A2.1 summarizes these 14 analyses, now organized by category. The index numbers in the left margin of the table refer to the itemized list above.

To compound the results of a set of analyses individually reported as $p$-values, we may take advantage of the fact that under the null hypothesis $p$ is uniformly distributed between 0 and 1, whence $-2 \log(p)$ is distributed as a $\chi^2$ with 2 degrees of freedom ($df$). The addition properties of $\chi^2$ then guarantee that a sum of $n$ such values is a $\chi^2$ with $2n$ $df$ (Rosenthal, 1984).

Considering first only the preplanned analyses that incorporate the entire database, we have $p_i = \{.022, .026, .051\}$. This results in $\chi^2 = 20.885$ on 6 $df$, yielding a composite meta-analytic $p = .0019$. Adding the three retrospective analyses that cover the entire database increases this $\chi^2$ to 32.651, now on 12 $df$ so the meta-analysis reaches $p = .0011$. Finally, including the four analyses based on single-laboratory contributions increases $\chi^2$ to 45.538 and $df$ to 20, yielding $p = .0009$. Thus while the various analyses, which might be considered questionable due to retrospective status or limitation to a single laboratory, increase the statistical significance, they do so only by a factor of 2 from the initial figure for preplanned, whole-database analyses.

Including retrospective analyses raises the issue of the "file-drawer effect," where the visible results might spuriously overestimate an effect by overlooking an unreported background of null results. The standard measure for considering the possible impact of unreported studies is the number of such studies,

TABLE A2.1
Summary of Analyses

| Item | Form of analysis | $p$-value |
|---|---|---|
| Preplanned; using all data | | |
| 1. Secondary parameters | Monte Carlo | .022 |
| 2. Series position | Independent | .026 |
| 10. Operator performance | Independent | .051 |
| Retrospective; using all data | | |
| 7. Interlab correlation | Independent | .243 |
| 12. Trial-level $\sigma$ | Independent | .049 |
| 14. Success-based $\sigma$ | Independent | .234 |
| Preplanned; single-laboratory data | | |
| 3. Experimenter effects | Independent; FAMMI only | .877 |
| 4. Control mode | Independent; GARP only | .684 |
| 5. Operator type | Independent; GARP only | .241 |
| Retrospective; single-laboratory data | | |
| 9. Operator consistency | Independent; PEAR only | .011 |
| Reanalysis of effects already analyzed | | |
| 6. Favored cells | Consequence of (1) | (.274) |
| 8. Unconfounded parameters | Consequence of (1) | (.031) |
| 11. Series $\chi^2$ | Consequence of (10) | (.021) |
| 13. Operator and series counts | Consequence of (1) | (.022) |

with null outcomes, that would need to be added to the reported database in order to reduce the overall result to nonsignificance. For the current result, this file-drawer number is 14. Given the difficulty of finding any other new and substantive analyses that are not in some way reexaminations of structural aspects already considered, and given that this file-drawer number is equal to the total number of analyses already reviewed, including several such "duplicates" (6, 8, 11, and 13), it would seem that there is little risk of file-drawer dilution of this survey statistic.

In conclusion, the aggregate interpretation for the PortREG analyses with all multiple-testing and redundancy concerns taken into account is $p = .0009$ against the null hypothesis that the data contain no anomalous structures, or $p = .0019$ if only preplanned complete-data analyses are included (which has the virtue of rendering file-drawer considerations completely moot).

## Acknowledgments

The authors acknowledge with deep gratitude the financial support of the Institut für Grenzgebiete der Psychologie und Psychohygiene, which allowed this Consortium to be formed and this research project to be accomplished. In the interpretation of the experimental data and its dialogue with theoretical models, and in critical editing of this report, our consultation with Dr. Harald Atmanspacher and his colleague Dr. Werner Ehm have been invaluable. We also express our thanks to the many operators who generated the experimental data and to the many staff persons who helped in implementing these studies and this report, most especially Ms. Lisa Langelier-Marks and Ms. Elissa Hoeger.

## References

Atmanspacher, H. (1997). Dynamical entropy in dynamical systems. In Atmanspacher, H., & Ruhnau, E. (Eds.), *Time, temporality, now* (pp. 327–346). Berlin: Springer.

Atmanspacher, H., Bösch, H., Boller, E., Nelson, R. D., & Scheingraber, H. (1999). Deviations from physical randomness due to human agent intention? *Chaos, Solitons, and Fractals, 10*(6), 935–952.

Atmanspacher, H., & Scheingraber, H. (2000). Investigating deviations from dynamical randomness with scaling indices. *Journal of Scientific Exploration, 14*(2), 1–18.

Bierman, D. J., & Houtkooper, J. M. (1975). Exploratory PK tests with a programmable high-speed random number generator. *European Journal of Parapsychology, 1*(1), 3–14.

Bierman, D. J., & Houtkooper, J. M. (1981). The potential observer effect or the mystery of irreproduceability. *European Journal of Parapsychology, 3*(4), 345.

Braud, W. G. (1993). On the use of living target systems in distant mental influence research. In Coly, L. (Ed.), *Psi research methodology: A re-examination*. New York: Parapsychology Foundation.

Braud, W. G., & Dennis, S. P. (1989). Geophysical variables and behavior: LVIII. Autonomic activity, hemolysis, and biological psychokinesis: Possible relationships with geomagnetic field activity. *Perceptual and Motor Skills, 68,* 1243–1254.

Dunne, B. J. (1991). *Co-Operator Experiments with an REG Device* (PEAR Technical Report No. 91005). Princeton, NJ: Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Dunne, B. J. (1998). Gender differences in human/machine anomalies. *Journal of Scientific Exploration, 12*(1), 3–55.

Dunne, B. J., Dobyns, Y. H., Jahn, R. G., & Nelson, R. D. (1994). Series position effects in random event generator experiments, with appendix by Angela Thompson. *Journal of Scientific Exploration, 8*(2), 197–215.

Dunne, B. J., & Jahn, R. G. (1992). Experiments in remote human/machine interaction. *Journal of Scientific Exploration, 6*(4), 311–332.

Dunne, B. J., Nelson, R. D., & Jahn, R. G. (1988). Operator-related anomalies in a random mechanical cascade. *Journal of Scientific Exploration, 2*(2), 155–179.

Grad, B. (1963). A telekinetic effect on plant growth. *International Journal of Parapsychology, 5*(2), 117–133.

Houtkooper, J. M. (1994). Does a meta-analysis demolition effect exist? *Abstracts of 18th International Conference of the Society for Psychical Research, September 2–4* (pp. 14–15). Bournemouth, UK: Society for Psychical Research.

Houtkooper, J. M. (1998). IGPP Mind-Machine Interaction Consortium: Giessen Anomalies Research Program, MMI/PortREG Replication Phase 1. *GARP Technical Report, Draft 19981001*. Giessen, Germany: Center for Psychobiology and Behavioral Medicine, Justus-Liebig-Universität Giessen.

Jahn, R. G., & Dunne, B. J. (1986). On the quantum mechanics of consciousness, with application to anomalous phenomena. *Foundations of Physics, 16*(8), 721–772.

Jahn, R. G., Dobyns, Y. H., & Dunne, B. J. (1991). Count population profiles in engineering anomalies experiments. *Journal of Scientific Exploration, 5*(2), 205–232.

Jahn, R. G., & Dunne, B. J. (1988). *Margins of reality: The role of consciousness in the physical world*. New York: Harper Brace Jovanovich.

Jahn, R. G., & Dunne, B. J. (1997). Science of the subjective. *Journal of Scientific Exploration, 11*(2), 201–224.

Jahn, R. G., Dunne, B. J., Dobyns, Y. H., Nelson, R. D., & Bradish, G. J. (2000). ArtREG: A random event experiment utilizing picture-preference feedback. *Journal of Scientific Exploration, 14*(3), 383–409.

Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1987). Engineering anomalies research. *Journal of Scientific Exploration, 1*(1), 21–50.

Jahn, R. G., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., & Bradish, G. J. (1997). Correlations of random binary sequences with pre-stated operator intention: A review of a 12-year program. *Journal of Scientific Exploration, 11*(3), 345–367.

Nelson, R. D., Bradish, G. J., & Dobyns, Y. H. (1989). *Random event generator qualification, calibration and analysis* (PEAR Technical Report No. 89001). Princeton, NJ: Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Nelson, R. D., Bradish, G. J., Dobyns, Y. H., Dunne, B. J., & Jahn, R. G. (1996). FieldREG anomalies in group situations. *Journal of Scientific Exploration, 10*(1), 111–141.

Nelson, R. D., Bradish, G. J., Jahn, R. G., & Dunne, B. J. (1994). A linear pendulum experiment: Operator effects on damping rate. *Journal of Scientific Exploration, 8*(4), 471–489.

Nelson, R. D., Dobyns, Y. H., Dunne, B. J., & Jahn, R. G. (1991). *Analysis of variance of REG experiments: Operator intention, secondary parameters, database structure* (PEAR Technical Report No. 91004). Princeton, NJ: Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Nelson, R. D., Jahn, R. G., Dobyns, Y. H., & Dunne, B. J. (2000). Contributions to variance in REG experiments: ANOVA models and specialized subsidiary analyses. *Journal of Scientific Exploration, 14*(1), 73–89.

Nelson, R. D., Jahn, R. G., Dunne, B. J., Dobyns, Y. H., & Bradish, G. J. (1998). FieldREG II: Consciousness field effects: Replications and explorations. *Journal of Scientific Exploration, 12*(3), 425–454.

Peoc'h, R. (1995). Psychokinetic action of young chicks on the path of an illuminated source. *Journal of Scientific Exploration, 9*(2), 223–229.

Radin, D. I. (1997). *The conscious universe: The scientific truth of psychic phenomena*. San Francisco, CA: HarperEdge.

Radin, D. I., & Nelson, R. D. (1989). Consciousness-related effects in random physical systems. *Foundations of Physics, 19*(12), 1499–1514.

Rhine, J. B., & Humphrey, B. M. (1944). The PK effect: Special evidence from hit patterns. II. Quarter distributions of the set. *Journal of Parapsychology, 8,* 287–303.

Rosenthal, R. (1963). Experimenter attributes as determinants of subjects' responses. *Journal of Projective Technique and Personality Assessment, XXVII,* 324–331.

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

Schlitz, M. J. (1986). An ethnographic approach to the study of psi: Methodology and preliminary data. *Proceedings of presented papers, The Parapsychological Association 29th Annual Convention.* Rohnert Park, CA: The Parapsychological Association, 187–204.

Schmidt, H., & Pantas, L. (1972). Psi tests with internally different machines. *Journal of Parapsychology, 36,* 222–232.

Schmidt, H. A. (1970). A quantum mechanical random number generator for psi tests. *Journal of Parapsychology, 34,* 219–224.

Schmidt, H. A., Morris, R., & Rudolph, L. (1986). Channeling evidence for a PK effect to independent observers. *Journal of Parapsychology, 50*(1), 1–16.

Shapin, B., & Coly, L. (Eds.) (1985). The repeatability problem in parapsychology. *Proceedings of the 32nd International Conference of the Parapsychology Foundation, held in San Antonio, Texas, 1983.* New York: Parapsychology Foundation.