

Time Normalized Yield: A Natural Unit for Effect Size in Anomalies Experiments

Roger D. Nelson

Princeton Engineering Anomalies Research
School of Engineering/Applied Science
Princeton University, Princeton, NJ 08544

Submitted for publication to the Journal of Scientific Exploration
November 2005

Abstract

Comparing the yields in different anomalies experiments is important for both theoretical and practical purposes, but it is problematic because the effects may be measured on differing scales. The units in which experiments are posed vary across digital and analog measures recorded in a wide range of uniquely defined trials, runs, and series. Even apparently fundamental units such as bit rates may lead to disparate calculated effect sizes and potentially misleading inter-experiment comparisons. This paper seeks to identify a study unit that can render the results from various types of anomalies experiments on a common scale. Across several databases generated in the consistent environment of the Princeton Engineering Anomalies Research (PEAR) laboratory, yield per unit of time is the most promising of several measures considered. The number of hours during which participants attempt to produce anomalous effects can be consistently defined, and the time normalized yield $Y(h) = Z/\sqrt{h}$ is demonstrably similar across a number of human/machine experiments, with a magnitude of about 0.2. On both practical and heuristic grounds, this constitutes a *prima facie* case for regarding the time normalized yield as a natural metric for anomalous effects of consciousness.

Application to a broad range of experiments, including examples from other laboratories, confirms the viability and utility of a time-based yield calculation. A χ^2 test across 12 local and remote databases from PEAR's human/machine experiments indicates strong homogeneity. Inclusion of the remote perception database, which has a significantly larger yield at $Y(h) = 0.6$, immediately renders the distribution of effect sizes heterogeneous. These and other applications return reasonable and instructive results that recommend the simple, time normalized yield as a natural unit for cross-experiment comparisons permitting an integrated view of anomalies research results.

Introduction

Because of the very small size of effects, and the consequently weak signal-to-noise ratio typical in anomalies research, especially human/machine interaction experiments, there is considerable impetus to search for experiments that are more sensitive. This search also produces a growing body of data on an array of potentially relevant parameters that may help define and understand the anomalous effects. However, a concomitant result of this otherwise desirable research development is a proliferation of differing data units or measures, with the result that it is difficult and apparently inappropriate to combine or compare results across experiments. Thus, ironically, what should in principle be a richer and more comprehensive picture becomes fragmented in such a way that important features of commonality and difference are obscured.

Over the past few decades, a problem similar to this in various fields has been addressed by developing procedures for meta-analysis, or quantitative review, within the literature of a particular discipline or experimental paradigm (Glass, 1977; Rosenthal, 1991). Meta-analysis treats each of a body of experiments or experimental subsets (categories) as a data-point, and creates thereby a “higher level” database that permits rigorous and quantitative assessment of the full concatenation of available information. The key to this approach is that the experiments must be posed in well-defined, common units so that effect sizes expressed in these units can be combined and compared. Such meta-analyses in anomalies research have demonstrated the importance of aggregation within carefully circumscribed protocols (Utts, 1991). But specifying the unifying measure is not a trivial task. Important questions and generalizations become accessible only if it is possible to find a common, or “natural” unit in which to express effects generated in differing experiments that have the common purpose of assessing anomalous interactions of human consciousness or intentions. The present exploration considers several potentially viable units to determine which of them may be most appropriate as the basis for a natural and broadly applicable measure of the anomalous yield.

The term “effect size” is used informally for a variety of different quantities, often with a unique, local definition. A frequent usage refers to a shift in the experimental distribution mean relative to a standard. This measure allows comparison of effects across subsets within a particular research protocol, but it does not embody information about reliability of the estimates, nor is it possible to compare distribution means from experiments with different measures.

Conversion of the mean shift to a Z -score normalizes it in terms of its own standard error of estimate, and hence expresses effects in a nominally comparable unit, but the magnitude of the Z -score is dependent on the size of the database from which the mean is estimated, making it useful only for significance comparisons addressing the certainty with which experimental effects can be distinguished from each other or from chance fluctuations.

In order to establish relationships and summarize findings across different experiments, and to incisively assess factors that influence variations, several other effect size measures have been developed, together with combination and comparison procedures. Special purpose measures of anomalous effects have been suggested by Schmidt (1970), Timm (1973), Tart (1983), and others, but these all apply only when experiments share a common experimental and statistical paradigm. More recently, for purposes of meta-analysis, the issue has been given serious consideration by statisticians. Generally, an effect size is constructed by relating the mean shift or its test of significance to the size of the study, and numerous specific examples have been proposed (Cohen, 1977; Glass, 1977). One that is widely used is Cohen’s d , which is the ratio of the difference in

means to the pooled estimate for the population standard deviation, $d = (M_1 - M_2) / \sigma$, but there are inconsistencies in its application for correlated and uncorrelated observations, and practical interpretation is not straightforward. Rosenthal (1991) argues that the most generally applicable, readily interpretable, and consistently defined of several roughly equivalent effect size measures is the Pearson product moment correlation coefficient, which can be computed from a variety of different original statistics. It is related to Z by the function $r = Z / \sqrt{N}$, where N is the number of study units on which the Z -score is based. This measure expresses the difference between experimental conditions in units of the standard deviation of the raw data (usually called trials) from the experiment. It has come to be regarded as a canonical measure, but as we will see, it is not an appropriate standard for inter-experiment comparisons because the practical meaning of a trial varies greatly across experiments.

The purpose here is to examine structural analogs of r calculated using other study units in addition to the original trials or data points, renormalizing the Z -score to express experimental results in terms of some common metric that yields a consistent measure of anomalous interactions across differing experimental protocols. The criterion for success in this search for what might be termed a “natural scale” is based on the assumption that conscious intention to change the distribution of experimental data should have a similar yield when tested in different ways, albeit with variations attributable to real differences in operator performance, experimental conditions, and other variables. It should be clear that this fundamental idea of expected similarity or homogeneity across experiments, although reasonable, can only be tested inductively, by accumulating indications that it supports consistent and sensible interpretations. We will therefore look for a transformation that produces the smoothest, or most similar array of yields across a comparable set of experimental databases, intending to test it further by applying it to make comparisons among a broader assortment of experiments.

Several bodies of data from human/machine interaction experiments and remote perception experiments conducted over 15 years in the Princeton Engineering Anomalies Research (PEAR) program provide a rich source for comparisons, since all the experiments have been conducted in a consistent environment, with the same philosophical framework, personnel, and style (Jahn, et al, 1987). PEAR has large databases from each of these experiments, in which most factors are kept constant, where there is no filedrawer of unreported experiments, and wherein there are statistically significant effects and demonstrable internal structure.

Procedure

Five study units were chosen for this assessment: bits, information, trials, series, and time. To simplify comparison of the different transformations, performance in each of the human/machine experiments was represented by the “bottom-line” difference between results in the two intentional conditions (e. g., HI – LO), expressed as a Z -score. For each of the five different study units, the yield, $Y(x) = Z / \sqrt{N(x)}$, where N is the number of units of type x , was calculated for a representative body of data from each of several experiments. In most cases, a standard subset comprising equal amounts of data from the most prolific operators was used, since the full databases have large imbalances in the sizes of individual operator contributions.

Calculations were made for: (1) the actual number of binary decisions, i.e., the raw bit count, (2) the Shannon-Weaver information content, called the effective bit count, (3) the number of trials, or basic data records, (4) the predefined complete series or experiment, and (5) a time-based unit, the number of hours invested in the experimental effort. Some of these measures need more

explanation. Trials are typically the basic data record and the smallest feedback unit for a given experiment. The trial-based yield corresponds to the unit used for calculating the product moment $r = Z/\sqrt{N}$, which is the canonical effect size expressing deviation in units of the trial standard deviation. The series or experiment amounts to a teleological measure, since operators know that it comprises the basic goal-directed task. That is, although the series definitions are arbitrary and may change, series are invariably followed by the terminal feedback that tells the operator and experimenter what happened as a result of the operator's effort. For the time-based unit, a measure of the operator's subjective time would be ideal, but is not feasible, so an objective and readily calculated approximation was specified: In all the human/machine interaction experiments, the time period during which the machine is running and the target system is therefore labile or potentially vulnerable is well defined. The total time for the two intentional conditions was used. For remote perception experiments, 15 minutes per trial, as suggested by the standard protocol, was used for the time-based calculation.

The Experiments

A brief description of the essential features distinguishing the five experiments used for our assessment will indicate how they differ with regard to the physical systems and the particular measures involved. For each experiment, a "standard subset" was specified to minimize the impact of variations in individual operator contributions; in most cases, this was accomplished by using equal contributions from the relatively prolific operators.

The Random Event Generator (REG) experiments have a recorded data unit of trials, approximately one second long, that are the sums of 200 bits, taken in series with lengths ranging from 1000 to 5000 trials per intention (Nelson, Dunne, and Jahn, 1984; Nelson, et al., 1991; Nelson, et al., 2000). For the REG experiment, the standard subset employed for the basic calculations and comparisons was the first 10,000 trials produced by 30 operators who generated at least that many, drawn from the subset of all local, diode-based trials. The bit in the REG experiments is the well-defined, classical binary decision, which leads to a clear theoretical model and straightforward calculations. The Shannon-Weaver "effective information" content of an REG trial corresponds to the base 2 log of 200, or 7.64 bits, and represents the number of binary decisions required to precisely specify a trial outcome. (The sum of 200 bits is normally distributed, so that a more conservative measure could be used, but for this argument the simpler procedure will suffice.) On its face, this is a very attractive unit, but as will be shown later, it produces an unreasonably broad range of effect size or yield estimates, suggesting that the Shannon-Weaver formalism does not represent the fundamental currency in which anomalous information transfer should be measured. The amount of time invested by operators was defined as a function of the number of trials, or, equivalently, the period of time during which the experiment provides online feedback.

The Random Mechanical Cascade (RMC) experiment records the bin into which a ball drops after bouncing through the pin array, and calculation indicates that there are about 40 binary equivalent decisions or raw bits per ball, where the bit is defined as the "decision" between adjacent bins (Dunne, et al, 1988). The effective bit count per ball is the base 2 log of 40, or 5.32 bits of information. Again, this is a simplified approximation that is sufficient for present purposes; a rigorous account would include details of the distribution. Data are taken in 12-minute runs of 9000 balls, in series of 10 or 20 runs per intention, and Z-scores are calculated from the difference between distribution means in pairs of runs. For the RMC experiment the standard subset used was the first 10 datasets for 25 operators meeting this minimum.

In the Linear Pendulum (PEND) experiment (Nelson, et al, 1994), the base unit is the swing-to-swing change in velocity, derived from interrupts timed by a 50 nanosecond clock, and recorded as differences in the damping rate over the 200 swings in a three minute run. This is fundamentally an analog measurement, making it difficult to define a bit-counting measure of the effect, and an arbitrary surrogate was calculated by assigning one bit per swing, as if the difference between conditions at each swing were either positive or negative, discounting magnitude. Series comprised five or nine sets of runs, and the standard subset used for PEND was the first 25 sets generated by 18 operators with this number or more.

The measurable in the microelectronic shift-register experiment (CHIP) is the error rate in one-second trials of 1000 bits (Nelson, et al, 1992). The information content of a trial is 9.97 effective bits. Data were taken in runs of 50 trials and series of 25 runs. For the CHIP experiment, all data from the reliable “trials” protocol were used as the standard subset.

In the Remote Perception (PRP) experiments, the basic data for computer analysis are recorded in the form of 30 binary descriptors per trial, chosen by each of the two participants (Dunne, Jahn, and Nelson, 1983; Dunne, Dobyms, and Intner, 1989). Both agent and percipient address the task in a free response mode, during which they are certainly processing a large amount of information that only later is coded into the arbitrary descriptor format from which a score is computed. If the 30 bits were all informative and independent, the description would specify one from more than a billion alternatives. Partial inter-descriptor redundancy reduces the effective bit count by about 25%, yielding an estimated information content of 22.5 bits per trial. The standard subset for the PRP experiment used all formal data in the randomly instructed, *ab initio* encoded subset.

Results

The five different yield normalizations were applied to each of these experiments, using the standard data subsets described above. Table 1 shows these calculations, giving a Z-score for the experiment, and for each of the five measures, the number of study units, N, and the renormalized effect size, Y(x).

Table 1: Comparison of Yield Calculations

Measure	REG	RMC	PEND	CHIP	PRP
Z-score	2.780	1.763	.994	.554	3.122
Raw Bits, N	3.4e7	1.8e8	180400	770000	2820
Yield, Y(r)	.00047	.00013	.0023	.00063	.059
Effective Bits, N	4.5e6	3.3e7	23601	76261	2115
Yield, Y(e)	.0013	.00031	.0065	.0020	.068
Trials, N	588400	492	902	760	94
Yield, Y(t)	.0036	.079	.033	.020	.322
Series, N	136	25	90	16	12
Yield, Y(s)	.238	.353	.105	.139	.901
Hours, N	138	49	45	11	2
Yield, Y(h)	.236	.251	.148	.170	.644

To help visualize the degree of variation across experiments, Table 2 compares the five different calculations as ratios of the yield in the other experiments to that of the REG as a standard. The results are visualized graphically in Figure 1 and Figure 2.

Table 2: Yield Ratios for Five Measures

Measure	REG	RMC	PEND	CHIP	PRP
Raw Bits	1	.28	4.89	1.54	125.53
Effective Bits	1	.24	5.00	1.53	52.31
Trials	1	30.38	9.17	5.56	89.40
Series	1	1.48	.44	.58	3.79
Hours	1	1.06	.63	.72	2.74

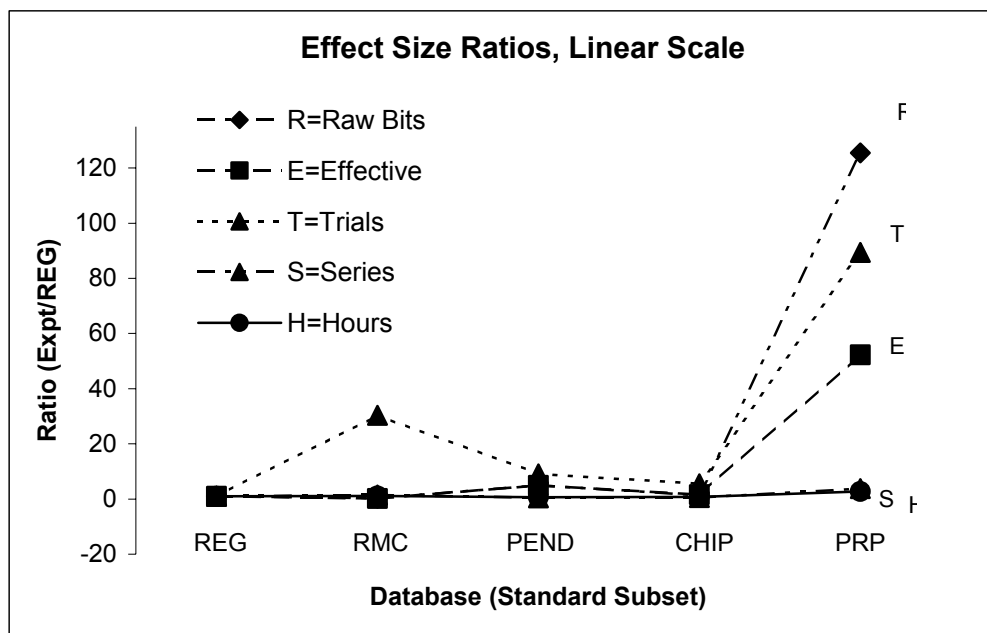


Figure 1: The ratio of the effect size for each of the five experiments is calculated relative to the REG effect size, and plotted on a linear scale.

In Figure 1, the linear scale allows a direct visual comparison of the relative consistency of the various measures. The yields calculated for both raw and effective bits range over two orders of magnitude across the five experiments, indicating that this apparently simple and fundamental measure cannot, in either form, serve as a general basis for inter-experiment comparisons, given the assumption that a natural scale should indicate homogeneity among scores purporting to measure the same phenomenon. Similarly, the trial, which is the basis for the nominal effect size, r , does not appear to provide a natural scale for anomalous effects. The figure makes it clear that variations in the definition of experimental units result in different patterns across the five yield calculations.

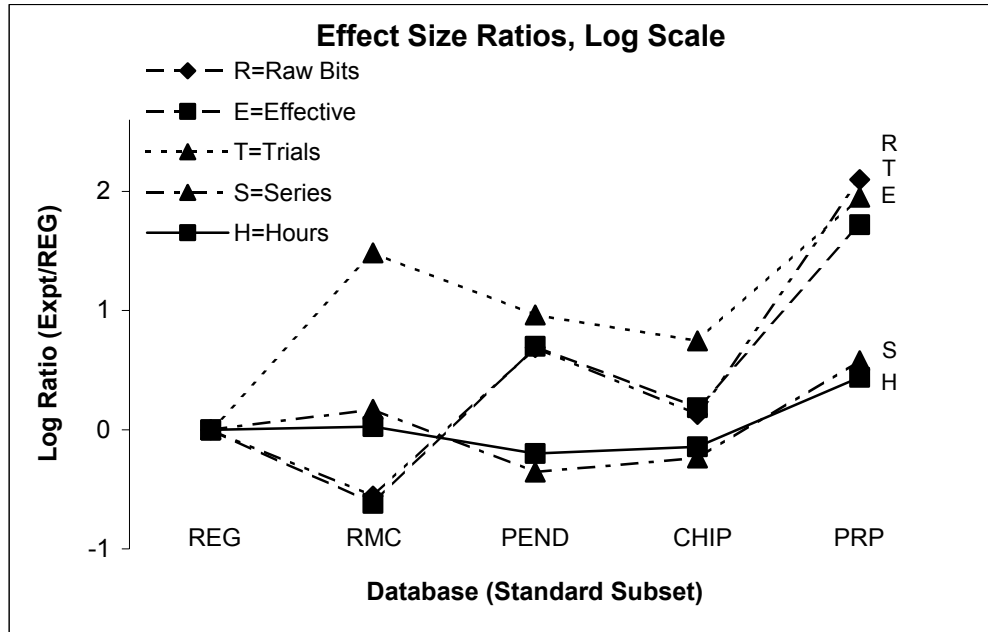


Figure 2: The ratio of the effect size for the five experiments is calculated relative to the REG effect size, and plotted on a logarithmic scale.

In Figure 2, a log scale is used for the same data, allowing a more detailed visual comparison of the relative consistency of the various measures. Here it is quite clear that there are orders of magnitude differences in the canonical, trial-based yield across experiments. This is the “effect size” that is most often published for anomalies experiments, and it is frequently invoked to compare experimental protocols (e.g., Targ, 2000). These results strongly suggest a need for careful reconsideration of such comparisons, and a search for an appropriate comparison standard; otherwise we may draw flawed conclusions about differences in effect size.

As noted, the bit and trial computations produce highly disparate results, but both the time-based and series-based calculations exhibit relatively similar yields across all experiments. This is a preliminary indication that the criteria for a useful standard might be met. The time-based measure presents the smoothest set of ratios. Now we must look more deeply to see whether its small advantage over the series unit is a substantial indication that results scale most naturally as a function of the time invested in their generation, or whether the teleological, goal-oriented measure represented by the completed experimental series is the fundamental unit in which anomalies might best be measured. This question can be quantitatively assessed by comparing data subsets where the predefined series length is changed within a particular experimental protocol, so that a given number of hours spent generating data is broken into differing numbers of series.

In the local, diode REG experiment at PEAR, series of 5000, 3000, 2500, and 1000 trials have been employed, and in the local RMC experiment, series of 20, 10, and 3 runs have been used. Table 3 and Figure 3 show the yield computations based on series, $Y(s)$, and time, $Y(h)$, with their standard errors (SE) for these seven datasets.

Table 3: Yield Transformed by Series and Time

Database	Z-score	N Series	N Hours	Y(s)	SE(s)	Y(h)	SE(h)
REG5000	3.472	17	40	.842	.243	.549	.158
REG3000	0.243	59	83	.032	.132	.027	.111
REG2500	1.359	86	102	.147	.107	.135	.099
REG1000	2.903	360	169	.153	.053	.223	.078
RMC20	3.335	26	208	.654	.196	.231	.069
RMC10	2.594	61	244	.332	.128	.166	.064
RMC3	-0.662	70	84	-.079	.120	-.072	.109

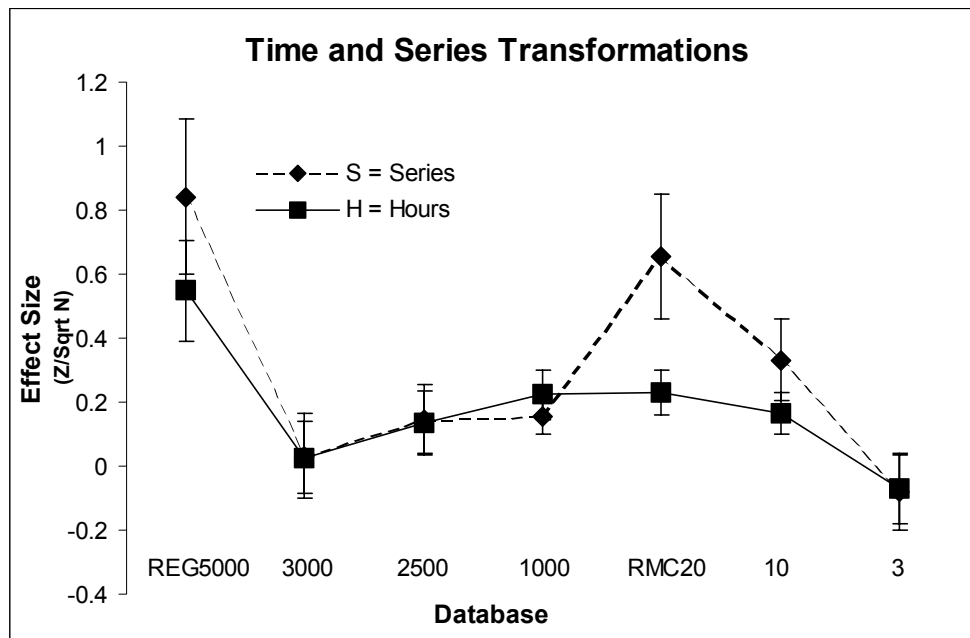


Figure 3: Yield computations for REG and RMC experiments with differing series lengths within otherwise consistent experimental protocols.

There is a significant positive correlation of the series-based yield, $Y(s)$, with the length of the series ($r = 0.845, p < 0.02$), whereas the corresponding correlation for the time-based yields, $Y(h)$, though positive, is not significant. A more direct test for our purposes, however, assesses the goodness-of-fit between the array of yield computations and our criterion of similarity, which can be modeled as a homogeneous distribution. Tests for homogeneity of the residuals from the mean across the seven yields shows that neither time nor series transformations can completely reconcile differences (χ^2 on 6 degrees of freedom = 17.4 and 27.3, respectively). However, two of the seven subsets have near-zero effects ($Z = 0.243$ and $Z = -0.662$, for the REG3000 and RMC3 experiments, respectively). Given the null effects, these cases are not useful in discriminating the series and time-based calculations. A common procedure used in meta-analysis to mitigate the effect of

outliers on estimations of effect size is to progressively exclude extreme values until a homogeneous distribution is achieved (Rosenfeld, 1975). This exercise identifies the REG3000 and RMC3 subsets as outliers, and if they are excluded, the picture sharpens: across the five remaining experiments, χ^2 for the time-based yields is 5.77 on 4 degrees of freedom, with $p = 0.23$, while the series based yields remain heterogeneous with $\chi^2 = 14.27$ and corresponding $p = 0.0035$. The time-based yields are statistically indistinguishable for four of the five remaining subsets, two from each experiment, while those based on the series measure show a component of variation proportional to the number of trials or length of the series, in addition to real differences that may exist among the subsets (e. g., the REG 5000 database has a relatively large effect size or yield by any standard).

Returning to the time normalized yield in the standard subsets, we find that none of the differences among the $Y(h)$ for the REG, RMC, PEND, and CHIP experiments approaches significance, and even that between PRP and a composite estimate for the others is only marginally significant. However, this latter difference appears to be real, as indicated by comparisons of the complete databases where error estimates are smaller. In these comparisons, the four human/machine experiments remain statistically indistinguishable from each other, while the PRP yield is significantly larger than REG ($Z = 3.59$), RMC ($Z = 3.51$), and PEND ($Z = 3.90$).

The calculation of time normalized yield, $Y(h)$, can be made with objectivity and repeatability, and it can be made with equal convenience not only for the various PEAR experiments, but for other laboratories as well, provided an adequate description of the experimental protocol is reported. It is encouraging that there is a demonstrable consistency across several quite different human/machine experiments. Only the PRP yield differs from the others, and it is a paradigm that differs in ways that should be instructive. Among other things, it is an information transfer experiment rather than a mind/matter interaction. It also involves two people, but even when the calculation is based on the time invested by both participants, the time normalized yield remains twice as large and significantly different from that in the human/machine experiments.

Using the fact that the yield per unit time is similar across a variety of related experiments to argue that the measure represents a natural scale for anomalous effects of consciousness is something of a bootstrapping operation, because the argument presumes an answer to one of the important questions for which an effect size or yield reconciliation is desired. Nevertheless, the balance of indications from these analyses, together with practical considerations, suggest that time normalization has broad generality. Analogy with the search for lawful relationships in the physical sciences suggests that an appropriate criterion for a useful metric is a simple functional relationship across a variety of applications, and time normalization does meet that criterion.

Applications

The time-based yield computation can be applied to a broader sample of experimental data, both to confirm its viability and to reveal some of the detailed information inherent in comparisons of experimental subsets within and across several research domains. The REG database is a primary resource, since it has a number of variants all using exactly the same basic design, but exploring parameters that give differing perspectives. Table 4 provides a comprehensive survey of the experiment, showing $Y(h) = Z/\sqrt{\text{Hours}}$ for the major variants and some of their subsets. In this and subsequent tables, an asterisk marks the standard subset used for the transformation comparisons previously shown in Tables 1 and 2.

Table 4: Time Based Effect Sizes, REG Experiment

Subset	Z-score	Ntrials	Hours	Y(h)	SE(h)
All Diode	4.379	2592450	609	.177	.041
First 10000*	2.780	588400	138	.236	.085
Local	3.809	1676450	394	.192	.050
Remote	2.045	792000	186	.150	.073
Diode A	3.103	1593200	374	.160	.052
Diode B	.849	124000	29	.157	.185
Diode C	1.173	618000	145	.097	.083
Diode D	2.153	174000	41	.337	.156
Diode X	3.519	83250	19	.796	.226
Oldreg 103	2.994	602450	142	.252	.084
Oldreg 87	3.615	522450	123	.326	.090
Remreg	1.541	1092000	257	.096	.062
Thoureg	3.289	898000	211	.226	.069
Co-operator	1.883	342000	80	.210	.112
Same-sex	-.815	158000	37	-.134	.164
Opposite-sex	3.324	184000	43	.505	.152
Bonded Pairs	2.976	60000	14	.794	.266
Unbonded	1.972	124000	29	.365	.185
Bonded Individuals	3.545	617150	145	.294	.083
Pseudo REG	1.988	293000	69	.240	.121
Ramp Frequency	2.765	247000	58	.363	.131
Fixed Frequency	-1.390	46000	11	-.423	.304
AT Pseudo (ATP)	-.444	964000	227	-.029	.066
Local	-.646	792000	186	-.047	.073
Remote	.897	128000	30	.164	.182
ATP B	-.866	44000	10	-.269	.311
ATP C	.836	122000	29	.156	.187
ATP D	.369	6000	1	.311	.842

* Data subset used for the Results section calculations.

A detailed description of the various subsets can be found in Nelson, et al (1991, 2000), but a brief accounting is in order. Three REG device types have been used, with the majority of experiments on a diode based “true” random source. Different locations include proximate (A), next room (B), remote (C), and remote, off-time (D). Some early experiments combined parameters within series (X). Oldreg, Remreg, and Thoureg are distinguished by the size of series and the general purposes of the experimental program. The subset names in the co-operator experiment are largely self-explanatory; the bonded individual subset is produced by the people who belong to bonded pairs, here working alone. The Pseudo REG (PREG) experiments use a 30-stage shift-register based pseudo-random sequence with a variable shift frequency (Ramp) or a fixed frequency (Fixed). Finally, the AT Pseudo (ATP) subsets use an algorithm seeded by a combination of the time-of-day and microsecond timer readings.

The majority of the subset yields clearly fall into the range for human/machine experiments shown in the last line of Table 1, with a few notable exceptions. The Diode X subset comprises the high scoring first series at the beginning of the research program (Dunne, et al, 1994), and reflects the performance of only a few individuals. The opposite-sex co-operators, especially the bonded pairs, also appear to generate larger effects, with $Z = 2.00$ for the difference between effects for bonded pairs and the standard subset; this is not due to the particular operators involved, since the difference compared with their combined individual databases is also impressive, with $Z = 1.79$. Even if the time for both operators is considered, reducing the calculated yield by a factor of $\sqrt{2}$, the opposite-sex yields remain relatively large. In contrast, the same-sex co-operators have a small negative result, significantly different from the standard subset ($Z = 2.00$). Other exceptions are the small or negative yields for ATP and for an exploratory database in the fixed frequency version of the hardware Pseudo experiment, both of which differ significantly from the standard subset ($Z = 2.46$ and 2.09 , respectively). It is instructive that the only major subset that shows an essentially null yield is the ATP database, which uses an algorithmic pseudorandom source. However, somewhat surprisingly, but of considerable theoretical interest, the remote subset of ATP shows a positive achievement comparable to the diode effect.

Looking at a finer level of detail within the REG database, some potentially instructive variations occur in the amount of operator time invested relative to the number of bits and trials, during explorations of different sample sizes and sampling rates. Table 5 shows $Y(h)$ in the full diode databases for sample sizes of 20, 200, and 2000 bits per trial, at sampling rates of 100, 1000, and 10000 bits per second.

Table 5: Time Based Yield: REG Diode, Sample Size and Rate

Size	Rate	Sec/Trl	Z-score	Ntrials	Hours	Y(h)	SE(h)
20	100	.792	-1.223	76000	17	-.299	.244
20	1000	.648	.820	6000	1	.789	.962
200	100	2.598	1.437	86900	63	.181	.126
200	1000	.846	3.848	2457150	577	.160	.042
200	10000	.648	.793	40000	7	.296	.373
2000	1000	2.640	2.634	34300	25	.525	.199
2000	10000	.846	.846	88000	21	.294	.220
20(010)	100	.792	-.248	30000	7	-.097	.389
100(010)	100	1.602	.502	12400	6	.214	.426
100(010)	1000	.744	.363	13250	3	.219	.604
200(010)	100	2.598	3.305	43750	32	.588	.178
200(010)	1000	.846	3.053	61400	14	.804	.263
2000(010)	1000	2.640	2.851	25300	19	.662	.232

Since some of the databases are quite small, and hence representative of only a few operators, the table also shows a set of results from one prolific operator, 010, in which variations due to differences among individuals are excluded.

This table indicates that Y(h) is of roughly similar magnitude in most of these sub-sets, with a trend toward larger yields for larger sample sizes. Similarly, there is a trend toward larger yields for faster rates, although few of the apparent differences approach significance. Figures 4 and 5 show these trends, using the full database calculations (except for the 100-bit sample size, which was explored only by operator 010). Neither the sample size nor the sampling rate trend is significant, although that for sample size has a Z-score of 1.60 for the slope coefficient, but they suggest structure and indicate that a closer look, disentangling the size and rate interaction, should be informative.

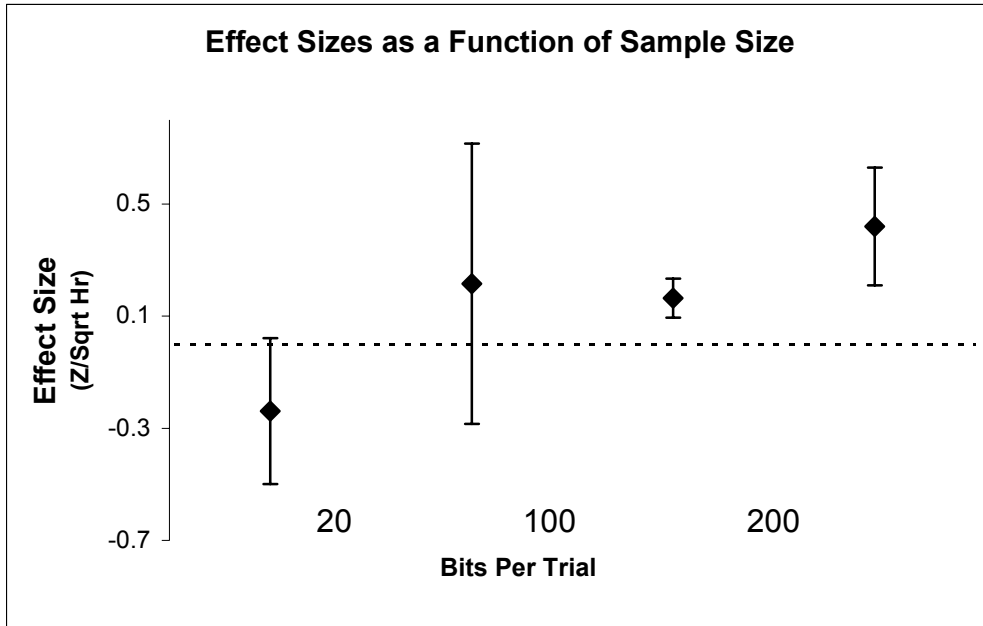


Figure 4: Time normalized REG yield, $Y(h)$, as a function of sample size.

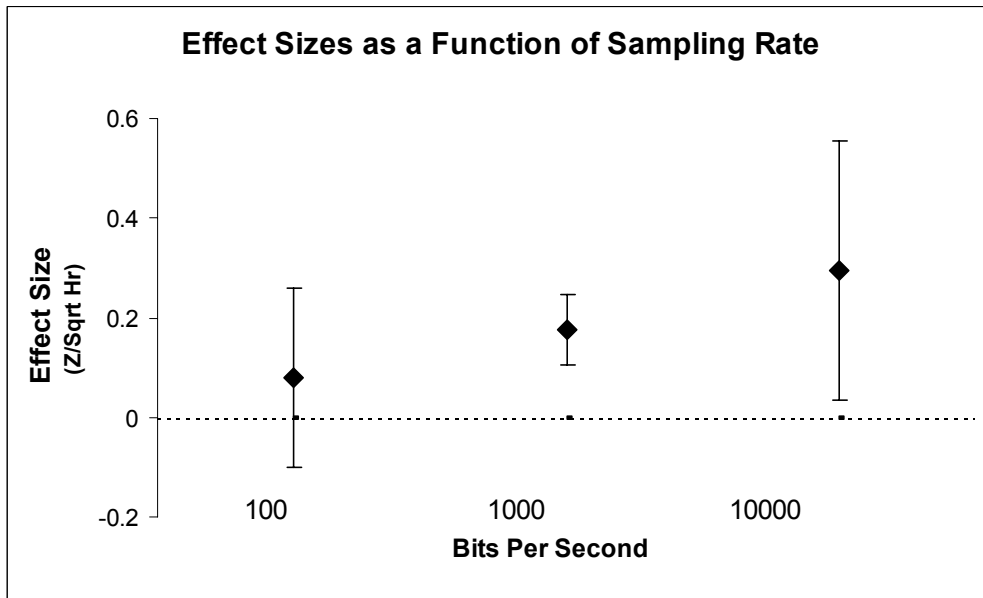


Figure 5: Time normalized REG yield, $Y(h)$, as a function of sampling rate.

The RMC experiment, shown in Table 6, was originally designed to have 20 sets of runs for a complete series. This was later shortened to 10 runs for operator comfort. 25 operators produced 87 series, with significant overall results (Dunne, et al, 1988). Subsequently, the nominal series was shortened still further to three sets, and a new exploratory database (RMC3) was started, with the goal of addressing certain questions inspired by the original experiment. In the latter, much smaller database, the overall effect is reversed, and has roughly the same magnitude as the positive effect in the original experiment. Despite the small size of the RMC3 database, the difference is significant, with $Z = 3.26$, but an attempt to interpret the difference is beyond the scope of this paper.

Table 6: Time Based Yield: RMC

Subset	Z-score	N trials	Hours	Y(h)	SE(h)
All 10, 20	4.264	2780	556	.181	.042
First 10 Sets*	1.763	246	49	.251	.143
Local	3.891	2262	452	.183	.047
Remote	2.139	518	104	.210	.098
All RMC3	-1.610	610	122	-.146	.091
Local	-0.662	420	84	-.072	.109
Remote	-1.914	190	38	-.310	.162
All RMC	4.063	3390	678	.156	.038
Local	3.813	2682	536	.165	.043
Remote	1.759	708	142	.148	.084

* Data subset used for the Results section calculations.

The Pendulum (PEND) experiment, presented in Table 7, has significant internal structure, even though the overall HI – LO difference is not significant. The largest contributions to the structure arise from the difference between subsets with volitional vs. instructed assignment of intention (Nelson, et al, 1994). Two versions of the experiment are presented in the table. The upper portion of Table 7 shows the full database as of February 1993, at which time the decision was made to close the global concatenation. The second part shows the database as of June 1992, which was described in a presentation to the annual meeting of the Society for Scientific Exploration (Nelson and Bradish, 1992). The bulk of the subsequent data are from one operator with a very large remote database (600 trials, more than half the new data) in which there is a marginally significant negative yield. The SSE database therefore may give a more representative indication of the effects in this experiment. The remote yield in the SSE subset is considerably larger than that in the local data, a difference that persists in the full database (although it is reduced by the hyper-prolific operator's contributions.)

Table 7: Time Based Yield: PEND

Subset	Z-score	N trials	Hours	Y(h)	SE(h)
All PEND	.713	3090	155	.057	.080
First 25 Runs*	.994	902	45	.148	.256
Prolific Only	1.785	2622	131	.156	.087
Local	.388	1830	92	.040	.104
Volitional	-1.505	842	42	-.232	.154
Instructed	1.958	988	49	.279	.142
Remote	.667	1260	63	.084	.126
SSE PEND	1.607	1902	95	.165	.103
Local	.709	1620	81	.079	.111
Volitional	-1.315	782	39	-.210	.160
Instructed	2.314	838	42	.357	.154
Remote	2.362	282	14	.629	.266

* Data subset used for the Results section calculations.

The PRP experiment has a number of instructive subset divisions, among which a particularly interesting one is the distinction between trials done in the volitional mode, where agents freely select targets in their location at the time specified for the trial, and instructed trials drawn randomly from a large prepared pool. A criticism of the PRP experiments (Hansen, et al, 1992) suggested that the volitional trials were vulnerable to “shared biases”. A detailed response (Dunne, et al, 1992) showed this concern to be unwarranted, and as may be seen in Table 8, the allegedly flawed volitional trials have a considerably smaller yield than those in the apparently safer instructed protocol.

Table 8: Time Based Yield: PRP, 15 Minutes per Trial

Subset	Z-score	N trials	Hours	Y(h)	SE(h)
All Formal	6.355	336	168	.693	.109
Instructed, ab initio*	3.122	94	47	.644	.206
Volitional	3.549	211	106	.489	.137
Instructed	5.771	125	63	1.032	.178
Ex post facto	5.792	59	30	1.508	.260
Ab initio	4.578	277	139	.550	.120

* Data subset used for the Results section calculations.

The table also provides a comparison of trials directly encoded in the binary descriptor list (ab initio) vs. those encoded from transcripts (post facto). If both the agent and percipient times are considered to be instrumental in this experiment, the yield and the standard error are both reduced by a factor of $\sqrt{2}$, but even in this case the overall yield remains a factor of two larger than is typical in the human/machine interaction experiments.

Finally, results from two relatively small human/machine experiments are shown in Table 9. These were both terminated as active experiments, even though they showed promise, before large databases could be obtained. The Fabry-Perot Interferometer experiment (FPI) proved to require too great a proportion of laboratory resources in order to provide adequate control of the environmental influences on the extremely sensitive instrument (Nelson, 1982). The microelectronic CHIP experiment could not be continued because the adequately controlled “trials” protocol was too demanding and uncomfortable for operators.

The “trials” subset of the CHIP database, although quite small, was generated in a fully competent and reliable experiment, and could therefore be included in the comparisons described in the Results section. The “runs” protocol was potentially vulnerable to large error rate fluctuations due to regime changes traceable to temporal variations in the microscopic behavior of electronic components, and the very large yield is suggestive of an artifactual inflation. This is an exemplary case showing how the comparison of $Y(h)$ with values typical of related experiments may help identify extreme outliers and lead to detection of design vulnerabilities.

Table 9: Time Based Yield: CHIP, FPI

Subset	Z-score	N trials	Hours	Y(h)	SE(h)
CHIP, Trials*	.554	760	11	.170	.306
CHIP, Runs	7.331	650	10	2.318	.316
FPI, Operator	2.258	60	10	.714	.316
FPI, Oper and Exptr	2.258	60	20	.505	.224

* Data subset used for the Results section calculations.

The FPI experiment used a bipolar protocol, making it potentially more vulnerable to artifacts than our standard tripolar experiments. Its yield appears to be larger than that of the other human/machine experiments, but the error estimate is commensurately large and the difference does not approach significance. The smaller yield shown in the last line of the table reflects the requirement in the FPI experiment for an experimenter to be present and to know the intention for the trial, and thus be a potential contributor, in the sense that he or she may also have an intention and at least unconsciously participate in the anomalous interaction.

Inter-Laboratory Explorations

As specific examples of the potential utility of the time normalized yield measure for exploration of the broad range of questions that might be asked in anomalies research, three calculations were made for non-PEAR research with commonalities and differences that are instructive. In all three cases, there is an expectation of a relatively large effect size or yield, based on the protocol.

Helmut Schmidt has a large body of REG-type experiments, addressing a number of issues common to the PEAR experiments, but using different approaches in some respects, most notably by pre-selecting subjects based on pilot tests. The question can be asked whether selected subjects actually produce larger yields, and if so, an estimate of their relative efficiency can be made, e. g., by comparing Schmidt’s time normalized yield with the PEAR results. One of the best protected of his experiments was done in collaboration with Morris and Rudolph (Schmidt, et al, 1986), and nicely excludes vulnerabilities to potential spurious effects and various criticisms through its multi-experimenter design and implementation. This experiment uses pre-recorded radioactive decay based seed numbers for an algorithmic pseudo sequence that determines the behavior of visual or auditory feedback.

It can be argued that there are more participants in Schmidt’s experiment than the nominal subject. The experimenter uses a true random event source to generate a set of seed numbers, hoping, one may presume, they will turn out to be interesting. The second observer generates a true random sequence of target assignments, probably with a similar state of mind. Finally, the subject spends on the order of a minute per trial, attempting to influence the outcome of the experiment. The upper part of Table 10 shows Schmidt’s yields calculated as if there were one, two, or three participants contributing to the anomalous result, using a time per trial in the middle of the range indicated in the published report. All of these are indeed larger than Y(h) for the standard PEAR REG database, but quite similar to those for some of the smaller subsets and for selected operators.

Table 10: Time Based Yield: Schmidt; Braud; Honorton

Experimenter	Subjects	Rate	Z-score	N trials	Hours	Y(h)	SE(h)
Schmidt*	Subject	.75min/trl	2.73	1040	13	.757	.277
Schmidt	Subject, Exper	.75min/trl	2.73	1040	26	.535	.196
Schmidt	Subj, Expr, Obsrvr	.75min/trl	2.73	1040	39	.437	.160
Braud*	Participant	16min/trl	1.97	960	16	.492	.250
Braud	Participant, Helper	16min/trl	1.97	960	32	.348	.177
Honorton*	Sender	6min/trl	3.89	355	36	.653	.168
Honorton	Receiver	30min/trl	3.89	355	178	.291	.075
Honorton	Julliard Student	6min/trl	2.20	20	2	1.556	.707
Honorton	Selected Subject	6min/trl	.69	7	.7	.824	1.195

* Data used in Figure 6

The second example is an exploration of a potentially more labile anomalous interaction in an experiment that assesses direct mental influence of one person on the activity of another (Braud, et al, 1995). It asks whether a participant’s ability to focus attention upon an object can be facilitated by a distant, isolated “helper”. Significant differences were found in the number of self-reported distraction episodes in randomly interspersed control and helping periods. Each session contained eight one-minute segments for each of the conditions, and the total time for both was used for the

yield calculation, shown in the second part of Table 10. The resulting $Y(h)$ is somewhat more than twice as large as the standard REG yield, and the difference is highly significant ($Z = 4.1$).

For the third example, claims of larger effect sizes depending on special conditions may be tentatively evaluated by cross-experiment comparisons of yield, in the absence of direct intra-experiment evidence. The Ganzfeld experimental program designed by Honorton (Bem and Honorton, 1994) has common elements with our PRP experiments, but again, some important differences. In particular, the Ganzfeld situation of reduced sensory input is held to be more conducive for anomalous information acquisition than simpler free-response protocols, although this expectation is based largely on theoretical considerations rather than on specific comparisons. The strongest set of these experiments is a database generated in a design meeting stringent criteria discussed in the Honorton-Hyman debates and description of ideal protocols (Hyman and Honorton, 1986). It is referred to as autoganzfeld, and incorporates excellent controls (Honorton, et al, 1990).

In this experiment also there are alternative ways to define the time invested. As in the PRP experiments, there are two participants, and the time spent by both might be included, but for this analysis only one person's time is counted. (As before, the two-person yields would be a factor of $\sqrt{2}$ smaller.) The receiver is in the Ganzfeld situation for 30 minutes, and the sender sees six one-minute presentations of the target over the course of the half hour. Calculations for both times are shown in the third part of Table 10. Also included are two data subsets from special or selected subject populations to indicate the range of yields in this experiment. One group (Julliard students) represents an artistic population; the other was selected on the basis of prior performance.

The Schmidt example provides moderate evidence that in structurally similar experiments, selected subjects can generate larger yields, by a factor of at least two. All three of the Schmidt yield estimates are larger than that for the standard PEAR REG, and although the error estimates are commensurately large, the difference is highly significant ($Z = 6.7$). In the Braud experiment, which is part of a program studying anomalous inter-actions with living systems, there is again a significantly higher yield compared with the REG experiment, by a factor of about two. We should note that Braud, et al's participants were friends and acquaintances of the helpers, and that some of the REG co-operator subsets have equal or larger yields, suggesting an alternative interpretation based on multiple-subject cooperation. The overall Ganzfeld yields are very much in line with PEAR's standard PRP results, and the largest, based on the 6 minutes/trial rate, is almost identical (.653 for Ganzfeld and .644 for PRP). Both experiments also show a similar range of yield variations across subsets. This constitutes suggestive evidence that the Ganzfeld procedure does not, as is widely believed, enhance anomalous information transfer over an unconstrained free-response approach; at least it indicates the question is open, and deserves direct scrutiny in appropriately designed research.

Discussion

A fundamental objective in all these experiments is to acquire data that address the anomalous interactions of consciousness with its environment. Considering the experiments from the point of view of the participants, one commonality is clear: there is a period of time during which the person is engaged in the experimental task, with intentions to produce anomalous results. Since the anomalies are correlated with these intentions, whether in the REG, RMC, PEND, CHIP or PRP experiments, a natural unit for comparable yield calculations is arguably the length of time spent by

the operator or percipient doing the experiment. This analysis shows that time normalization does give sensible results, specifically, a high degree of consistency among calculated yields for a variety of human/machine experiments. In contrast, the binary and information measures, and the trial unit all indicate yields ranging over orders of magnitude, which does not seem sensible for experiments that all attempt to establish and measure essentially the same phenomenon. Our teleological unit, the series, approaches the consistency of the time-based measure, but detailed examination shows it is correlated with the size or length of the series. Moreover, it is impractical because it is arbitrarily defined and not generally applicable.

Figure 6 graphically displays the uniformity of the time measure $Y(h)$ across a broad spectrum of independent subsets of the human/machine and information transfer experiments, as well as the stark exceptions to the rule. It includes local and remote variants of the REG, ATP, RMC, and PEND experiments, the PseudoREG, RMC3, and CHIP databases, the PRP database, and the three examples drawn from non-PEAR research, Schmidt (HS), Honorton (CH), and Braud (WB). A χ^2 test across the 12 local and remote databases from PEAR's human/machine experiments yields 5.78 on 11 degrees of freedom, indicating strong homogeneity. The distribution of yield measures becomes heterogeneous if the PRP database is added ($\chi^2 = 28.5$ on 12 df, $p = 0.0046$). Adding the non-PEAR databases singly does not produce significant heterogeneity, but the combined effect of these and the PRP database produces a highly significant χ^2 of 44.6 on 15 df.

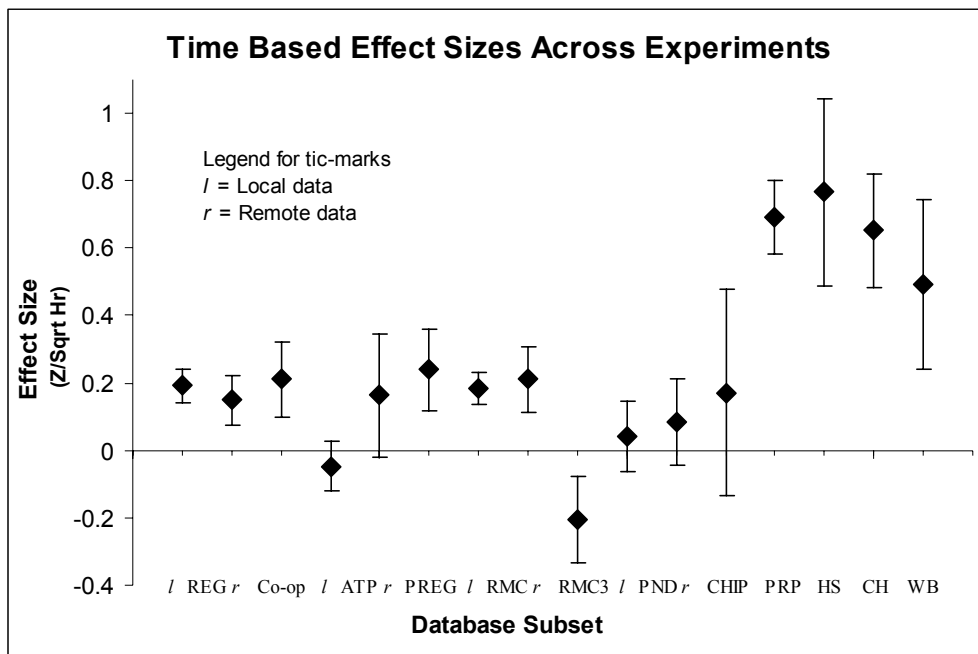


Figure 6: Time normalized yields, $Y(h)$, across a wide range of experiments.

A few examples of applications for the time normalization approach to cross-experiment comparisons suggest the power and flexibility of this perspective.

1. One of the motivating questions for the development of the PEND experiment was whether an analog device might be more accessible or vulnerable to anomalous interactions than digital experiments. The answer suggested in this analysis is that there is no such advantage.
2. The CHIP experiment in its best-controlled form could not be pursued beyond a pilot database for technical reasons, and it did not establish a persuasive level of significance. These comparisons show, however, that it did have a yield comparable to the other human/machine experiments, suggesting that the behavior of a fundamental electronic device such as a shift register may be vulnerable to an influence of consciousness. Results generated in the less completely controlled protocol were shown by this analysis to be outliers, demonstrating the need for experimental refinement.
3. Application of this strategy on an operator specific basis to the comparison of yields across several experiments could provide another incisive test of the viability of time normalization, as well as insight into the relative vulnerability of different physical systems: does the particular device matter, or are operators' effects independent of the device? Preliminary work shows that there is indeed consistency of the time normalized yield across multiple experiments for individual operators. A computation based on all data from operators who have generated databases in two or more PEAR experiments indicates a Bayes factor of 11 (roughly equivalent to a p-value of 0.03 for a chance result) supporting the intra-operator consistency hypothesis.
4. Remembering that other moderators may need to be considered, the two or three times larger yield in the PRP experiment suggests that it is more efficient, implying greater statistical power to detect anomalous interactions. It may be possible to determine whether this is a function of the protocol or the involvement of two participants, by direct comparison of the PRP experiment with otherwise similar remote perception experiments involving only a single participant. Indeed, it should be instructive to compare yields in various one and two person anomalies experiments, for example, within the PEAR database of multiple operator experiments, and by examining the telepathy vs. clairvoyance literature in parapsychology. As noted earlier, the overall REG co-operator yield closely resembles the single operator REG yield, suggesting that remote perception (information transfer) effects may be inherently larger as a function of the particular task.
5. With the caveat that a broader survey is required, the exploratory applications to other researchers' work promise useful, quantitative results. Comparisons of $Y(h)$ from selected and unselected subject populations suggests a considerable difference in yield for the former, and an implied commensurate research efficiency. Anomalous interaction with physiological systems, in this case a two-person interaction, also appears to promise a substantial increase in yield. Finally, the remote perception and Ganzfeld protocols appear not to differ, despite widespread belief in the efficacy of the Ganzfeld, but both show a factor of two or three larger yield than is typical for human/machine experiments.

These examples from intra- and inter-laboratory comparisons are interesting in their own right, and they provide tentative answers to questions of considerable importance for anomalies research. In addition, the results seem reasonable, and as such constitute a substantial inductive argument for the viability of $Y(h)$ as a time-based natural scale for anomalous effects.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Bem, D. J., and Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, **115**, 4-18.

Braud, W., Shafer, D., McNeill, K., and Guerra, V. (1995). Attention focusing facilitated through remote mental interaction. *J. Amer. Soc. Psychical Research*, **89**, 2, 103-115.

Dunne, B. J. (1991). *Co-operator experiments with an REG device*. Technical Note PEAR 91005, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

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

Dunne, B. J., Dobyns, Y. H., Jahn, R. G., and 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., and Nelson, R. D. (1983). *Precognitive Remote Perception*. Technical Note PEAR 83003, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Dunne, B. J., Nelson, R. D., and Jahn, R. G. (1988). Operator-related anomalies in a Random Mechanical Cascade. *Journal of Scientific Exploration*, **2**, 155-179.

Dobyns, Y. H., Dunne, B. J., Jahn, R. G., and Nelson, R. D. (1992). Response to Hansen, Utts, and Markwick: Statistical and methodological problems of the PEAR Remote Viewing (sic) experiments. *Journal of Parapsychology*, **56**, 115-146.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, **5**, 351-379.

Hansen, G., Markwick, H., and Utts, J. (1992). Critique of the PEAR Remote Viewing experiments. *Journal of Parapsychology*, **56**, 97-114.

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., and Ferrari, D. C. (1990) Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology*, **54**, 99-139.

Hyman, R. and Honorton, C. (1986). A Joint communique: The psi ganzfeld controversy. *Journal of Parapsychology*, **49**, 3-49.

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

Nelson, R. D. and Bradish, G. J. (1992). *A Linear Pendulum experiment: Operator effects on damping rate*. Internal Document PEAR 92003, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Nelson, R. D., Bradish, G. J., Jahn, R. G., and Dunne, B. J. (1994). A linear pendulum experiment: Operator effects on damping rate. *Journal of Scientific Exploration*, **8**, 4, 471-489. (Also available as Technical Note PEAR 93003).

Nelson, R. D., Dobyns, Y. H., Dunne, B. J., and Jahn, R. G. (1991). *Analysis of variance of REG experiments: Operator intention, secondary parameters, database structure*. Technical Note PEAR 91004, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Nelson, R. D., Dunne, B. J., and Jahn, R. G. (1982). Psychokinesis studies with a Fabry-Perot interferometer. *Research in Parapsychology, 1981*. Metuchen, NJ: Scarecrow Press.

Nelson, R. D., Dunne, B. J., and Jahn, R. G. (1984). *An REG experiment with large database capability, III: Operator related anomalies*. Technical Note PEAR 84003, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Nelson, R. D., Jahn, R. G., Dobyns, Y. H., and Dunne, B. J. (2000). Contributions to Variance in REG Experiments: ANOVA Models and Specialized Subsidiary Analyses. *Journal of Scientific Exploration*, **14**, 1, 473-89.

Nelson, R. D., Ziemelis, U. O., and Cook, I. A. (1992). *A Microelectronic Chip experiment: Effects of operator intention on error rates*. Technical Note PEAR 92003, Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science.

Rosenfeld, A. H. (1975). The Particle Data Group: Growth and operations. *Annual Review of Nuclear Science*, **25**, 555-599.

Rosenthal, R. (1991). *Meta-analytic procedures for social research (Revised ed.)*. Newbury Park, CA: Sage.

Schmidt, H. (1970). The psi quotient (PQ): An efficiency measure for psi tests. *Journal of Parapsychology*, **34**, 210-214.

Schmidt, H., Morris, R., and Rudolph, L. (1986). Channeling evidence for a PK effect to independent observers. *Journal of Parapsychology*, **50**, 1-15.

Targ, R. (2000). Remote Viewing in a Group Setting. *Journal of Scientific Exploration*, **14**, 107-114.

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

Timm, U. (1973). The measurement of psi. *Journal of the American Society for Psychical Research*, **67**, 282-294.

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science*, **6**, 363-403.

Acknowledgements

The Princeton Engineering Anomalies Research program is supported by grants from the John E. Fetzer Institute, the McDonnell Foundation, the Ohrstrom Foundation, Mr. Laurance S. Rockefeller, and Donald Webster, along with other philanthropic agencies and individuals. Special thanks are extended to my colleagues, Robert Jahn, York Dobyns, and Brenda Dunne, for valuable discussions.