

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*

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.

Keywords: effect size—random event generator—random number—REG—RNG—normalization—inter-experiment comparison—meta-analysis—statistics—experimental yields—bits—trials—time normalization

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 thereby creates 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 meanshift 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, 1988; 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 (PRP) 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 file drawer 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 composed of 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 pre-defined 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 during the two intentional conditions when the target system was active and labile in this sense was used. For PRP 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 at PEAR are the longest running and most deeply studied paradigm. There are several variations, but a basic description applies generally and will give an idea of the conduct of all

our experiments. The design is called "tripolar" to reflect three conditions of intention: high, low, and baseline. This means that an operator (PEAR's name for the "subject" or participant) tries to get the REG to produce results either higher or lower than expectation according to an instruction for the current trial or run, or to let the REG produce uninfluenced baseline trials. The experiment takes place in a comfortable setting, with the operator sitting in a chair roughly a meter distant from the REG itself for the basic local trials. There is usually feedback presented in a dedicated numerical display or by computer graphics, although there are a number of options including no feedback. After an introduction and general instruction, the experimenter withdraws to allow the operator to focus on the REG and develop his or her own strategy for interaction with the machine. The operators are not told how to achieve the intended results, but are allowed to develop their own strategies. Most report that they wish for or envision the desired outcome, and that they try to become attuned to the device, to be resonant or friendly with it. All data recordings, and issues of security and integrity, are managed automatically by the hardware and software.

All the REG experiments have a recorded data unit of "trials", approximately 1-second long, that are the sums of 200 bits, taken in series with lengths ranging from 1000 to 5000 trials per intention (Nelson et al., 1984, 1991, 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 is a large machine, 6 feet wide and 10 feet high, built into the wall opposite a couch. In a single 12-minute run, 9000 $\frac{3}{4}$ -inch balls fall from a central opening at the top through an array of 330 pins into 19 collecting bins. Operators sit on the couch and try to shift the mean of the resulting quasi-Gaussian distribution to the right or left compared to a baseline run. Software 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 a tripolar protocol, 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), operators sit in a comfortable chair in front of an aesthetically designed pendulum consisting of a 30-inch long fused silica shaft and a quartz crystal bob 2 inches in diameter. It is enclosed in a clear acrylic case, and feedback is provided by changing the color of light to represent degree of success in keeping the pendulum swinging or damping it, relative to baseline. The measured 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 3-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 consisted of 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 (CHIP) experiment is the error rate in 1-second trials of 1000 bits (Nelson et al., 1992), which operators try to increase or decrease. 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 (in which the intention assignment was randomly changed for every trial) were used as the standard subset.

In the PRP experiments, one person, the percipient, tries to envision the scene visited by a second person, the agent. There is typically a verbal description and sketches, but 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 et al., 1983, 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

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

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, $Y(x)$.

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 Figures 1 and 2.

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.

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,

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

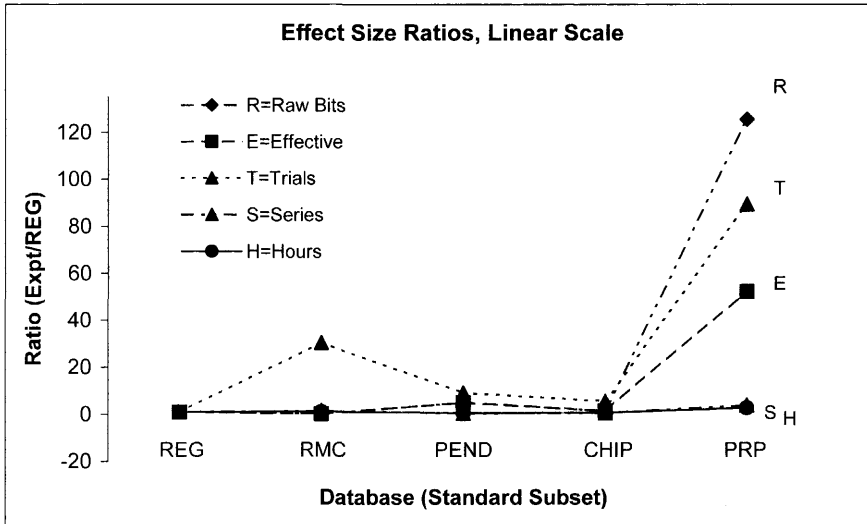


Fig. 1. The ratio of the effect size for each of the five experiments is calculated relative to the REG effect size and plotted on a linear scale.

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 pre-defined 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.

